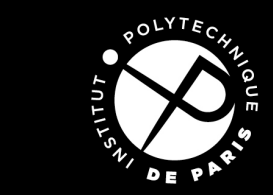
A study of some trade-offs in statistical learning : online learning, generative models and fairness

Nicolas Schreuder



INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAG004

Thèse de doctorat

A study of some trade-offs in statistical learning: online learning, generative models and fairness

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Nationale de la Statistique et de l'Administration Économique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 5 Octobre 2021, par

NICOLAS SCHREUDER

Composition du Jury :

Alexandre Tsybakov Professeur, ENSAE (CREST)	Président
Aurélien Garivier Professeur, Ecole Normale Supérieure de Lyon (UMPA)	Rapporteur
Massimiliano Pontil Professeur, IIT & University College London	Rapporteur
Gérard Biau Professeur, Sorbonne Université (LPSM)	Examinateur
Lorenzo Rosasco Professeur, Università di Genova	Examinateur
Arnak Dalalyan Professeur, ENSAE (CREST)	Directeur de thèse
Victor-Emmanuel Brunel Professeur, ENSAE (CREST)	Co-directeur de thèse

On doit être un logicien ou un grammairien rigoureux,
et être en même temps plein de fantaisie et de musique.

Hermann Hesse

Remerciements

> Soyons reconnaissants aux personnes qui nous donnent du bonheur ; elles sont les charmants jardiniers par qui nos âmes sont fleuries.
>
> *Marcel Proust*

En premier lieu, je tiens à remercier Arnak de m'avoir fortement encouragé à poursuivre en thèse puis d'avoir accepté de devenir mon directeur. Tu m'as permis de dépasser un sous-ensemble important de mes (nombreux) doutes en me redonnant confiance lorsque j'étais en proie à la lassitude. Merci de m'avoir transmis, par l'exemple, ta vision de la recherche et ton goût pour les articles bien écrits. Je tâcherai d'être à la hauteur de ce que tu m'as appris. Finalement, merci de m'avoir accordé tant de liberté pendant ma thèse pour explorer ce qui m'intéressait. Le chemin a parfois été rude mais, grâce à ton encadrement, j'en ressors assurément grandi, fier et prêt à affronter la suite.

Je tiens ensuite à remercier Victor-Emmanuel de m'avoir apporté sur un plateau mon premier problème de thèse. Je garde un plaisant souvenir de nos moments passés à réfléchir au tableau de choses plus ou moins claires pour moi - mais toujours excitantes ! - et de ton enthousiasme à présenter de nouveaux problèmes mathématiques.

Спасибо большое Evgenii, my amazing co-author. Since Saint-Flour it has been a pleasure to share ideas with you and to learn - so much - from you. Thank you for patiently introducing to the topic of statistical fairness and for proposing me to work with you. I am delighted that we still have plenty of papers to write together! Спасибо also to Monika and Lufi for the nice moments spent together in our neighborhood.

Je tiens à exprimer ma reconnaissance aux membres du jury, dont la présence m'honore profondément. Un grand merci à Aurélien Garivier d'avoir accepté d'être rapporteur de ma thèse puis d'avoir rédigé un rapport si détaillé, généreux et encourageant. Thank you very much Massimiliano for taking the time to review my thesis despite your overloaded schedule and thank you for the kind review. I hope to meet you someday in Genova! Merci Sacha de

m'avoir fait découvrir l'élégance des statistiques mathématiques à l'ENSAE. Vos enseignements et vos travaux ont toujours été - et continueront à être - une source fondamentale d'inspiration. Merci Gérard d'avoir eu la gentillesse d'accepter de faire partie de mon jury et d'avoir eu une influence indirecte mais stimulante sur cette thèse. Tante grazie Lorenzo for joining my jury from Boston. I am looking forward to working with you in Genova soon!

Cette aventure doctorale a été enrichissante aussi bien sur le plan scientifique que sur le plan humain. Pendant ces trois années, j'ai eu la chance de rencontrer et de côtoyer de nombreuses personnes qui m'ont inspiré, encouragé et avec qui j'ai pu partager des moments agréables et mémorables. Je remercie les doctorants et les chercheurs du CREST qui, chacun à leur manière, participent à en faire un environnement de travail accueillant et dynamisant. Je tiens particulièrement à remercier Guillaume de sa bienveillance, de m'avoir initié à la recherche en apprentissage statistique pendant mon master et finalement de m'avoir permis de poursuivre en thèse au CREST. Merci Pierre pour tes excellents cours à l'ENSAE et pour l'entrain que tu arrives à répandre dans le labo, même depuis le Japon. Merci Anna et Jaouad de nous servir de modèle en étant à la fois si sympathiques et si brillants. Merci Cristina, Matthieu et Nicolas pour les déjeuners et cafés partagés au Magnan et sur la terrasse du laboratoire. Merci Vianney pour ta capacité à - bien - saler les conversations. Merci Christophe, Katia et Mohamed de m'avoir irradié de votre bonne humeur depuis le bureau adjacent au mien.

Merci Jules pour ta gestion exceptionnelle des emplois du temps, pour les encadrements de statapp, pour tous nos échanges autour de la littérature, du jazz et de la vie. Merci aussi à Fabien de m'avoir souvent accueilli dans votre bureau et au A Bout de Souffle. Merci Meyer pour toutes les pauses-café à rallonge, à discuter de tout et surtout de notre prochaine idée révolutionnaire. J'espère que nous finirons un jour par lancer notre start-up ! Merci François-Pierre d'avoir régulièrement partagé avec moi - et allégé - les galères de la thèse, merci aussi d'avoir participé avec Martin à l'organisation - parfois laborieuse - du séminaire le plus ambitieux du labo. Merci à mes co-bureaux, Gabriel, Lucie, Alexander pour la bonne ambiance permanente en 3032. Merci Badr, le chef historique du bureau, pour ton enthousiasme contagieux et pour tes conseils. Merci Amir et Avetik, mes frères de thèse sur qui j'ai toujours pu compter. Merci Flore, Julien et Suzanne d'insuffler de votre dynamisme et de votre gaieté dans le labo. Merci Yannick d'être un tel phénomène. Merci Simo pour tes précieux conseils. Merci Geoffrey de m'avoir aidé quand j'étais perdu dans les processus empiriques. Je souhaite bon courage aux nouvelles générations de doctorants en les remerciant pour les agréables déjeuners partagés : Clara, Hugo, Etienne, Gabriel, Nayel, Yannis... Merci Arnaud, Edith et Pascale d'avoir pris le temps de m'expliquer avec patience les démarches administratives à suivre quand je n'y comprenais rien.

Je tiens à aussi remercier mes amis, à qui cette thèse et moi-même devons énormément. Merci Laura de ton précieux et indéfectible soutien tout au long de la thèse, notamment pendant les trajets en voiture entre Paris et Palaiseau et dans des bars aux tarifs excessifs. Merci aussi à Armand et toute la bande ajaccienne pour les excellents moments passés ensemble. Merci à mon ami Sholom que je remercierai de vive voix pour lui faire plaisir. Merci Pierre/gros pour nos innombrables discussions, pour nos aventures à vélo et pour l'introduction à la théorie des topos. Merci Johan pour les machines de Turing clockables et pour Bukowski. Merci Chloé pour nos discussions littéraires à toute heure de la journée - et de la nuit. Merci Diane de m'avoir éduqué sur le plan juridique. Merci Jeanne et Quentin, mes partenaires de bibliothèque à Jussieu. Merci Simon pour nos échanges intensément revigorants. Merci à

Thibaut, avec qui j'ai partagé mes premières joies mathématiques, pour les tours de barque sur le lac d'Annecy à disserter sur la fin du monde. J'en profite pour exprimer ma profonde gratitude à Mr. Lerasle qui m'a magistralement introduit à la beauté des mathématiques au lycée Berthollet. Merci Laetitia pour les centaines de kilomètres de balades le long de la Seine après une bonne journée de travail - ou de procrastination. Merci Anthony pour nos séances matinales et méditatives de pêche. Merci Caroline d'avoir partagé - et aggravé - ma passion pour les plats asiatiques ainsi que de m'avoir régulièrement soutenu et encouragé toutes ces années. Merci Claire d'avoir été et de continuer à être si attentionnée.

Je tiens à remercier sincèrement Jean-Michel Lasry d'avoir pris le temps de partager avec moi ses inestimables conseils et d'avoir su me convaincre d'avancer lorsque j'hésitais. J'en profite pour remercier Clément, Matthieu, Vincent et le reste de l'équipe ML de Kayrros. Travailler avec vous a été un véritable plaisir !

Merci à celles et ceux que j'ai moins eu l'occasion de voir ces derniers temps, mais dont la présence me ravit toujours : Jean, Parvati, Caroline, Eléonore, Ségal, Yoan, Ignacio, Kris, Laurie, Léo, Suzanne, Laura, Marylou, Camille, Joris...

En dernier lieu, je tiens à remercier toute ma famille dont l'intérêt persistant pour ma thèse m'a toujours étonné, mais surtout réjoui. Je remercie mon grand-père de m'avoir transmis le goût des lettres, des sciences, de la montagne et de tant d'autres belles choses. Je suis infiniment reconnaissant à mes parents de m'avoir soutenu pendant mes études, de m'avoir constament conforté dans mes choix ainsi que d'avoir permis à ma curiosité de croître sans limite. Enfin, je remercie mon petit frère dont le parcours me rend si fier.

A mon grand-père,
Dr. Claude Riondel.

Contents

List of Symbols

$\|\cdot\|$	the Euclidean norm on $\mathbb{R}^p$
$\|\cdot\|_\infty$	the supremum norm (on the relevant space)
$\mathcal{N}(\mu, \sigma^2)$	the Gaussian distribution with mean μ and variance σ^2
$f\sharp\mu$	the pushforward (or image) measure of measure μ by the measurable map f
$\mathrm{Lip}_1(\mathcal{X})$	the set of 1-Lipschitz real-valued functions on the metric space $\mathcal{X}$
$\mathsf{W}_p(P, Q)$	the p-Wasserstein metric between measures P and Q
$\mathbb{1}_A(\cdot)$	the indicator function of the set A
$a \vee b$	the maximum between real numbers a and b
$a_n \lesssim b_n$	$a_n \leq cb_n$ where c is a constant independent of n
$a_n \asymp b_n$	$a_n \lesssim b_n$ and $b_n \lesssim a_n$
$a_n = o(b_n)$	$a_n/b_n \to 0$ as $n \to +\infty$

CHAPTER 1

An introduction to statistical learning problems

On the most fundamental level, the goal of a statistician is to infer patterns and identify structures underlying a given phenomenon from observational data. More formally, the statistician is given observations $Z_1, \ldots, Z_n$ taking values in an abstract space $\mathcal{Z}$ as well as a task related to these data points. Throughout this report, we will make the assumption that the observations are *independent* realizations of a random variable Z, distributed according to a probability distribution P^*. This corresponds to the standard i.i.d. ("*independent and identically distributed*") assumption which underlies a large portion of the statistical theory literature. Given those observations, the statistician wants to infer some properties — determined by the task to be solved — of the unknown probability distribution P^*.

Contents

1.1 A typology of learning problems

There is a large variety of learning tasks and problems that can be tackled by a statistician (see, *e.g.,* Friedman, Hastie, and Tibshirani (2001)). In this thesis, we will present theoretical contributions covering several distinct frameworks studied in the statistical learning literature. For the sake of clarity, we propose a rather coarse and non-exhaustive typology of statistical learning problems to give a more precise feeling of what this thesis is about. The distinctions we make stem from three usual questions of interest in any learning task: *(i)* does the statistician have access to all the data at once ? *(ii)* are the data points labelled ? *(iii)* should all the features be treated the same ?

Online and batch learning. The first distinction concerns the way the data are received and treated by the statistician. Perhaps the batch learning setting is better known since it is usually the first one to be taught in most introductory machine learning courses. In this setting, the statistician has access to the whole dataset $Z_1, \ldots, Z_n$ *at once* and proposes a learning procedure which is based on all the data points. But this is not always the case: in some applications, data points are received *sequentially* and the statistician needs to develop a learning procedure which *adapts* to every new data point. This corresponds to the online learning setup. For example, in web advertisement placement, a given ad can be shown to a bunch of successive visitors, generating a stream of data which consists of the available characteristics of the users as well as whether or not they clicked on the ad. In such case, the statistician wants to learn which ad should be shown to which kind of user and adapt his/her strategy as more data becomes available. Note that it might also be the case that the statistician has to treat the data sequentially for computing power and/or storage constraints.

Supervised and unsupervised learning. The second distinction concerns the presence (or absence) of a *label* for each data point. In supervised learning, the random variable Z can be decomposed into a set of features X and a label Y. For instance, X could be some measurements from a flower (*e.g.,* the lengths and widths of its sepal and petal) and Y the specie of the flower, as in the well-known Fisher's Iris data set (Fisher 1936). The statistician's goal is then to learn a function that maps an input feature to an output label based on example input-output pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ where X_i and Y_i are, respectively, the features and the label of the i-th observation. In Fisher's Iris dataset example, the labels were determined by botanists according to specific criteria.

There are two important points we would like to underline: first, the features might not contain all the information about the criteria used by botanists to classify the flowers of interest, hence the learning procedure might have to make decisions with incomplete information on the label of interest; second, labelling flowers requires human knowledge and time which are sparse — and therefore costly — resources. An empirical consequence of the last remark is that most of the time, no explicit label is given or available in a dataset, leading to the *unsupervised learning setting*. The goal of the statistician is then to find some structure in the data. For instance, continuing with Fisher's Iris dataset example, the statistician might only have access to measurements from unidentified flowers and has to group, in an automatic manner, flowers that are "similar" in some sense to be determined.

Fair learning. Finally, the third distinction that has attracted a lot of attention in recent years and plays an important role in the present manuscript, concerns the way the features (which can be roughly defined as components of the data) are considered. It might be the case that some features need to be treated differently from the others for moral or ethical reasons. For instance, if a statistician's task is to develop an algorithm for hiring decisions in the US or in France, the statistician has to make sure that it does not discriminate individuals based on their gender or their ethnicity. Fair learning is about making decisions while being aware that some aspects of the data have to carefully be taken into account.

The typology of statistical learning problems we have introduced is deliberately coarse. We refer the reader to the books (Bishop 2006; Shalev-Shwartz and Ben-David 2014) for a broad covering of the most common statistical learning problems. Nevertheless, the goal of our typology is to give the possibility to most readers to grasp what the chapters of this thesis are about. In the next introductory chapters, we will refine our definitions and introduce mathematically sound frameworks to formalize those concepts.

Before diving into the details, let us clarify what we mean by statistical learning theory. Machine learning algorithms are celebrated for their impressive performance on many tasks that we thought were dedicated to human minds, from handwritten digits recognition (LeCun et al. 1990) to cancer prognosis (Kourou et al. 2015). Statistical learning theory is the branch of machine learning which aims at providing *(i)* a powerful modelling formalism for inference problems *(ii)* a better understanding of the statistical properties of learning algorithms. Statistical learning theory sits at the intersection of statistics, probability theory, functional analysis, among other domains, and constitutes a general sound mathematical framework for learning.

Recently, the striking development and use of machine learning algorithms in the real world raised an important question in the author's mind: why bother studying in theory procedures that works well in practice ? Wouldn't it be more beneficial to focus on the implementation of high-performance algorithms such as deep neural networks ? In the author's humble opinion, matured by those three years of PhD, statistical learning theory participates to this effort since the statistical learning theory framework allows to *(i)* get a better understanding of the cases in which an algorithm performs well *(ii)* quantify trade-offs inherent to learning for better-informed algorithmic choices *(iii)* provide insights to develop new algorithms which will eventually outperform existing ones or tackle new tasks. We hope that the reader will find in this thesis a satisfactory illustration of those three key points.

The introduction of this manuscript is split into two chapters. The rest of this first chapter is devoted to the presentation of two traditional learning problems considered in this thesis. In Section 1.2, we introduce a framework for online learning problems, in particular we focus on the quantification of uncertainties of estimators. Section 1.3 concerns generative models as an unsupervised learning problem for sampling. Finally, Chapter 2 serves as a general introduction to (supervised) fair learning.

1.2 Online learning and anytime deviation bounds

In the online learning setup, it is often the case that the sample size n of an experiment is unknown in advance. For instance, if we assume that the acquisition of each data point x has a cost $\psi(x) \geq 0$, this number n might be given by

$$n = \max\{k \in \mathbb{N}^* : \psi(X_1) + \ldots + \psi(X_k) \leq B\},$$

where $B > 0$ is a given available budget.

In the following we will assume that random variables $X, X_1, X_2, \ldots$ are independently drawn from a probability distribution P^* on some space $\mathcal{X}$. As usual, the statistician wants to infer some property of the distribution P^* such as its mean, its median, etc., given that such quantities are well-defined. We will denote by θ^* the quantity of interest and assume that it belongs to some subspace Θ of the Euclidean space $\mathbb{R}^p$. As we have pointed out, unlike in the usual (offline) setup, the size n of observed data is not fixed once and for all. In particular, n can be an arbitrarily complex function of the past observations. Therefore, the statistician has to propose a *sequence* of estimators $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ of the quantity of interest θ^* such that, for any positive integer n, $\widehat{\theta}_n$ is an estimator of θ^* based on the first n observations, *i.e.,* a measurable function of the data $X_1, \ldots, X_n$.

Of course, as taught in the most basic statistics courses (Wasserman 2013), a point estimation is void if it does not come with a measure of uncertainty and any careful statistical estimation procedure must provide a way of computing confidence intervals (also referred to as deviation bounds in the learning literature) around a proposed estimator. For a given level of confidence $\delta \in (0, 1)$ and a given norm $\|\cdot\|$ on $\mathbb{R}^p$, it can be formalized through a sequence $(c(n, \delta))_n$ of positive real numbers such that

$$\mathbb{P}_{(X_1,\ldots,X_n)}\left(\left\|\widehat{\theta}_n - \theta^*\right\| \leq c(n, \delta)\right) \geq 1 - \delta, \quad \forall n \geq 1. \tag{1.1}$$

It goes without saying that the goal of the statistician is to make the terms of the sequence $(c(n, \delta))_n$ as small as possible. There are mainly two levers for this task: the choice of the estimator and the ability to compute uncertainty precisely.

Most of the literature on statistical inference provides bounds such as the ones in Eq. (1.1). Those are well-suited for the usual offline setup but, as we will see, may come short in the online learning setup. When the sample size n is random and eventually depends on the observed data , an analogous to Eq. (1.1) is given by

$$\mathbb{P}_{(X_1,X_2,\ldots)}\left(\forall n \geq 1, \ \left\|\widehat{\theta}_n - \theta^*\right\| \leq c(n, \delta)\right) \geq 1 - \delta. \tag{1.2}$$

Importantly, to take into account the fact that the sample size n is random, the probability measure depends on the whole sequence $X_1, X_2, \ldots$ and the inequalities must hold for any sample size n on the *same* high-probability event.

A simple trick for offline-to-online transformation of deviation bounds is to apply the union bound trick. Indeed, assume that the statistician needs to provide deviation bounds for the first $m \geq 1$ observations but only has access to bounds from Eq. (1.1). Using a simple union bound argument yields

$$\mathbb{P}_{(X_1,\ldots,X_m)}\left(\forall 1 \leq n \leq m, \left\|\widehat{\theta}_n - \theta^*\right\| \leq c(n, \delta)\right) \geq 1 - m\delta.$$

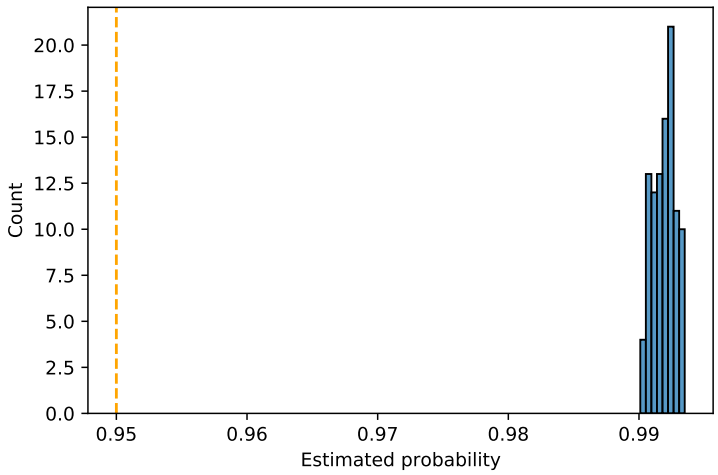


Figure 1.1: Looseness of the union bound probability. In this experiment we set $\delta = 0.05$ and $n = m = 100$. The probability is estimated computing empirical frequencies over 10000 independent draws of $n = 100$ standard normal variable. The estimation procedure is independently repeated 100 times to obtain an empirical distribution for the empirical frequencies. The broken vertical line represents the theoretical lower bound on the probability obtained with the union bound trick.

There may be two problems with this trick: first, if one does not adapt the sequence $(c(n, \delta))_n$ the confidence goes to $-\infty$ as m grows. To overcome this benign issue, one can obtain a confidence of level at least $1 - \delta$ taking the sequence $(c(n, \delta/m))_{n\in[m]}$. However, the resulting coverage might be too loose as illustrated in the following simple example.

Assume that P^* is the standard normal distribution $\mathcal{N}(0, 1)$. Denoting by q_α the standard normal percentile of order $\alpha \in (0, 1)$, the sequence $c(n, \delta) = \frac{q_{1-\delta/2}}{n}, n \geq 1$, is tight in the sense that

$$\mathbb{P}_{(X_1,\dots,X_n)}\left(|\bar{X}_n| \leq c(n, \delta)\right) = 1 - \delta, \quad \forall n \geq 1.$$

As shown in Figure 1.1, the uniform bound obtained with the union bound trick is highly conservative: while the union bound gives a lower bound on the probability of level 95%, the actual probability is greater than 99%. Intuitively, there is no reason that the union bound trick preserves tightness of the sequence $(c(n, \delta))_n$ since the union bound is a very general trick. In particular it does not take into account any particular structure in the sequence of estimators $(\widehat{\theta}_n)_n$ and the resulting probability ignores the potential dependencies between successive events.

This toy example illustrates that the simplicity of the union bound trick comes at the price of a looser control and, therefore, that it is necessary to carefully design confidence bounds for a sequence of estimators tailored for the online setting. The crucial difference is that, while deviation bound in the usual setting holds for a given sample size with high probability, we will be looking for deviation bounds which hold *for any sample size* with high probability. We will call such bounds *anytime* or *uniform* bounds.

The contribution we will present was inspired from Jamieson et al. (2014), who obtained anytime bounds for the empirical means of i.i.d. data. The rate of their bounds matches that of the law of iterated logarithm, which we will recall in the next section. The authors exploited their bound to prove optimality of a new procedure for the best-arm identification problem, a sub-problem of the multi-armed bandit framework. The literature on the multi-armed bandit is vast and since it concerns a small fraction of this manuscript, we do not intent to present it here. We refer the reader to the excellent book (Lattimore and Szepesvári 2020) for a general presentation of the theory and the algorithms developed for this framework.

We begin by recalling the asymptotic and finite-sample theory of sample averages before moving on to more general estimators, the so-called convex M-estimators. We then present our contributions which are anytime bounds for a general class of convex M-estimators.

1.2.1 Asymptotic theory of sample averages on the real line

Let us precise the considered setting and introduce some basic definitions. Random variables $X, X_1, X_2, \dots$ are independently drawn from a probability distribution P^* on some measurable subset $\mathcal{X}$ of the real line $\mathbb{R}$. We denote their mean by $\mu := \mathbb{E}[X]$ which we assume to be finite. We consider the sequence of sample averages $\bar{X}_n := n^{-1} \sum_{i=1}^n X_n, n \geq 1$. Note that we consider here the univariate setting for the sake of simplicity. The results we will present generalize to multivariate settings.

Two fundamental theorems on the asymptotic behaviour of the sample means. Perhaps the most fundamental theorems in statistics are the law of large numbers (LLN) and the central limit theorem (CLT). They state that a sample average converges almost surely or in probability to the population average, and if one zooms in by multiplying by a square root factor, a much weaker form of stochastic convergence still holds, namely, convergence in distribution[1] towards a Gaussian law. For the sake of completeness, let us formally recall those two fundamental theorems. We refer the reader to Jacod and Protter (2012, Chapter 20 & 21) for proofs of these results.

Theorem (Law of large numbers)**.** *Assume that* $\mathbb{E}|X| < +\infty$*. Then, the sequence* $(X_n)_n$ *converges* almost surely *to* μ*,* i.e.,

$$\mathbb{P}\left(\lim_{n \to +\infty} \bar{X}_n = \mu\right) = 1. \tag{LLN}$$

Theorem (Central limit theorem)**.** *Assume that* $\sigma^2 := Var[X] < \infty$*. Then*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n \to +\infty]{d} \mathcal{N}(0, \sigma^2). \tag{CLT}$$

The law of iterated logarithm. An intermediate result, known as the law of iterated logarithm (LIL), shows what happens in between the two scales. By zooming in slightly less

[1] Given a probability distribution P with Cumulative Distribution Function (CDF) F and sequences of random variables $Y_1, Y_2, \dots$ with associated CDFs $F_1, F_2, \dots$, we say that the sequence $(Y_n)_n$ converges in distribution to P (which we write $X_n \xrightarrow[n \to +\infty]{d} P$), if for any number $x \in \mathbb{R}$ at which F is continuous, $F_n(x) \to F(x)$ as $n \to +\infty$.

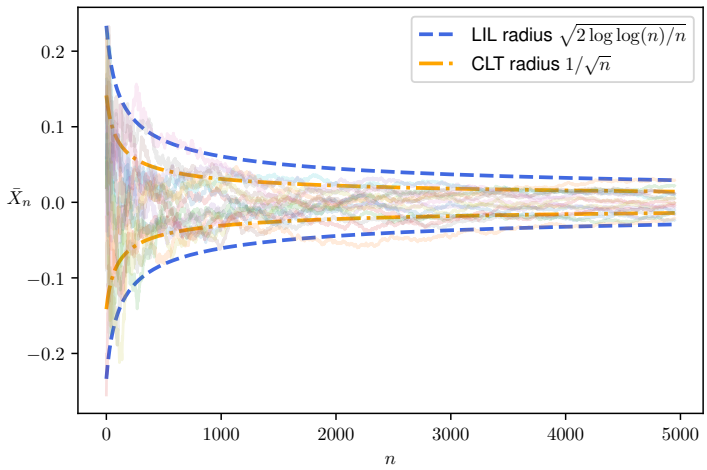


Figure 1.2: Illustration of the central limit theorem (CLT) and the law of iterated logarithm (LIL). Each transparent line represents a sequence of empirical averages from i.i.d. centered Bernoulli random variables.

than in the CLT, *i.e.*, by rescaling the sample average with a slightly smaller factor than in the CLT, it is possible to gain a guarantee for infinitely many sample sizes, almost surely. The precise statement of the law of iterated logarithm, discovered by Khintchine (1924) and Kolmogoroff (1929) almost a century ago, is as follows:

Theorem (*Law of iterated logarithm*)**.** *Assume that* $\sigma^2 := Var[X] < \infty$*. Then*

$$-1 = \liminf_{n\to\infty} \frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma\sqrt{2\ln\ln n}} \leq \limsup_{n\to\infty} \frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma\sqrt{2\ln\ln n}} = 1, \quad \textit{almost surely.} \tag{LIL}$$

This provides a guarantee on the deviations of the sample average as an estimator of the mean μ since it yields that, with probability one, for any constant $c > 1$, there exists a random integer $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$

$$|\bar{X}_n - \mu| \leq c\sigma\sqrt{\frac{2\ln\ln n}{n}}.$$

As compared to the deviation guarantees provided by the central limit theorem, the one resulting from the law of iterated logarithm has the advantage of being *valid for any sample size large enough*. This advantage is gained at the expense of a factor $(\ln\ln n)^{1/2}$. Akin to the classic version of the CLT, the applicability of the LIL is limited by the fact that it is hard to get any workable expression of n_0.

1.2.2 From asymptotic theory to finite-sample bounds

The results we have stated in the previous section are all asymptotic in nature: they yield insights on the behaviour of the sample averages as the sample size n approaches $+\infty$. Such a

situation is purely theoretical since in practice one only has access to a finite number of samples. There are mainly two ways to overcome this gap between theory and practice: pretend that those results (roughly) hold for n large enough; develop non-asymptotic counterparts to those results *i.e.,* obtain quantitative control of the sample averages for finite sample size n. We will focus on the latter option in what follows.

Non-asymptotic counterpart to the CLT. In the case of the CLT and its use in statistical learning, the drawback related to n_0 was lifted by exploiting concentration inequalities, such as Hoeffding or Bernstein inequalities, that can be seen as non-asymptotic versions of the CLT. We refer the reader to (Boucheron, Lugosi, and Massart 2013) for a general presentation of the topic of concentration inequalities. Perhaps the simplest of those bounds, known as Hoeffding's inequality (Hoeffding 1994) deals with bounded random variables. For clarity of exposition we state a particular case, the more general statement can be found in (Vershynin 2018, Theorem 2.2.6).

Theorem (*Hoeffding's inequality for bounded random variables*)**.** *Assume that X is bounded in $[0,1]$ almost surely. Then, for any $n \geq 1$ and $\delta \in (0,1)$,*

$$|\bar{X}_n - \mu| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}, \quad \textit{with probability} \geq 1-\delta.$$

It is easy to show that if X is distributed according to a Gaussian distribution with standard deviation $1/2$, then the sample averages $\bar{X}_n$ satisfy the inequalities of the previous theorem. Hence such results are not limited to bounded distributions. Actually, one can define a large class of distributions which satisfy such inequalities, known as sub-Gaussian distributions: we say that the random variable X has a sub-Gaussian distribution with parameter $\sigma > 0$ if, for all $t > 0$,

$$\mathbb{P}\left(|X - \mu| > t\right) \leq 2e^{-t^2/2\sigma^2}. \tag{1.3}$$

The sub-Gaussian property is particularly handy because it can equivalently be expressed in many different ways such as moment conditions or conditions on the moment generating function. We refer the reader to Vershynin (2018, Section 2.5) for a collection of results on sub-Gaussian distributions, such as the following theorem.

Theorem (*Hoeffding's inequality for sub-Gaussian random variables*)**.** *If X is σ-sub-Gaussian, then, for any $n \geq 1$ and $\delta \in (0,1)$,*

$$|\bar{X}_n - \mu| \leq \sigma\sqrt{\frac{2\ln(2/\delta)}{n}}, \quad \textit{with probability} \geq 1-\delta.$$

Non-asymptotic counterpart to the LIL. Combining the union bound trick presented in the previous section with Hoeffding's inequality for sub-Gaussian random variables, one can easily obtain the following anytime bound on the sample means: for any $\varepsilon > 0$ and for any $\delta \in (0,1)$, it holds, with probability at least $1 - (1+\varepsilon^{-1})\delta$,

$$|\bar{X}_n - \mu| \leq \sigma\sqrt{\frac{2\ln(n^{1+\varepsilon}/\delta)}{n}}, \quad \forall n \geq 1. \tag{UB}$$

Note that unlike Hoeffding's bound which presents the same $1/\sqrt{n}$ rate as the CLT, the rate obtained in (UB) is slower than the $\sqrt{\frac{\ln \ln n}{n}}$ rate of the LIL. Is it then possible to obtain an anytime bound whose width has the rate given in the LIL ? Several works (Jamieson et al. 2014; Kaufmann, Cappé, and Garivier 2016; Howard et al. 2018) provide a positive answer to this question. For instance, let us state the following result from Jamieson et al. (2014) which served as a starting point for our work.

Theorem (Jamieson et al. 2014, Lemma 3)**.** *If X is σ-sub-Gaussian, then for any $\varepsilon \in (0,1)$ and $\delta \in (0, \log(1+\varepsilon)/e)$, we have*

$$\mathbb{P}\left(\forall n \geq 1, \bar{X}_n \leq c_\varepsilon \sigma \sqrt{\frac{1}{n}\log\left(\frac{\log((1+\varepsilon)n)}{\delta}\right)}\right) \geq 1 - c'_\varepsilon \delta^{1+\varepsilon},$$

where $c_\varepsilon, c'_\varepsilon$ are constants which only depends on ε.

The width of the last theorem matches that of the LIL; hence, one cannot hope to obtain a tighter dependence on n.

Up until now we have focused on the estimation of the mean using the empirical mean, which is probably the most well-known in statistics. The mean is not necessarily the only quantity of interest in statistics. For instance, one might be interested in the quantiles of a given probability distribution such as the median. In particular, the mean is not defined for some heavy-tailed distributions such as the Cauchy distribution, while the quantiles are always well-defined. In the next section, we introduce convex M-estimators, a large class of popular estimators which encompasses, among others, the sample mean, and we present known results on such estimators.

1.2.3 Convex M-estimators

In a foundational paper of the theory of robust statistics (Huber 1964), Huber proposed a generalization of the maximum-likelihood estimation procedure motivated by the estimation of a location parameter from contaminated Gaussian data. The resulting estimators, presented in the next paragraph, are known as M-estimators.

Let $\phi : \mathcal{X} \times \Theta \to \mathbb{R}$ be a given loss function. Throughout this work, we make the tacit assumption that the random variable $\phi(X, \theta)$ has a finite expectation for all $\theta \in \Theta$. The population and the empirical risks are then defined, respectively, by the formulas

$$\Phi(\theta) = \mathbb{E}_{X \sim P^*}\left[\,\phi(X,\theta)\,\right], \quad \widehat{\Phi}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i, \theta),$$

where $n \geq 1$ is an integer. We denote by θ^* a minimizer of the function Φ on Θ. The M-estimators associated to the loss function ϕ are then defined as minimizers of the empirical risk:

$$\widehat{\theta}_n \in \arg\min_{\theta \in \Theta} \left\{ \widehat{\Phi}_n(\theta) \coloneqq n^{-1} \sum_{i=1}^{n} \phi(X_i, \theta) \right\}. \tag{1.4}$$

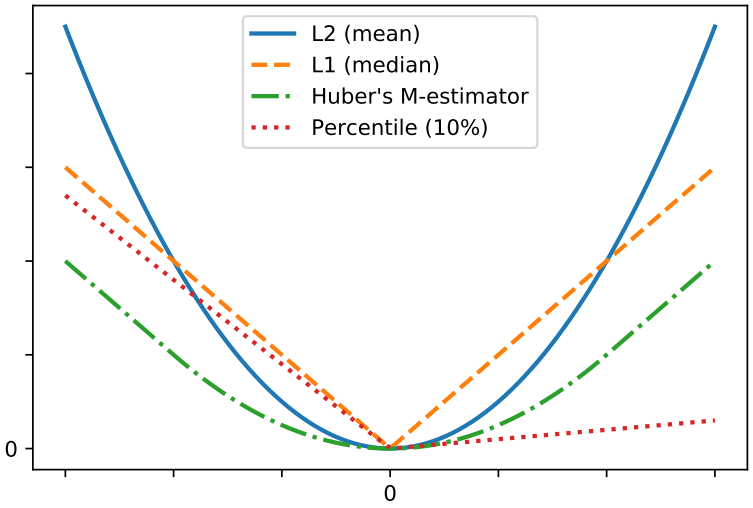


Figure 1.3: Four common choices for the loss ϕ.

We will consider the class of *convex* M-estimators which are those M-estimators associated to a loss function ϕ such that the mapping $\phi(X, \cdot)$ is convex for P^*-almost all values of X.

Let us now remind some popular convex M-estimators and the associated loss function ϕ. Considering the squared loss $\phi(x, \theta) = (x - \theta)^2$, $\widehat{\theta}_n$ defined in Eq.(1.4) is the sample average. The absolute loss $\phi(x, \theta) = |x - \theta|$ yields the estimation of the median with empirical median. Finally, we shall present another popular convex M-estimator known as Huber's M-estimator (Huber 1964): for $c > 0$, define the mapping

$$g_c(x) = \begin{cases} x^2 & \text{if} |x| \leq c \\ c(2|x| - c) & \text{if} |x| > c \end{cases},$$

and let $\phi(x, \theta) = g_c(x - \theta)$. See Figure 1.2.3 for an illustration of common loss function ϕ.

Under mild assumptions, M-estimators are both consistent and asymptotically normal (Haberman 1989; Niemiro 1992), *i.e.*, suitably adapted versions of the LLN and the CLT apply to them. Moreover, some versions of the LIL were also shown for M-estimators (Arcones 1994; He and Wang 1995). They suffer, however, from the limitations explained above for the standard LIL; namely their asymptotic nature which makes it hard to use them in practice. Is it possible to obtain anytime deviation bounds whose width has the rate given in the LIL for a general class of convex M-estimators ? Up to our knowledge, no such results were available.

1.2.4 Contributions

Non-asymptotic LIL for M-estimators. Our contributions allow to circumvent the limitations exposed in the previous section by providing general anytime deviation bounds whose width has the rate given in the LIL for a general class of convex M-estimators.

In Chapter 4, we extend the anytime bounds for sample averages obtained by Jamieson et al. (2014) to a large class of M-estimators containing, among others, the median, the quantiles, and Huber's M-estimator. Under mild assumptions on the loss function ϕ and the data distribution P^*, Theorem 4.2.4 states that there exists an explicit positive constant $c > 0$ such that the M-estimators $(\widehat{\theta}_n)$ defined in (1.4) satisfy

$$\mathbb{P}\left(\forall n \geq n_0,\ |\widehat{\theta}_n - \theta^*| \leq c\sqrt{\frac{\ln\ln n + \ln(1/\delta)}{n} + o\left(\frac{1}{n}\right)}\right) \geq 1-\delta, \qquad \forall \delta \in (0,1).$$

Note that the presence of the asymptotic term $o\left(\frac{1}{n}\right)$ in the formulation above was added for the sake of simplicity. The result stated in Theorem 4.2.4 is fully non-asymptotic and provides explicitly all the constants and terms hidden in the little-o notation. Theorem 4.3.5 provides an extension of our univariate result to penalized multivariate M-estimators for the problem of predicting a real-valued label given a d-dimensional feature vector. Our formalism is applicable to settings such as the maximum a posteriori approach and penalized empirical risk minimization.

Application to the problem of Best Arm Identification. For a given set of K (unknown) probability distributions $\mathbb{P}_1, \ldots, \mathbb{P}_K$, Best Arm Identification (BAI) in the fixed confidence setting consists in identifying, for a given confidence level, the distribution which has the highest expected outcome while minimizing the number of samples (sequentially) drawn from those distributions (Audibert and Bubeck 2010; Gabillon, Ghavamzadeh, and Lazaric 2012; Kaufmann, Cappé, and Garivier 2016). Naturally, the same problem can be formulated for finding the distribution with the largest median, or the largest quantile of a given order. In particular, such a formulation of the problem might be of interest in cases where the expectations of the outcomes of each arm may not be defined (rewards are heavy-tailed) or are not meaningful (rewards are subject to some arbitrary contamination) (Altschuler, Brunel, and Malek 2018). We show that the univariate bounds we established can be converted into an extension of lil'UCB algorithm from Jamieson et al. (2014) with provably optimal theoretical guarantees: one can replace the empirical mean by any M-estimator satisfying our assumptions while preserving optimality of the procedure for best-arm identification according to the corresponding M-estimator.

1.3 Generative models

A short introductory story. On 27 July 1890, aged 37, Vincent Van Gogh hopelessly shot himself in the chest with a 7mm Lefaucheux pin fire revolver (Sweetman 1990). More than a century later, he is acclaimed as one of the most revolutionary painters of human history. He left us with around nine hundred known paintings, which are now exposed all around the world. This number might still grow slightly if, for instance, someone discovers a forgotten painting of Vincent Van Gogh in his attic [2]. Nevertheless, the total number of paintings that Vincent Van Gogh created is finite and fixed forever. Despite this undeniable fact, one might wonder: how would Van Gogh have painted a modern scene such as a street from Arles in France nowadays ? Modern painters could try to mimic his style to fulfil such curiosity but it requires a lot of work and will potentially be quite expensive. Can we automatize such a task, to overcome this limitation ? With all its promises, could modern artificial intelligence continue Vincent Van Gogh's work and fill us with paintings similar in style to his ? Since most painters and artists in general train themselves by mimicking their masters, designing an AI with such abilities would constitute a first step in the direction of developing creative artificial intelligence.

Figure 1.4: A famous painting by Vincent Van Gogh, "Terrasse du café le soir" depicting a café terrace at night in Arles, France. It was painted in mid-September 1888 according to the Wikipedia page dedicated to this painting.

[2] This actually happened in 2013 as related in the New York Time's article "A Van Gogh's Trip From the Attic to the Museum".

1.3.1 Problem presentation and formalization

Problem presentation. The problem of learning generative models has attracted a lot of attention during the last 5 years in machine learning and artificial intelligence. The most prominent example is generating artificial images that look similar to actual photographs, by means of generative adversarial networks (GANs) (Goodfellow et al. 2014). The general formulation of generative adversarial networks can be given as a game between a generator which aims to draw samples similar to the observed samples and a discriminator that learns to distinguish the true samples from the simulated ones. Those two players are usually implemented in practice as neural networks. As we will see, a trade-off appears in this competition: the quality of the overall model depends on the relative ability of the players to achieve their task. We postpone a more thorough presentation of GANs and the associated adversarial generative models to the next subsection.

Let us provide a few reasons to study generative models and GANs in particular. Perhaps the most prominent and possibly straightforward application of GANs concerns *data augmentation*, a set of techniques used to generate new data from existing one to increase the number of available training data points (Antoniou, Storkey, and Edwards 2017). Such techniques are increasingly useful as modern algorithms, such as deep learning-based algorithms which require a massive amount of data to be trained. As argued by Goodfellow (2016), GANs may also prove useful to improve performance of model-based reinforcement learning. More generally, generative models can be used for likelihood-free inference, when intractability issues arise (see, *e.g.,* Diggle and Gratton 1984; Briol et al. 2019). Last but not least, modern generative models could be used for generating art, such as paintings, poems, music, texture of sounds, etc. For instance, WaveNet, a deep-neural network architecture, is able to generate original and often highly realistic musical fragments (Oord et al. 2016). Just as supervised learning algorithms are used as a tool for decision-making, generative models could be used as a tool for creating art.

Problem formalization. Assume that we are given a collection of i.i.d. random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from some probability distribution P^* supported on a subset $\mathcal{X}$ of $\mathbb{R}^D$, where D is typically a large integer (e.g., $D = 28 \times 28 = 784$ for the `MNIST` dataset (LeCun 1998)). Those data points could represent, for example, a collection of digitalized paintings from Vincent Van Gogh or a set of financial features from different individuals, depending on the reader's personal taste. Morally, we would like to generate new data points $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ such that those could have been drawn from the distribution P^*.

For example, in the one-dimensional case, we could ask that a Kolmogorov-Smirnov test fails to distinguish between the set of generated samples and the observed samples at a prescribed level. We shall give more precise definitions of how we measure the closeness of the generated samples to the original samples in a subsequent paragraph. One (naive) way of achieving our goal is to sample with replacement from the collection of observed samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, which will mimic the sampling process from P^* such as in bootstrap approaches. However, the reader will agree that this procedure is far from desirable in our case, as it does not bring anything new than what we already had. Thus, a desirable first property of a sampling procedures that we will be looking for is the ability to *generalize* beyond the observed data, *i.e.,* to generate data points which were not observed but are similar, in some sense, to the observations.

Remark 1.3.1 (Some comments on sampling)**.** *Sampling is a fundamental problem in statistics which covers a wide variety of topics. In this thesis we focus on a specific sampling problem in which we observe data from an unknown probability distribution and would like to sample new data points. We refer the reader to the book of Devroye (1986) for general techniques such as inversion and rejection to generate/sample random variables given an analytic expression for their distribution. We also mention that there is a whole body of literature on Markov chains based sampling techniques, known as Markov Chain Monte Carlo (MCMC) (Robert and Casella 2004) which is in part motivated by Bayesian statistics applications (Robert 2007).*

Prescribed and implicit probabilistic models. In order to go beyond purely empirical observations, we need to carefully choose a probabilistic model to describe our data. We will follow Mohamed and Lakshminarayanan (2016) who proposes an interesting distinction between two classes of probability models: *prescribed* and *implicit* probabilistic models.

The former class of so-called *prescribed probabilistic models* requires an *explicit* parametric description of the distribution of the observed data points through a likelihood function. Those models are ubiquitous in theoretical statistics, statistical learning, etc. In particular, they underpin the well-known maximum likelihood approach: for a given parametric specification of the likelihood function $(f_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$, find a parameter $\boldsymbol{\theta}$ inside a set $\Theta \subset \mathbb{R}^p$ which is the most likely to have generated the data:

$$\widehat{\boldsymbol{\theta}}_n \in \arg\max_{\boldsymbol{\theta}\in\Theta} \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{X}_i).$$

The main limitation of prescribed probabilistic models is that one needs to be able to explicitly provide a (preferably relevant) likelihood function to model the observed data – an increasingly difficult task as the complexity of the data grows. Moreover, even when the parameter θ is known, sampling a random variable according to the density $f_{\boldsymbol{\theta}}$ might still constitute a challenging task (Robert and Casella 2004).

Taking a rather different but complementary approach, *implicit probabilistic models* define a stochastic procedure that directly generates data. In particular, they do not require a likelihood function for the observed data nor the existence of a density with respect to a prescribed measure. Such models are usually based on a latent variable procedure: first, a d-dimensional latent variable $\boldsymbol{Z}$ is drawn according to some easy-to-sample-from distribution (*e.g.,* Gaussian distribution or uniform distribution on the unit hypercube); then a chosen function $g : \mathbb{R}^d \to \mathbb{R}^D$ is applied to the latent variable $\boldsymbol{Z}$ to transform it into another variable $\boldsymbol{X} := g(\boldsymbol{Z})$. Importantly, the dimension d of the latent space is usually less than or equal to the ambient space dimension D and can be thought of as the intrinsic dimension of the observed data. Returning to the `MNIST` example, typical values of d would be between 10 and 15 (Costa and Hero 2004b; Facco et al. 2018), while, we recall that $D = 784$.

The assumption that the intrinsic dimension is smaller than the ambient one stems from an empirical observation, known as the "manifold hypothesis" (see (Fefferman, Mitter, and Narayanan 2016), and references therein), that real world data tend to lie close to a low dimensional manifold embedded in a high-dimensional ambient space. It supports the idea that, even though the observed samples are usually of very high dimension, they may exhibit significant structures such as the harmonic and rhythmic schemes followed by a melody or a

poem, or the presence of simple shapes in an image. Such an assumption is at the core of the manifold learning literature (Cayton 2005; Pless and Souvenir 2009) and could explain the ability of some algorithms to learn the structure of the data even when the observed data is extremely high-dimensional. We also refer the reader to the recent work of Goldt et al. (2020) which, in a similar flavour, proposes the "hidden manifold model" to study the influence of data structure on learning in neural networks.

Furthermore, implicit probabilistic models subsume a large class of well-known procedures. For instance, in one dimension, we recover the inversion method for sampling (Devroye 1986) by defining the mapping g as the inverse CDF of the target variable $\boldsymbol{X}$ and $\boldsymbol{Z}$ as a uniform random variable on the interval $[0, 1]$. Implicit probabilistic models are particularly convenient for data distributions which are supported on low-dimensional manifolds such as distributions which can be expressed as the pushforward measure (or image measure) of a distribution supported on a low-dimensional subspace. Such distributions do not admit a density with respect to the Lebesgue measure as a consequence of Sard's theorem, which we recall now.

Theorem 1.3.2 (Sard's theorem, Sard (1942))**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^D$ be k-times continuously differentiable (where $k \geq \max(d - D + 1, 1)$). Let $Jf(\boldsymbol{x})$ denote the Jacobian of f at the point $\boldsymbol{x}$ and let $\mathcal{C} \coloneqq \{\boldsymbol{x} : rank(Jf(\boldsymbol{x})) < D\}$ denote the critical set of f. Then the image $f(\mathcal{C})$ has Lebesgue measure zero in $\mathbb{R}^D$.*

Implicit probabilistic models are good candidates for our sampling task. Indeed, they provide a way of generating data which do not admit a simple likelihood modelization and/or do not admit a density with respect to a known measure. Such flexibility is necessary for our task, since, following the manifold hypothesis, natural data is expected to live close to an unknown low-dimensional manifold. Providing a likelihood function in such case would require estimating the manifold to choose a relevant dominating measure on/around this manifold — a challenging task which necessitates additional assumptions on the data (Genovese et al. 2012). Furthermore, implicit probabilistic models are better suited than parametric models because they yield a distribution which is close to the unknown target distribution *and* easy-to-sample from. We note that, in an iconoclastic work, Richardson and Weiss (2018) tried to fit Gaussian Mixture Models on high-dimensional images data and generated new samples from the fitted models. Even though the results are better than one could have expected, they are still far from the impressive results obtained by procedures based on implicit probabilistic models. Let us now see how can one learn implicit probabilistic models.

Likelihood-free learning in implicit probabilistic models. Mohamed and Lakshminarayanan (2016) insist on *testing* and *density estimation-by-comparison* as principles for learning in implicit generative models. In particular, they identify four ways to perform likelihood-free inference:

1. **class probability estimation**: the ratio of densities is estimated by training a classifier that discriminates real data from generated data;
2. **divergence minimisation**: use a divergence (Ali and Silvey 1966; Csiszár 1967) between the true density and the model as an objective to drive learning of the generative model;
3. **ratio matching**: minimise error between the true density ratio and an estimate of it

obtained through, *e.g.,* least squares importance fitting, see, *e.g.,* Sugiyama, Suzuki, and Kanamori (2012);

4. **moment matching**: evaluate whether the moments of the true distribution and that of the model are the same.

The authors argue that the second and the third ways are not particularly well suited for learning generative models. In the next subsection, we will focus on examples of procedures which follow either the first or the last principle for learning implicit probabilistic models.

1.3.2 Generative adversarial networks: foundations and practice

GAN foundations. Generative adversarial networks were introduced in the seminal paper (Goodfellow et al. 2014). In this initial framework, two models, implemented as deep neural networks, compete against each other: a generative model $G : \mathbb{R}^d \to \mathbb{R}^D$ and a discriminative model $D : \mathbb{R}^D \to \mathbb{R}$. Goodfellow et al. (2014) give a nice metaphor to think about those models: the generative model G can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminator model D is analogous to the police, trying to detect the counterfeit currency. The generator transforms low-dimensional latent variable $\boldsymbol{Z} \sim P_{\boldsymbol{Z}}$ (usually uniform or Gaussian) into fake data $G(\boldsymbol{Z})$, distributed according to the generating distribution $P_{\text{learner}} \coloneqq G\#P_{\boldsymbol{Z}}$. The discriminator receives data (real and fake) and needs to guess for each data point whether it comes from the generator or from the observed data. In practice, the two models are trained simultaneously through backpropagation (see, *e.g.,* Goodfellow et al. 2016, for an introduction to deep learning and deep neural networks). Figure 1.5 illustrates the general functioning of GANs. The hope is that the training leads to an equilibrium in which the discriminator is not able anymore to distinguish between generated and real data. The whole framework can be formalized as a two-player zero-sum game in which the generator G (respectively the discriminator D) maximizes (respectively minimizes) the objective

$$\mathbb{E}_{\boldsymbol{X}\sim P^*}[D(\boldsymbol{X})] + \mathbb{E}_{\boldsymbol{Z}\sim P_{\boldsymbol{Z}}}[1 - D(G(\boldsymbol{Z}))]. \tag{GAN}$$

Since the introduction of GANs, impressive results have been obtained in practice for generating realistically looking images (see, *e.g.,* Radford, Metz, and Chintala 2015; Brock, Donahue, and Simonyan 2018) but also for more complex tasks such as image-to-image translation (Isola et al. 2017) and image super-resolution (Ledig et al. 2017). In addition to its striking empirical performance, one of the main advantages of GANs is the "low" computational cost of generating new samples once the generator is learned: it amounts to drawing a low-dimensional Gaussian or uniform random vector and passing it through the generator network. A large part of subsequent works to Goodfellow et al. (2014) focused on proposing new neural network architectures or finding training heuristics and tricks to improve the quality of the generated images. We refer the reader to "The GAN zoo" GitHub repository for an index of hundreds of GAN variations[3]. As we will see in the next section, some works also introduced new formulations for the GAN objective which resulted in popular adaptation of the initial framework.

[3]There were more than 500 named GAN paper at the time of writing this manuscript.

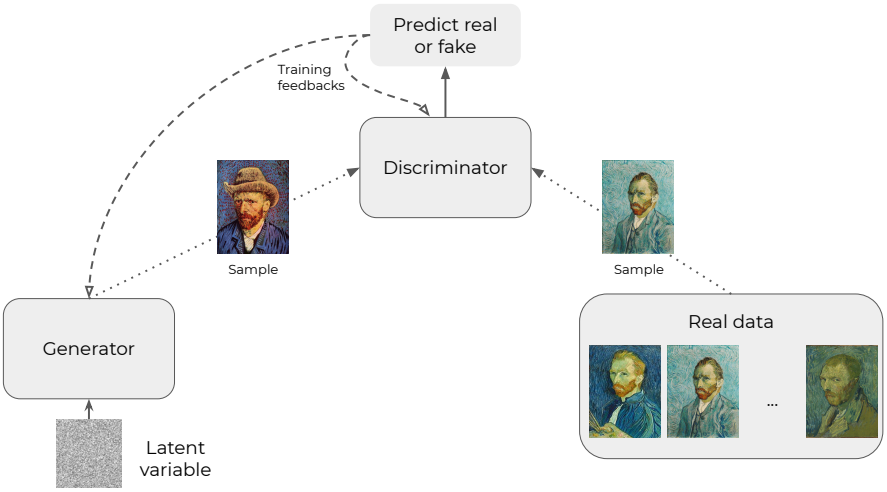


Figure 1.5: Illustration of the original GAN model on Vincent Van Gogh's self-portraits. During the training phase, real data and generated data are fed to the discriminator (dotted arrows) which in turn must predict which data is real and which is fake. Feedback (in the form of gradients of the loss) are then sent to the generator and the discriminator (broken arrows) based on predictions from the latter to update their parameters (through back-propagation in the case of neural networks). Note that the generator cannot see the real data.

Limitations of GANs and challenges. Despite their impressive empirical performance, GANs are notoriously hard to train. Even though some fixes have been proposed (Salimans et al. 2016), several problems are yet to be fully understood and solved such as, for instance, the lack of originality of generated samples or the so-called "mode collapse" problem in which the generator is only able to sample from one mode while the true distribution P^* is multi-modal. Furthermore, as for most deep learning models, a sound mathematical theory for understanding GANs is lacking. In particular, a fully satisfactory statistical framework for studying GANs is yet to be developed. In the next section we present some theoretical works on (or inspired by) the GAN framework.

1.3.3 Generative adversarial model theory

In what follows, we call *generative adversarial model* any abstraction of the original GAN model in which a generator learns against a general (and potentially abstract) adversary.

Since GANs initially emerged from the deep learning community, the first line of work primarily relied on empirical insights and general mathematical intuitions. Later on, a parallel line of work tackled the GAN problem from the statistical perspectives (Biau, Sangnier, and Tanielian 2020; Biau et al. 2018; Chen et al. 2020; Liang 2018; Singh et al. 2018; Luise, Pontil, and Ciliberto 2020; Uppal, Singh, and Poczos 2019) as well as optimization and algorithmic viewpoints (Pfau and Vinyals 2016; Kodali et al. 2017; Liu, Bousquet, and Chaudhuri 2017; Nagarajan and Kolter 2017; Genevay, Peyré, and Cuturi 2017; Genevay et al. 2018; Liang and Stokes 2018; Nie and Patel 2020). We begin this section by explaining how to obtain a general formulation for the generative adversarial model objective before presenting the main

Table 1.1: Some popular IPMs.

Metric name	Class $\mathcal{F}$
Total Variation distance	$\{f : \|f\|_\infty \leq 1\}$
Dudley metric	$\{f : \|f\|_\infty + \|f\|_L \leq 1\}$
Kolmogorov distance	$\{\mathbb{1}_{(-\infty,t]}(\cdot) : t \in \mathbb{R}^d\}$
Maximum Mean Discrepancy (MMD)	$\{f : \|f\|_{\mathcal{H}} \leq 1\}$ for RKHS $\mathcal{H}$

known statistical results.

From the GAN discriminator to Integral Probability Metrics. Goodfellow et al. (2014) showed that minimizing the objective (GAN) with respect to the generator G against optimal discriminator $D^*(G)$ (*i.e.,* the discriminator which maximizes (GAN)) amounts to minimizing (*w.r.t.* G) the Jensen-Shannon (JS) divergence between the generated data distribution $G\sharp P_{\boldsymbol{Z}}$ and the real sample distribution P^*. Arguing that the topology induced by the JS divergence is rather coarse, Arjovsky, Chintala, and Bottou (2017) proposed to replace this divergence by the Wasserstein-1 distance, leading to the so-called *Wasserstein GAN*. More precisely, the goal of the generator G in this variant is to generate data from a distribution that is as close as possible, w.r.t. the Wasserstein-1 distance, to the empirical distribution of the original data. This leads to the (empirical) objective

$$\mathsf{W}_1(G\sharp P_{\boldsymbol{Z}}, \widehat{P}_n) = \sup_{f \in \mathrm{Lip}_1(\mathcal{X})} \left| \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) - \mathbb{E}_{\tilde{\boldsymbol{X}} \sim G\sharp P_{\boldsymbol{Z}}}[f(\tilde{\boldsymbol{X}})] \right|, \tag{1.5}$$

where $Lip_1(\mathcal{X})$ is the set of 1-Lipschitz real-valued functions on the (metric) space $\mathcal{X}$. In view of this relation, which follows from the Kantorovitch-Rubinstein duality theorem (Villani 2008, Theorem 5.9, Remark 6.5), the Wasserstein distance admits a nice interpretation as a sampling error. Furthermore, replacing the class of Lipschitz functions by an arbitrary functional class $\mathcal{F}$, we obtain a pseudo-metric over the space of Borel probability measures $\mathcal{P}$,

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{\boldsymbol{X} \sim P}[f(\boldsymbol{X})] - \mathbb{E}_{\boldsymbol{Y} \sim Q}[f(\boldsymbol{Y})]|, \quad P, Q \in \mathcal{P}. \tag{IPM}$$

All such pseudo-metrics over the space of probability measures constitute the so-called Integral Probability Metrics (IPM) (Müller 1997). Depending on the choice of $\mathcal{F}$, IPM encompass a large family of popular probability metric as illustrated in Table 1.1. We refer the reader to Sriperumbudur et al. (2012), Liang (2019) and reference therein for theoretical results on IPM such as sample complexities and computationally tractable consistent estimators. An interesting fact is that Csiszár ϕ-divergences (Csiszár 1964), which comprise the well-known Kullback-Leibler divergence, and IPM are fundamentally different and only intersect at the Total Variation distance.

An IPM can naturally be interpreted as an adversarial loss: to compare two probability distributions, it seeks for the function f^* in $\mathcal{F}$ for which the expectations of $f(\boldsymbol{X})$ under the two distributions have the largest discrepancy. The class $\mathcal{F}$ and the induced IPM can then be seen as an abstraction of the discriminator from the initial GAN framework which compares distributions via a (generalized) moment matching principle. We can now state a general

(population) objective for abstract adversarial generative models: given a class of admissible generators $\mathcal{G}$ and a class of discriminators $\mathcal{F}$, find a minimizer $G \in \mathcal{G}$ of

$$d_{\mathcal{F}}(G\sharp P_{\boldsymbol{Z}}, P^*) = \sup_{f\in\mathcal{F}} |\mathbb{E}_{\boldsymbol{X}\sim P^*}[f(\boldsymbol{X})] - \mathbb{E}_{\boldsymbol{Z}\sim P_{\boldsymbol{Z}}}[f(G(\boldsymbol{Z}))]|.$$

In practice we usually do not have access to the true distribution P^* which we replace by the empirical measure $\mathbb{P}_n$ as a proxy to obtain the empirical objective. This substitution opens the way to many interesting statistical questions.

1.3.4 Existing statistical results

From a statistical perspective, the usual goal is to obtain a bound on the discrepancy between the learned distribution P_{learner} and the true distribution P^* of the data $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$, with respect to a given evaluation metric d. A particularly relevant task is the quantification of the rate of convergence to zero of this discrepancy as the sample size n grows to infinity. Given a family of candidate distributions $\mathcal{P}$, typical bounds are of the form

$$\mathbb{E}_{(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)\sim P_{\text{obs}}}\left[\mathsf{d}(P_{\text{learner}}, P^*)\right] - \inf_{P\in\mathcal{P}} \mathsf{d}(P, P^*) \lesssim n^{-r(\alpha,\beta,d,D)}.$$

for some exponent $r(\alpha, \beta, d, D) > 0$, where the parameter α characterises the *complexity* of the discriminator, β represents the *smoothness* of the generator, d is the intrinsic dimension of the data and D is the ambient dimension (*e.g.*, the number of pixels in an image). Since D is typically much larger than d, it is desirable to avoid any decaying dependence on D in the exponent $r(\alpha, \beta, d, D)$ — potentially avoiding the curse of dimensionality.

The statistical results on generative adversarial models can be split in three categories depending on the considered loss: vanilla GAN (Biau et al. 2018), IPM-based loss (Liang 2018; Chen et al. 2020; Uppal, Singh, and Poczos 2019; Singh and Póczos 2018) and optimal transport based loss (Genevay, Peyré, and Cuturi 2018; Biau, Sangnier, and Tanielian 2020; Luise, Pontil, and Ciliberto 2020). The last two intersect at Wasserstein-1 loss for GANs.

In the first category, Biau et al. (2018) is among the first to provide mathematical and statistical insights on the "vanilla" GAN problem. In particular, the authors propose a sound mathematical analysis of the connection between the Jensen-Shannon divergence and the GAN population-level objective, as well as the role of the discriminator family in this connection. Finally, they derive several large sample properties of the estimated parameters and distributions in a parametric statistics style such as asymptotic normality of the estimated generator's parameters. Finite-sample guarantees are yet to be derived for this setting.

Most of the subsequent works focused on different losses, fostered by improvements of variants of GAN's initial objective. We now present results based on IPM-based loss. Liang (2018) obtained both parametric and non-parametric rates of convergence for learning distributions under IPM losses. For instance, Liang (2018, Theorem 1) can informally be stated as follows.

Theorem (Informal)**.** *Assume that the target density as well as the IPM loss $d_{\mathcal{F}}$ are characterized by D-dimensional Hilbert-Sobolev balls of smoothness α and β respectively. Then, the minimax optimal rate is given by*

$$\inf_{\tilde{\nu}_n} \sup_{\nu\in\mathcal{G}} \mathbb{E}[d_{\mathcal{F}}(\tilde{\nu}_n, \nu)] \asymp n^{-\frac{\alpha+\beta}{2\alpha+D}} \vee n^{-\frac{1}{2}},$$

where $\tilde{\nu}_n$ is any estimator for ν based on n i.i.d. *drawn samples $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim \nu$. The minimax rate is achieved by a thresholded Fourier decomposition estimator.*

We notice that the rate is similar to the usual minimax rates obtained for non-parametric estimation problems (Tsybakov 2008) and presents a typical curse of dimensionality behaviour in the low smoothness regime $\beta < D/2$. Crucially, the proof heavily relies on the assumption that the target distribution admits a Lebesgue density on the D-dimensional hypercube $[0,1]^D$ to leverage the rich toolboxes from Fourier theory and non-parametric estimation. As a consequence, it does not generalize to scenarios when the target distribution does not admit a simple density, such as when it can be expressed as the pushforward of a low-dimensional latent distribution — a foundational assumption in GAN's applications. We note that Liang (2018) also provides upper bounds on the expected risk when functions of interest lie in RKHS balls such as in MMD GAN. Interestingly, in this latter case, the rate depends on the intrinsic dimension of the RKHS rather than the ambient dimension of the data. In a similar flavour, Chen et al. (2020, Theorem 2) obtained the following rate of convergence in the Hölder case:

$$\mathbb{E}[d_{\mathcal{F}}(\tilde{\nu}_n, \nu)] \lesssim n^{-\beta/(2\beta+D)} \log^2 n.$$

Surprisingly their rate does not scale with the smoothness of the target density, unlike Liang (2018). According to the authors, this comes from the fact that their estimator is an empirical risk minimizer: the target distribution μ is replaced by the empirical measure $\widehat{\mu}_n$ which does not contain any information regarding smoothness.

Finally, generalizing results from Liang (2018) and Singh et al. (2018), Uppal, Singh, and Poczos (2019) derived minimax rates for the estimation of non-parametric probability density in Besov spaces (which include L^p, Hölder, Hilbert-Sobolev spaces) under Besov IPM.

A common thread of the aforementioned works is the assumption of absolute continuity of the target distribution *w.r.t.* to some known measure (such as the Lebesgue measure) and, thus, they study GANs as a (non-parametric) density estimation under IPM loss problem. They yield insightful theoretical results on the performance of GANs as a non-parametric estimation procedure. Yet, they do not fully explain the impressive ability of GANs to generate high-dimensional data and seemingly avoid the curse of dimensionality. Overcoming the dependence of their bounds on the dimension D of the ambient space should necessarily pass by *(i)* dropping the density assumption *(ii)* taking advantage of the particular structure of the candidate distributions which is arguably one of the defining features of of GANs: they are expressed as pushforward measures of an easy-to-sample-from low-dimensional distribution into the high-dimensional sample space. In particular, as a consequence of Sard's theorem (see Theorem 1.3.2), we have seen that the distributions induced by a GAN generator do not admit a density with respect to the Lebesgue measure of the sample space. Hence, we would like to work with metrics which can provide meaningful measure of distances between distributions without requiring the existence of a *known* dominating measure. Optimal transport based metrics fulfil this requirement, among others, and appear as good candidates for this task. For instance, the Wasserstein-GAN variation (Arjovsky, Chintala, and Bottou 2017) relies on the Wasserstein-1 metric which is at the same time an IPM and an optimal transport metric. In this direction, Biau, Sangnier, and Tanielian (2020) provided a theoretical analysis of the Wasserstein-GAN parametrized by neural networks. In particular they obtained parametric rate of convergence under the IPM induced by a parametric class of discriminators contained in the set of 1-Lipschitz functions.

Instead of approximating the class of 1-Lipschitz functions, the last category of results — general optimal transport based loss — rely on a different, but more convenient formulation of Wasserstein distance, whose evaluation is known to be a hard computational problem. To circumvent this issue, Cuturi (2013) introduced a regularized version of the optimal transport metrics known as Sinkhorn divergences, also referred to as entropic regularized optimal transport. Following this track, Genevay, Peyré, and Cuturi (2018) proposed the first tractable method to train large-scale generative models using Sinkhorn loss. Later, Luise, Pontil, and Ciliberto (2020) provided statistical guarantees for learning generative models under the Sinkhorn loss. Interestingly and unlike the rest of the theoretical literature, the authors advocated optimizing simultaneously for the generator as well as for the latent distribution. They obtained parametric rates of convergence $n^{-1/2}$ under the Sinkhorn loss when the generator belongs to a Hölder class with smoothness larger than $d/2$. To our knowledge, Luise, Pontil, and Ciliberto (2020) is the only work, other than ours, which establishes statistical guarantees under the assumption that the data generating process is a smooth transformation of a low-dimensional latent distribution.

1.3.5 Contributions

As we have seen, most statistical results on generative adversarial models are either stated under the stiff parametric framework or under the assumption that the target distribution is absolutely continuous *w.r.t.* a base measure such as the Lebesgue measure. However, positing that the target distribution P^* has a density with respect to the Lebesgue measure, or any other dominating σ-finite measure μ on $\mathbb{R}^D$, is, in general, incompatible with the fact that P^* is inherited from a low-dimensional latent variable and supported by a low-dimensional manifold (see Sard's theorem, Theorem 1.3.2). As a consequence of the restriction to dominated distributions, the available statistical results fail to assess the benefits of the reduced dimension d of the latent space (as compared to the ambient dimension D) on the quality of the generative model and present a typical curse of dimensionality behaviour.

In Chapter 6, we propose a convenient framework for studying adversarial generative models from a statistical perspective to assess the impact of dimension reduction on the error of the generative model. In this work, the latent distribution is chosen as the uniform on the d-dimensional hypercube $\mathcal{U}_d$. For a smooth class of real-valued functions $\mathcal{F}$ on the sample space $\mathbb{R}^D$ and the Integral Probability Metric $\mathsf{d}_{\mathcal{F}}$, the risk of a generator $G : [0,1]^d \to \mathbb{R}^D$ is defined by

$$R_{\mathsf{d}_{\mathcal{F}},P^*}(G) \coloneqq \mathsf{d}_{\mathcal{F}}(G\sharp\mathcal{U}_d, P^*) \coloneqq \sup_{f\in\mathcal{F}} |\mathbb{E}_{\boldsymbol{U}\sim\mathcal{U}_d}[f(G(\boldsymbol{U}))] - \mathbb{E}_{\boldsymbol{X}\sim P^*}[f(\boldsymbol{X})]|.$$

In particular, this formulation encompasses that of Wasserstein-1 GAN.

Assuming that the distribution of the training samples, up to some noise and adversarial contamination, is a smooth transformation of the uniform distribution on the d-dimensional hypercube, we establish non-asymptotic risk bounds for the Empirical Risk Minimizer (ERM) for which the exponent of the rate depends on the latent space dimension d, while the ambient dimension D only enters as a constant multiplicative factor – hence our rate does not exhibit the usual non-parametric curse of dimensionality. Furthermore, our new bounds, which are of independent interest, leverage both the smoothness of the distribution of the samples and that of the functions in the IPM class $\mathcal{F}$.

In Chapter 5, we provide upper bounds on the expectation of the supremum of empirical processes indexed by Hölder classes of any smoothness and for any distribution supported on a bounded set in $\mathbb{R}^d$. These results can alternatively be seen as non-asymptotic risk bounds, when the unknown distribution is estimated by its empirical counterpart, based on n independent observations, and the error of estimation is quantified by an IPM indexed by a Hölder class. These results interpolate between two well-known extreme cases: the rate $n^{-1/d}$ corresponding to the Wasserstein-1 distance (the least smooth case) and the fast rate $n^{-1/2}$ corresponding to very smooth functions (for instance, functions from a RKHS defined by a bounded kernel). Those theoretical results enable us to obtain our bounds in Chapter 6.

CHAPTER 2

An introduction to fair learning

> If knowledge can create problems, it is not through ignorance that we can solve them.
>
> *Isaac Asimov*

Statistical algorithms trained on personal data take pivotal decisions which influence our lives on a daily basis. Recent studies show that a naive use of these algorithms in sensitive domains may lead to unfair and discriminating decisions, often inheriting or even amplifying biases present in data (Barocas and Selbst 2016). Consider an example from a recent survey on the subject (Barocas, Hardt, and Narayanan 2017): "Amazon uses a data-driven system to determine the neighbourhoods in which to offer free same-day delivery. A 2016 study found stark disparities in the demographic make-up of these neighbourhoods: in many U.S. cities, white residents were more than twice as likely as black residents to live in one of the qualifying neighbourhoods.". This example highlights a worrying trend that data-driven algorithms can lead to unfair decisions in much more sensitive domains such as, for instance, court decisions[1], school/university admissions, loan approvals, etc. Therefore, there is a growing need for ensuring that practical algorithms are not contradicting neither moral nor legal grounds while still being useful. This issue sits at the intersection of several domains such as political philosophy, sociology, statistics, computer science, etc. Thus, its full comprehension must ultimately involve interdisciplinary discussions. Following the wisdom of the Latin expression "*Sutor, ne ultra crepidam*"[2], we will primarily focus on aspects of this issue which lie inside our domain of expertise. In particular, we will rely on recent advances in learning theory which allow to address some aspects of this problem within a rigorous statistical framework. We

[1] See Propublica's study of the risk assessment software COMPAS that is used in US courts to assess the likelihood of a defendant becoming a recidivist.

[2] The expression literally means "Shoemaker, not beyond the shoe" and was used to warn people to avoid passing judgment beyond their expertise (Wikipedia).

refer the reader to Mehrabi et al. (2019) and Barocas, Hardt, and Narayanan (2019) for a general introduction on algorithmic fairness and to Oneto and Chiappa (2020) and Barrio, Gordaliza, and Loubes (2020) for reviews of the most recent theoretical advances.

In this chapter of the manuscript we will focus on the problem of fairness-aware learning (Menon and Williamson 2018a). Once again, we do not pretend or aim to derive a general theory of fairness nor to debate about what is fair and what is not. This is a political debate, which should be held on the level of society. Our goal will be much more modest. As presented in the next section, inspired by recognized legal or philosophical concepts, (mainly) computer scientists have proposed several mathematical formalizations of discrimination of decision rules. Our goal as statisticians will be to evaluate, given a measure of performance and fairness criteria, what is the best one can hope to achieve from a learning perspective regarding fairness and performance. In particular, we will see that a natural trade-off appears between satisfying fairness constraints and achieving good (predictive) performance. Our approach to fairness-related issues does not discuss the relevance of a given choice (such as the choice of a fairness criterion and a measure of risk), which is ultimately left to the decision-maker, but yields a better understanding of the consequences of this choice. Such an approach is in phase with what Weber (1992) defined as one of the goals of science in general – not a substitute for human's judgment but a tool for informed decision-making.

Contents

2.1 Problem formalization and definitions

In what follows, we place ourselves in the supervised learning setting: a statistician, which is given couples of feature and label variables, aims to express the label as a function of the feature variables in order to predict correctly the label associated to new and unseen feature variables. The fairness-aware (supervised) learning setting slightly differs from the usual setting in that we do not treat all features equally. In particular, we distinguish between two types of features: a set of (nominally) *unsensitive* features $\mathcal{X}$ and a set of *sensitive* features $\mathcal{S}$. The reader can think of the latter as, for instance, gender, ethnicity, and/or age. Importantly, the set of sensitive features $\mathcal{S}$ contains those features against which we want to control potential discriminations. The set $\mathcal{S}$ will generally be a finite set in this manuscript. We would like to point out that in the considered framework, the choice of sensitive attributes belongs to the decision-maker, potentially incentivized by legal or ethical motives. It is not the statistician's task to determine which feature should be regarded as sensitive.

More formally, the statistician observes independent copies of a triplet $(\boldsymbol{X}, S, Y) \sim P$ where $\boldsymbol{X}$ is the feature vector, Y the label variable and S is a sensitive feature (*e.g.* gender, ethnicity or age). Those random variables take their values in the sets $\mathcal{X}, \mathcal{Y}$, and $\mathcal{S}$, respectively. For full generality, we will consider predictors of the form $f : \mathcal{Z} \to \mathcal{Y}$ where, following the notation from Donini et al. (2018), the set $\mathcal{Z}$ is either the set of unsensitive features $\mathcal{X}$ or the whole set of features $\mathcal{X} \times \mathcal{S}$, depending on whether the statistician is allowed to access the sensitive attribute for prediction. Analogously, we define $\boldsymbol{Z}$ as $\boldsymbol{X}$ or $(\boldsymbol{X}, S)$ depending on the type of predictions at hand. Note that any predictor f induces group-wise distributions of the predicted outcomes $\text{Law}(f(\boldsymbol{Z}) \mid S{=}s)$ for $s \in \mathcal{S}$. The general goal of the statistician is two-fold: *maximize prediction performance* by minimizing a given risk while *satisfying one or several given fairness constraints*. Let us now dive into the mathematical definitions of fairness to specify this general goal.

Individual and group fairness. Basically, the mathematical definitions of fairness can be divided into two groups (Dwork et al. 2012): *individual fairness* and *group fairness*. The former notion reflects the principle that similar individuals must be treated similarly, which translates into Lipschitz type constraints on possible (often randomized) prediction rules. The latter defines fairness on population level via (conditional) statistical independence of a prediction from sensitive attribute S (*e.g.,* gender, ethnicity). The high-level idea of *group fairness* notions can be seen as bounding or diminishing an eventual discrepancy between group-wise distributions of the predicted outcomes. In this thesis we will focus on the *group fairness*. We refer the reader to the seminal paper of Dwork et al. (2012) for an introduction to individual fairness and to Jung et al. (2019), Dwork, Ilvento, and Jagadeesan (2020), and Mukherjee et al. (2020) for examples of recent works in this context.

Disparate Treatment. Before introducing the main group fairness definitions, let us define a first natural (and naive) notion of fairness which essentially restricts the class of predictors to those which do not take as input the sensitive attribute S.

Definition 2.1.1 (Disparate Treatment)**.** *Any function $f : \mathcal{X} \to \mathcal{Y}$ that cannot receive the sensitive attribute S in its functional form does not produce* Disparate Treatment.

Gajane and Pechenizkiy (2017) refer to this definition as *fairness through unawareness*, as opposed to *fairness through awareness*, introduced by Dwork et al. (2012). The latter type of prediction allows one to build separate model for each sensitive attribute, while the former obliges one to fix a single model which is later applied across all groups. This property might be desirable for obvious legal and/or privacy reasons (Primus 2003; Barocas and Selbst 2016; Gajane and Pechenizkiy 2017; Lipton, McAuley, and Chouldechova 2018). For instance, in France it is forbidden to use ethnicity as a sensitive attribute for any statistical treatment[3] (hence in particular for prediction).

Remark 2.1.2. *For some authors, treatment disparity encompasses a larger phenomenon. For instance, Lipton, McAuley, and Chouldechova (2018) consider that disparate treatment addresses intentional discrimination which includes decisions explicitly based on protected characteristics, as in Definition 2.1.1, but also intentional discrimination via proxy variables (e.g*

[3]See "Décision n° 2007-557 DC du 15 novembre 2007 du Conseil Constitutionnel, Loi relative à la maîtrise de l'immigration, à l'intégration et à l'asile" (link to English version).

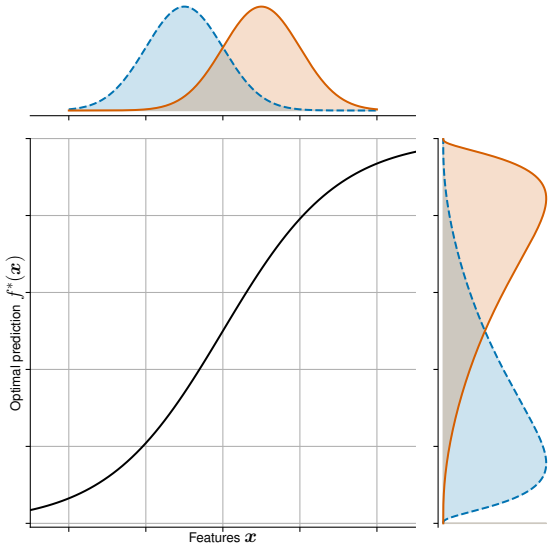


Figure 2.1: A simple example of an (optimal) predictor which does not produce Disparate Treatment but whose predictions differ across sensitive groups. The sensitive attribute S can take two values $S = 1$ and $S = 2$ with equal probability. The feature X is distributed according to a Gaussian mixture model: $X|S = 1 \sim \mathcal{N}(-1, 1), X|S = 2 \sim \mathcal{N}(1, 1)$. The group-wise feature distributions are represented on top in dotted blue ($S = 1$) and solid orange ($S = 2$). The label Y is obtained as $Y = f^*(X)$ where $f^*(x) = \frac{1}{1+e^{-x}}$. The distribution of predictions from f^* across groups are plotted on the right in dotted blue ($S = 1$) and solid orange ($S = 2$).

literacy tests for voting eligibility). Since the mathematical formalization of such phenomenon is not evident, we prefer to stick to the definition we introduced.

Importantly, the absence of Disparate Treatment does not guarantee the prediction to be *statistically independent* from the sensitive attribute S because of correlations and, more generally, dependencies between the sensitive attribute S and the feature vector $\boldsymbol{X}$ (Pedreshi, Ruggieri, and Turini 2008). Indeed, consider the Bayes optimal predictor $\boldsymbol{x} \mapsto f^*(\boldsymbol{x})$ defined as

$$f^*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}].$$

By definition, it does not take as input the sensitive attribute and achieves the lowest possible squared risk among predictions avoiding Disparate Treatment. Yet, the predictor f^* might still promote disparity between sensitive groups if the distributions of features $\boldsymbol{X}$ differ between groups. An example of such scenario is given in Figure 2.1: even though the predictor does not produce DT (it is the same for both groups), the distribution of the predictions for the first group significantly differ from that of the second group because of the distributions of features.

Moreover, in the classification setting, Lipton, McAuley, and Chouldechova (2018) showed that avoiding Disparate Treatment (DT) is not necessarily desirable, even when combined with other fairness criteria as in Disparate Learning Processes, a class of learning procedures

which can access sensitive attribute during training but cannot use it for prediction. They prove that a decision rule that does not produce DT cannot achieve a better accuracy than an optimal rule that uses this information; leading them to the conjecture that any rule which avoids DT yields a suboptimal trade-off between fairness and performance. Furthermore, the authors provide empirical evidence that avoiding DT leads to indirect treatment disparity, via proxy variables, or within-class discrimination. Several questions stem from those insightful observations. Is it possible to provide satisfactory theoretical justifications of their results ? Are those phenomenon limited to the classification setting ? How fair and accurate can one hope to be by allowing to treat subgroups differently ? Chapter 7 and 8 will provide a tentative theoretical answer to those questions in the regression setting.

Disparate Impact. As we have seen, given a predictor f, the prediction distributions $(\text{Law}(f(\boldsymbol{Z})|S=s))_{s\in\mathcal{S}}$ may differ across sensitive subgroups and, importantly, this phenomenon may happen *unintentionally*, *i.e.,* not because of a pernicious statistician, but as a *by-product* of the learning process. The decision-maker might deem as undesirable such difference; for instance, if a predictor f is used to determine the salaries of employees in a company, the board might want (or have to) choose a predictor which yields similar salary distributions across subgroups (*e.g.,* men and women). Hence, most fairness notions focus on controlling the discrepancy between the distributions of the predictions across groups. We will now provide some of the most popular fairness definitions regarding the distributions of the predictions. Since most of the literature focuses on the fair *classification* problem (see Calders, Kamiran, and Pechenizkiy (2009) and previous references), the subsequent fairness definitions were initially given in this framework. For clarity we will provide those definitions in the binary classification with binary-valued sensitive attribute setting. We will explain, when needed, how to extend those definitions to regression problems and to non-binary sensitive attributes. Those extensions will be based on the independence property between random variables. We will use the notation $A \perp\!\!\!\perp B$ to express the independence between random variables A and B.

First of all, it might be the case that the risk of a predictor is small on average across the whole population but with a high group-wise discrepancy in risk. Such a situation could be considered as discriminatory for groups with high level of risk. To prevent such issues, we introduce a first definition of fairness in binary classification, which asks for equality of group-wise risks.

Definition 2.1.3 (Equality of Group-wise Risks)**.** *A classifier* $f : \mathcal{Z} \to \{0, 1\}$ *achieves* Equal Group Wise risk *with respect to the distribution* $\mathbb{P}$ *of* $(\boldsymbol{X}, S, Y)$ *if*

$$\mathbb{P}(f(\boldsymbol{Z}) \neq Y \mid S = 0) = \mathbb{P}(f(\boldsymbol{Z}) \neq Y \mid S = 1).$$

Buolamwini and Gebru (2018) argues that it corresponds to the requirement that all group receive good service and Zafar et al. (2017) advocates for this notion to avoid disparate mistreatment. Note that this definition immediately generalizes to any expected risk notion. A relaxed formulation of this fairness notion was considered in the context of regression by Agarwal, Dudík, and Wu (2019).

Next, we introduce the main fairness notion that this thesis will focus on. It was formally introduced by Calders, Kamiran, and Pechenizkiy (2009) and essentially asks for the prediction to be (statistically) independent from the sensitive attribute.

Definition 2.1.4 (Demographic Parity)**.** *A classifier* $f : \mathcal{Z} \to \{0, 1\}$ *achieves Demographic Parity with respect to the distribution* $\mathbb{P}$ *of* $(\boldsymbol{X}, S, Y)$ *if*

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1).$$

Simply put, this definition assigns equal chances of positive decision for both groups. It acts as an incentive to promote diversity through affirmative action (Mouzannar, Ohannessian, and Srebro 2019). We can equivalently express Demographic Parity as the (probabilistic) independence between the distribution of the prediction $f(\boldsymbol{Z})$ and that of the sensitive attribute S, which we denote by

$$f(\boldsymbol{Z}) \perp\!\!\!\perp S. \tag{DP}$$

Such characterization immediately generalizes the initial definition to any supervised learning problem and any sensitive attribute.

As argued by Hardt, Price, and Srebro (2016), if deployed in reality, such a notion may cause more negative effects than positive ones in some cases. An example they give is connected with credit landing, where $Y = 1$ means that an individual $(\boldsymbol{X}, S)$ is able to pay back the loan and $f(\boldsymbol{Z}) = 1$ means the bank approves the credit. In case when the paying abilities of two groups are drastically different, providing equal chances of getting a credit without looking at the paying ability Y sets less privileged individuals to the path of default. To circumvent the above issue Hardt, Price, and Srebro (2016) proposed the following two definitions.

Definition 2.1.5 (Equalized Odds)**.** *A classifier* $f : \mathcal{Z} \to \{0, 1\}$ *achieves* Equalized Odds *with respect to the distribution* $\mathbb{P}$ *of* $(\boldsymbol{X}, S, Y)$ *if*

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0, Y = y) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1, Y = y), \quad \forall\, y \in \{0, 1\}.$$

This fairness notion asks for the prediction $f : \mathcal{Z} \to \{0, 1\}$ to equalize True Positive and True Negative rates across both groups. One immediately sees that this definition can be expressed in a more general form as

$$(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y. \tag{EOdd}$$

It means that, given the true label of an individual, knowing the value of the sensitive attribute does not bring any information on the distribution of the prediction. Typically, the equalization of True Negatives is not necessary in practice, since $f(\boldsymbol{Z}) = 1$ is interpreted as a positive decision and it can be natural to focus on such decisions. In this way we arrive at the definition of Equal Opportunity.

Definition 2.1.6 (Equal Opportunity)**.** *A classifier* $f : \mathcal{Z} \to \{0, 1\}$ *achieves* Equal Opportunity *with respect to the distribution* $\mathbb{P}$ *of* $(\boldsymbol{X}, S, Y)$ *if*

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0, Y = 1) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1, Y = 1).$$

Equal Opportunity is less constraining than Equalized Odds as it just asks for the True Positive rates to be the same across all groups. In the credit landing example, Equal Opportunity means that among the clients who are able to pay back their loan, the proportion of attributed

loans across groups should be the same. It is easy to see that this definition is equivalent to the more general conditional independence constraint

$$(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y = 1. \tag{\textbf{EOpp}}$$

Note, however, that the extension of this notion to the regression setting is not obvious at all. That is because the definition of Equal Opportunity relies on the fact that $Y = 1$ can be interpreted as a positive decision (*e.g.,* attribute a loan) and there is no immediate equivalent of a "positive decision" in the regression setting. It is still an open question to find a transposition of the notion of Equal Opportunity outside of the classification setting.

All the fairness definitions introduced up until now are expressed as the conditional probability of a positive prediction $f(\boldsymbol{Z}) = 1$ given the value sensitive attribute S and eventually the true label Y. Unlike the previous definitions, the next one deals with conditional probabilities *w.r.t.* to the event of a positive prediction ($f(\boldsymbol{Z}) = 1$).

Definition 2.1.7 (Test fairness)**.** *A classifier* $f : \mathcal{Z} \to \{0, 1\}$ *satisfies* Test Fairness *with respect to the distribution* $\mathbb{P}$ *of* $(\boldsymbol{X}, S, Y)$ *if*

$$\mathbb{P}(Y = 1 \mid S = 0, f(\boldsymbol{Z}) = 1) = \mathbb{P}(Y = 1 \mid S = 1, f(\boldsymbol{Z}) = 1).$$

As for the previous definitions, the last one can naturally be expressed as the conditional independence condition

$$(Y = 1 \perp\!\!\!\perp S) \mid f(\boldsymbol{Z}) = 1. \tag{\textbf{Test fairness}}$$

Test fairness is also referred to as *predictive parity*. It imposes equality across sensitive groups of the rates of positive outcomes $Y = 1$ among those who received a positive prediction $f(\boldsymbol{Z}) = 1$. It is tightly related to the concept of calibration (Barocas, Hardt, and Narayanan 2017, Chapter 2).

We have given five definitions of fairness which encompass different conceptions of discrimination. Considering those definitions as constraints, we can now clarify the goal of fairness-aware learning: minimize a given risk over the class of predictors which satisfy (a subset of) those definitions. Note that all the constraints are *distribution-dependent i.e.,* they depend on the joint distribution $(\boldsymbol{X}, S, Y)$. Hence, the resulting constrained optimization problems are quite different from the usual shape-constrained problems in which distribution-independent constraints are imposed on the predictors, such as smoothness constraints.

Let us now briefly expose an illuminating well-known fairness case study to point out some issues one might encounter. COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions) is a risk-assessment software developed by Northpointe Group, a technology and management consulting firm which is used to predict recidivism risk in US courts. A study by ProPublica (Angwin et al. 2016) showed that COMPAS yields a high discrepancy regarding False Positive and False Negative rates across ethnicity groups, hence it does not satisfy Equalized Odds. However, Northpointe's refutation (Dieterich, Mendoza, and Brennan 2016) claimed that COMPAS satisfies (a generalized notion of) predictive parity. This example highlights that the choice of the fairness criterion is crucial when discussing discriminatory behaviour of a decision rule. Furthermore, it brings the question of whether one should stick to a particular fairness criterion or try to satisfy several criteria simultaneously.

Of course, ideally (and naively), one would like to obtain a predictor which satisfies as many fairness constraints as possible. However this is not necessarily desirable since the class of predictors might be small or even empty. Indeed, Chouldechova (2017) showed an impossibility result which states that unless $\mathbb{P}(Y = 1|S = s)$ is the same across groups, no classifier $f : \mathcal{Z} \to \{0, 1\}$ can simultaneously satisfy Test fairness and Equalized Odds. What lessons can we take out of this ? First of all, one indeed needs to be careful about chosen fairness criteria for a given task; secondly, it might indicate that the fairness notions we consider are too stiff and it would be worth exploring ways of relaxing those constraints. This will be the subject of the next section.

A word on bias in data and causality. Before going on to the next section, let us discuss an important point to us. An attentive reader might have noticed that all the definitions we gave focus on potential discrimination on the *prediction* level while nothing has been said on the potential biases in the data. Why did we make such a choice ? In a nutshell, we are interested in the outcome of the algorithms, not in the identification of biases in the data, because the reasons for which a procedure or an algorithm can have discriminatory behaviour are, most of the time, far from obvious and potentially impossible to unveil unless we are ready to make strong assumptions about our data. For example, if features are high-dimensional images, it might be an extremely difficult task to provide or test a causal model for the data because of its complexity (Bühlmann 2013). Nevertheless, we would like to mention a complementary growing sub-branch of algorithmic fairness literature which tries to incorporate causal reasoning into fairness-aware learning (see,*e.g.,* Kilbertus et al. 2017; Kusner et al. 2017; Loftus et al. 2018; Makhlouf, Zhioua, and Palamidessi 2020). They provide new formalizations of fairness notions and new procedures to overcome potential bias and discriminations using the machinery of causal learning. The setting that is of interest to us is one in which providing a causal model for the problem of interest is too difficult, too time consuming or too expensive. Hence we focus on controlling bias in prediction, which is something we can observe and test.

2.2 Relaxation and trade-offs.

In the previous section, we have provided some popular ways of formalizing what it means for a predictor to be fair, *i.e.,* to avoid particular discriminations. We have broadly stated the fairness-aware learning problem as that of finding a predictor f which has a low risk $\mathcal{R}(f)$ while satisfying one or several fairness criteria, such as the ones we have defined. Following this paradigm, fairness-aware learning problem can be expressed a constrained optimization problem:

$$\arg\min_{f:\mathcal{Z}\to\mathcal{Y}} \{\mathcal{R}(f) : f \text{ is “fair” } \},$$

where the constraint could be, for instance, imposing that the predictors satisfy one or several of the definition we have introduced such as Definitions 2.1.4 and 2.1.5. Given this general formulation of the fairness-aware learning problem, a natural question arises: what is the impact of introducing the fairness constraint on the risk of the best fair predictors ? In order to form a relevant answer, we will argue why *relaxation* of the fairness constraints is needed and provide different ways of achieving it. This will allow us to derive a general framework for studying the *trade-off between performance and fairness.*

Relaxation of fairness constraints. There are four main reasons why relaxation is necessary. First, mathematical fairness definitions are imperfect transposition of qualitative ideas, thus it might not be desirable to aim at exact satisfaction of the resulting constraints. Second, two sources of randomness might blur the satisfaction our definitions: we have introduced fairness definitions for fixed predictors while in practice we will be working with estimators *i.e.* predictors which depends on a finite number of samples from the joint distribution $(\boldsymbol{X}, S, Y)$; furthermore, the quantities of interest (such as the False Positive Rate for instance) also have to be estimated from finite samples. Those unavoidable uncertainties and the resulting errors needs to be taken into account in the constraints. Third, a drastic change of a prediction policy can have a negative impact: for instance, if the task consists in setting wages in a company, employees will probably not tolerate an abrupt (negative) change in their salary. We might want to introduce fairness in a continuous manner hence we need a way to interpolate smoothly between fair and unfair situations. Finally, the introduced definitions are too stiff as do not allow to compare the discriminatory behaviour of two predictors which do not satisfy the fairness constraint(s): a predictor is either considered as fair or unfair. As for the risk, we would like to have some kind of order relation on the predictors regarding their fairness.

Hence the need to find a way to relax the initial definitions and provide meaningful notions of unfairness, defined as the violation of the fairness constraints. In what follows, we will focus on relaxations of the Demographic Parity constraint as all works from this thesis are based on this constraint. Recall that DP requires equality of two quantities of interest, $\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0)$ and $\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1)$.

A first natural way to relax exact fairness constraints is to ask for the difference of the quantities of interest to be small. In the case of Demographic Parity, this amounts to

$$|\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0) - \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1)| \leq \varepsilon,$$

for some prescribed threshold level $\varepsilon \geq 0$. One can also consider other types of relaxations; for instance, multiplicative instead of additive. Note that those ideas easily generalize to other notions of fairness we have introduced in the classification with binary-valued sensitive attribute setting, $\mathcal{S} \times \mathcal{Y} = \{0, 1\}^2$. However, it is not clear how to proceed for other settings such as regression or general sensitive attribute.

Another way of relaxing fairness criteria relates to the alternative (conditional) independence definitions we have provided. As we have seen, three popular fairness constraints can be expressed as (conditional) independence conditions depending on the joint distribution of the triplet $(f(\boldsymbol{Z}), Y, S)$. In order to formalize this idea, any metric/divergence d on the space of probability distributions (such as Kolmogorov-Smirnov distance, total variation distance, Kullback-Leibler divergence, etc.) can be used to measure the discrepancy between group-wise predictions as

$$d\big(\operatorname{Law}(f(\boldsymbol{Z})|S = s), \operatorname{Law}(f(\boldsymbol{Z})|S = s')\big),$$

for any values $s, s' \in \mathcal{S}$ of the sensitive attribute. In particular, if d satisfies the "identity of indiscernible" (namely, for any probability distributions P and Q, $d(P, Q) = 0 \implies P = Q$) Demographic Parity is satisfied if and only if

$$d\big(\operatorname{Law}(f(\boldsymbol{Z})|S = s), \operatorname{Law}(f(\boldsymbol{Z})|S = s')\big) = 0, \quad \forall s, s' \in \mathcal{S}.$$

A natural idea is to define a functional $\mathcal{U}$ on the set of predictors which quantifies the violation of the DP constraint using a chosen metric and to declare a prediction approximately fair if this functional does not exceed a user pre-specified threshold. In recent years a large variety of such relaxations has been proposed: correlation based (Baharlouei et al. 2019; Mary, Calauzènes, and El Karoui 2019; Komiyama et al. 2018); Kolmogorov-Smirnov distance (Agarwal, Dudík, and Wu 2019); Mutual information (Steinberg et al. 2020; Steinberg, Reid, and O'Callaghan 2020); Total Variation distance (Oneto, Donini, and Pontil 2019b; Oneto et al. 2019); Equality of means and higher moment matching (Raff, Sylvester, and Mills 2018; Fitzsimons et al. 2019; Calders et al. 2013; Berk et al. 2017; Olfat et al. 2020; Donini et al. 2018); Maximum Mean Discrepancy (Quadrianto and Sharmanska 2017; Madras et al. 2018); Wasserstein distance (Chiappa et al. 2020; Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c; Gordaliza et al. 2019).

The most common relaxations of the Demographic Parity constraint are based on the Total Variation (TV) and the Kolmogorov-Smirnov (KS) distances (Agarwal, Dudík, and Wu 2019; Oneto, Donini, and Pontil 2019a; Agarwal et al. 2018; Chzhen et al. 2020a). There are various ways to use the TV or KS in order to build a functional $\mathcal{U}$, which quantifies the violation of the DP constraint. To compare those measures of discrepancy with the one that we introduce in our work, we define $\mathcal{U}_{\mathrm{TV}}$ and $\mathcal{U}_{\mathrm{KS}}$ as follows

$$\textbf{TV unfairness:} \qquad \mathcal{U}_{\mathrm{TV}}(f) := \sum_{s\in[K]} \mathrm{TV}\left(\mathrm{Law}(f(\boldsymbol{X}, S) \mid S = s), \mathrm{Law}(f(\boldsymbol{X}, S))\right),$$

$$\textbf{KS unfairness:} \qquad \mathcal{U}_{\mathrm{KS}}(f) := \sum_{s\in[K]} \mathrm{KS}\left(\mathrm{Law}(f(\boldsymbol{X}, S) \mid S = s), \mathrm{Law}(f(\boldsymbol{X}, S))\right).$$

Using these notions, one wishes to study those predictors f which satisfy relaxed fairness constraint $\mathcal{U}_{\square}(f) \leq \varepsilon$, where $\square$ is KS or TV and $\varepsilon \geq 0$ is a user specified parameter. Note that since both KS and TV are metrics, setting $\varepsilon = 0$ is equivalent to the DP constraint. Meanwhile, for $\varepsilon > 0$ these formulations allow some slack. It is known that the TV distance is rather strong and extremely sensitive to small changes in distributions which is the major drawback of the TV unfairness. This limitation can be addressed by the KS unfairness due to an obvious relation $\mathcal{U}_{\mathrm{KS}}(f) \leq \mathcal{U}_{\mathrm{TV}}(f)$.

The price of fairness and the risk-fairness trade-off. Now that we have defined a quantitative notion of unfairness, we can study the impact of incorporating a fair constraint in the risk minimization problem. As usual we define the Bayes optimal predictor f^* as an unconstrained minimizer of the risk

$$f^* \in \arg\min_{f:\mathcal{Z}\to\mathcal{Y}} \mathcal{R}(f).$$

For clarity of exposition we will assume in the following that the minimum is indeed attained. The Bayes predictor, and its associated level of risk, give us a sense of the best (risk) performance one can hope to achieve for a given problem. Note that the level of unfairness of the Bayes optimal predictor depends on the problem at hand, not on the statistician's choices. In particular, if one wants to optimize for risk and unfairness simultaneously, new notions of optimality need to be defined. For a given metric $\square$ (*e.g.,* TV or KS) and a positive real number $\varepsilon > 0$, we define an optimal ε-fair predictor as a minimizer f^*_ε of the risk whose

unfairness, measured through the functional $\mathcal{U}_\square$, does not exceed the threshold *varepsilon*:

$$f_\varepsilon^* \in \arg\min\{\mathcal{R}(f) : \mathcal{U}_\square(f) \leq \varepsilon\}.$$

Since the Bayes optimal predictor achieves the smallest risk among all predictors, we must have

$$\mathcal{R}(f_\varepsilon^*) - \mathcal{R}(f^*) \geq 0,$$

for any predictor ε-fair predictor f_ε^* and for any $\varepsilon > 0$, showing that the incorporation of the fairness constraint can only increase the risk. Furthermore, since we have a natural order on the constraints, the difference in risk in Eq. (2.1) decreases with respect to the threshold ε. Thus, a natural trade-off arises between the fairness condition and the risk and it can entirely be measured through the excess risks

$$\mathcal{R}(f_\varepsilon^*) - \mathcal{R}(f^*), \quad \varepsilon \geq 0.$$

One of our goals will be to precisely understand and quantify this trade-off in the regression setting to answer questions such as: for a prescribed level of risk, what is the best achievable fairness level ? Or what is the cost in risk of switching from fairness threshold ε to ε' ?

2.3 Fair regression and optimal transport

Unlike its classification counterpart, the problem of fair regression has received far less attention in the literature. However, as argued by Agarwal, Dudík, and Wu (2019) classifiers only provide binary decisions, while in practice final decisions are taken by humans based on predictions from the machine. In this case a continuous prediction is more informative than a binary one and justifies the need for studying fairness in the regression framework. Until very recently, contributions on fair regression were almost exclusively focused on the practical incorporation of proxy fairness constraints in classical learning methods, such as random forest, ridge regression, kernel based methods to name a few (Calders et al. 2013; Komiyama and Shimao 2017; Berk et al. 2017; Pérez-Suay et al. 2017; Raff, Sylvester, and Mills 2018; Fitzsimons et al. 2018). Several works empirically study the impact of (relaxed) fairness constraints on the risk (Bertsimas, Farias, and Trichakis 2012; Zliobaite 2015; Haas 2019; Wick, Panda, and Tristan 2019; Zafar et al. 2017). Yet, the problem of precisely quantifying the effect of such constraints on the risk has not been tackled.

More recently, statistical and learning guarantees for fair regression were derived (Agarwal, Dudík, and Wu 2019; Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c; Chiappa et al. 2020; Fitzsimons et al. 2019; Plečko and Meinshausen 2019; Chzhen et al. 2020a). The closest works to our contribution are that of (Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c; Chiappa et al. 2020), who draw a connection between the problem of exactly fair regression of demographic parity and the multi-marginal optimal transport formulation (Gangbo and Święch 1998; Agueh and Carlier 2011). In particular, (Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c) derive the form of optimal fair prediction, provide statistical guarantees on plug-in type estimators, and establish the exact value of the risk of the optimal fair prediction. Let us cite in full their result which served as a starting point for our work.

Theorem 2.3.1 (Le Gouic, Loubes, and Rigollet (2020) and Chzhen et al. (2020c))**.** *Assume that for any $s \in [K]$, the random variable $(f^*(\mathbf{X}, S)|S = s)$ has a finite second moment and is non-atomic. Then,*

$$\min\left\{\mathcal{R}(f) \,:\, (f(\boldsymbol{X}, S) \mid S = s) \overset{d}{=} (f(\boldsymbol{X}, S) \mid S = s') \;\forall s, s' \in [K]\right\} = \mathcal{U}(f^*).$$

Moreover, the distribution of the minimizer of the problem on the l.h.s. *is given by*

$$\operatorname*{arg\,min}_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{K} w_s \mathsf{W}_2^2 \left(\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S{=}s), \nu\right).$$

This result is important for two reasons: it puts a measure of performance the risk $\mathcal{R}$, and a measure of unfairness $\mathcal{U}$ on the same scale. Moreover it expresses a deep connection between the Wasserstein-2 barycenter problem and the fairness-aware learning problem squared risk under Demographic Parity.

We note that the use of optimal transport tools in the study of fairness is relatively recent. Initially, contributions in this direction were mainly dealing with the problem of binary classification (Gordaliza et al. 2019; Jiang et al. 2019). Later on, the tools of the optimal transport theory migrated to the setup of fair regression (Chiappa et al. 2020; Chzhen et al. 2020c; Le Gouic, Loubes, and Rigollet 2020). As we will see, we made extensive use of the Wasserstein-2 metric to derive more results in the fair regression setup.

2.4 Contributions

2.4.1 Risk-fairness trade-off in the regression setup

Chapter 7 introduces a theoretical framework for rigorous analysis of regression problems under fairness requirements. Within this framework we precisely quantify the risk-fairness trade-off and derive general lower bound for learning under the Demographic Parity constraint.

We study the regression problem when a sensitive attribute is available. The statistician observes triplets $(\boldsymbol{X}_1, S_1, Y_1), \ldots, (\boldsymbol{X}_n, S_n, Y_n) \in \mathbb{R}^p \times [K] \times \mathbb{R}$, which are connected by the following regression-type relation

$$Y_i = f^*(\boldsymbol{X}_i, S_i) + \xi_i, \qquad i \in [n], \tag{2.1}$$

where $\xi_i \in \mathbb{R}$ is a centered random variable and $f^* : \mathbb{R}^p \times [K] \to \mathbb{R}$ is the regression function. Here for each $i \in [n]$, $\boldsymbol{X}_i$ is a feature vector taking values in $\mathbb{R}^p$, S_i is a sensitive attribute taking values in $[K]$, and Y_i is a real-valued dependent variable. We define the risk of a prediction function f via the $\mathbb{L}_2$ distance to the regression function f^* as

$$\mathcal{R}(f) := \|f - f^*\|_2^2 := \sum_{s=1}^{K} w_s \mathbb{E}\left[(f(\boldsymbol{X}, S) - f^*(\boldsymbol{X}, S))^2 \mid S = s\right], \qquad \textbf{(Risk measure)}$$

where $\mathbb{E}[\cdot \mid S{=}s]$ is the expectation *w.r.t.* the distribution of the features $\boldsymbol{X}$ in the group $S = s$ and $\boldsymbol{w} = (w_1, \ldots, w_K)^\top \in \Delta^{K-1}$ is a probability vector, which weights the group-wise risks.

For any $s \in [K]$ define ν_s^* as $\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S{=}s)$ – the distribution of the optimal prediction inside the group $S = s$. Throughout this work we make the following assumption on those measures, which is, for instance, satisfied in linear regression with Gaussian design.

Assumption 2.4.1. *The measures* $\{\nu_s^*\}_{s\in[K]}$ *are non-atomic and have finite second moments.*

As we have seen, exact DP is not necessarily desirable in practice and it is common in the literature to consider *relaxations* of this constraint. In this work we introduce the α-*Relative Improvement* (α-RI) constraint – a novel DP relaxation based on our unfairness measure.

Following the works of (Chzhen et al. 2020c; Le Gouic, Loubes, and Rigollet 2020) which linked the problem of regression under (exact) Demographic Parity constraint to a Wasserstein barycenter problem, we propose a new measure of unfairness

$$\mathcal{U}(f) \coloneqq \min_{\nu\in\mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{K} w_s \mathsf{W}_2^2(\mathrm{Law}(f(\boldsymbol{X}, S) \mid S{=}s), \nu),$$

which can be used to relax the Demographic Parity constraint: $\mathcal{U}(f)$ is equal to zero if and only if the distributions of the prediction are the same across all sensitive feature groups, *i.e.* if the predictor f satisfies the Demographic Parity constraint. Otherwise it is positive and quantifies the violation of this constraint. Equipped with this unfairness measure, we introduce a collection $\{f_\alpha^*\}_{\alpha\in[0,1]}$ of *oracle* α-*relative improvement* (α-RI) indexed by the parameter α as

$$f_\alpha^* \in \arg\min \left\{ \sum_{s=1}^{K} w_s \mathbb{E}\left[(f(\boldsymbol{X}, S) - f^*(\boldsymbol{X}, S))^2 \mid S = s\right] : \mathcal{U}(f) \le \alpha \mathcal{U}(f^*) \right\}.$$

Note that we consider relative unfairness, *w.r.t.* the Bayes predictor, to make the parameter α more interpretable. For $\alpha = 0$ the predictor f_0^* corresponds to the optimal fair predictor in the sense of DP studied by (Chzhen et al. 2020c; Le Gouic, Loubes, and Rigollet 2020) while for $\alpha = 1$ the corresponding predictor f_1^* coincides with the regression function f^*. Those two extreme cases have been previously studied but, up to our knowledge, nothing is known about those "partially fair" predictors which correspond to $\alpha \in (0, 1)$. Importantly, the fairness requirement is stated relatively to the unfairness of the regression function f^*, which allows to make a more informed choice of α. Our study of the family $\{f_\alpha^*\}_{\alpha\in[0,1]}$ serves as a basis for our statistical framework and analysis. It also reveals the intrinsic interplay of the fairness constraint with the risk measure.

Formally, for a fixed $\alpha \in [0, 1]$, the goal of a statistician in our framework is to build an estimator $\widehat{f}$ using data, which enjoys two guarantees (with high probability)

$$\alpha\textbf{-RI guarantee:} \quad \mathcal{U}(\widehat{f}) \le \alpha \mathcal{U}(f^*) \qquad \text{and} \qquad \textbf{Risk guarantee:} \quad \mathcal{R}(\widehat{f}) \le r_{n,\alpha,f^*}.$$

The former ensures that $\widehat{f}$ satisfies the α-RI constraint. In the latter guarantee we seek the sequence r_{n,α,f^*} being as small as possible in order to quantify *two effects*: the introduction of the α-RI *fairness constraint* and the *statistical estimation*. We note that r_{n,α,f^*} depends on the sample size n, the fairness parameter α, as well as the regression function f^* to be estimated, we clarify the reason for this dependency later in the text.

The contributions of this work can be roughly split into three interconnected groups:

1. We provide a theoretical study of the family of oracle α-RI $\{f_\alpha^*\}_{\alpha\in[0,1]}$ from which we derive a precise quantification of the risk-fairness trade-off on the population level as illustrated in Figure 2.2.

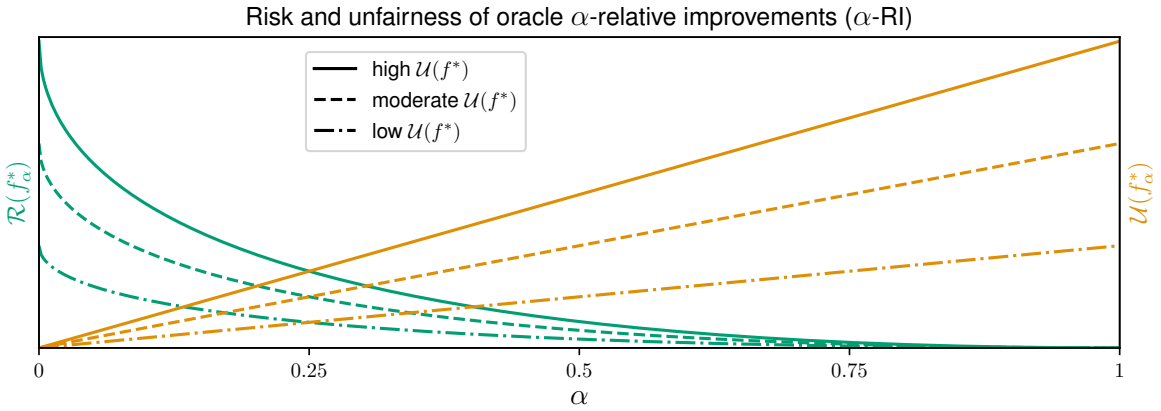


Figure 2.2: Risk $\mathcal{R}$ and unfairness $\mathcal{U}$ of α-RI oracles $\{f^*_\alpha\}_{\alpha\in[0,1]}$. Green curves (decreasing, convex) correspond to the risk, while orange curves (increasing, linear) correspond to the unfairness. Each pair of curves (solid, dashed, dashed dotted) corresponds to three regimes: high, moderate, and low unfairness of the regression function f^* respectively.

2. In order to quantify the *statistical* price of fairness, we introduce a minimax statistical framework and derive a general problem-dependent minimax lower bound for the problem of regression under the α-RI constraint. In particular, we show in Theorem 7.5.3 that *any estimator* $\widehat{f}$ *satisfying the* α*-RI* constraint with high probability must incur

$$\mathcal{R}(\widehat{f}) \geq \delta_n \vee (1-\sqrt{\alpha})^2\mathcal{U}(f^*),$$

where δ_n is the rate one would obtain *without* restricting the set of possible estimators.

3. In order to demonstrate that the general problem-dependent lower bound we derived does indeed yield minimax optimal rates, we derive such rates for the statistical model of linear regression with systematic group-dependent bias and Gaussian design under the α-RI constraint.

2.4.2 Demographic Parity without Disparate Treatment in the regression setting

As we have seen in Section 2.1, the sensitive feature might be inaccessible to the statistician for legal reasons and one may have to build fair regressor which does not produce Disparate Treatment (Definition 2.1.1). As discussed earlier, Lipton, McAuley, and Chouldechova (2018) provided an insightful study of this issue for classification problems. However, very little is known in the regression setting about the predictions which avoid Disparate Treatment and achieve Demographic Parity, even in the infinite sample regime. Actually, even the existence of non-trivial regression prediction strategies satisfying the two constraints is unclear.

In Chapter 8, we make progress towards the mathematical understanding of the latter problem. We make the following contributions: we propose a large family of prediction functions which achieve Demographic Parity without producing Disparate Treatment (not taking as input the sensitive feature); we identify a specific function within this class which additionally equalizes

the group-wise risks. To the best of our knowledge this is the first explicit construction of a non-constant predictor simultaneously satisfying those three fairness constraints. Even though the proposed prediction rule achieves several desirable formal group-fairness notions, we argue that this prediction is not suitable for real-world scenarios. We show on simple scenarios that, even though those predictors satisfy desirable fairness constraints, they may violate an intuitive conception of fairness, namely that some kind of group-wise order on the individuals should be preserved by fair procedures. In contrast, a prediction that is allowed to produce Disparate Treatment can alleviate these drawbacks. In the context of binary classification, similar conclusions were reached by Lipton, McAuley, and Chouldechova (2018).

2.4.3 Fair classification with abstention

In Chapter 9 we come back to the classification setting and introduce the possibility for decision rules to "reject" a fixed proportion of predictions (*i.e.* abstains from giving certain predictions), as it is done in usual classification with abstention/rejection procedures (Lei 2014b). In this setting a classifier is a mapping $g : \mathbb{R}^d \times [K] \to \{0, 1, r\}$. That is, any classifier g is able to provide a prediction in $\{0, 1\}$, or to abstain from prediction by outputting r. Our hope is that such an abstention can potentially overcome the usual risk-fairness trade-off.

More precisely, Chapter 9 combines and extends previous results in abstention framework with recent results on fair binary classification. Namely, similarly to Denis and Hebiri (2020), we aim at minimizing misclassification risk under a control over *group-wise* reject rates. As we would like to avoid disparate impact, we explicitly add the Demographic Parity constraint in our framework.

Formally, given reject rates $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top \in [0, 1]^K$ over the K sensitive groups, our goal is to find a solution of the problem

$$\begin{aligned} &\min_{g:\mathbb{R}^d\times[K]\to\{0,1,r\}} \mathbb{P}(Y \neq g(\boldsymbol{X}, S) \mid g(\boldsymbol{X}, S) \neq r) \\ &\text{s.t. } ,\forall s \in [K], \begin{cases} \mathrm{NAb}_s(g) = \alpha_s \\ \mathrm{PT}_s(g) = \mathrm{PT}(g) \end{cases} , \end{aligned} \tag{DPWA}$$

where we defined

$$\begin{aligned} \mathrm{NAb}_s(g) &:= \mathbb{P}(g(\boldsymbol{X}, S) \neq r \mid S = s), \\ \mathrm{PT}_s(g) &:= \mathbb{P}(g(\boldsymbol{X}, S) = 1 \mid S = s, g(\boldsymbol{X}, S) \neq r), \\ \mathrm{PT}(g) &:= \mathbb{P}(g(\boldsymbol{X}, S) = 1 \mid g(\boldsymbol{X}, S) \neq r)). \end{aligned}$$

Abstention framework has not yet received a lot of attention in the context of fair learning. Notable exceptions are work of Madras, Pitassi, and Zemel (2018) and Jones et al. (2020). The latter demonstrates that an imprudent use of abstention might amplify potential disparities already present in the data. In particular, they show that in the framework of prediction without disparate treatment (Zafar et al. 2017) the use of the same rejection threshold across sensitive groups might result in a large group-wise risks disparities. As a potential remedy, our work offers a theoretically grounded way to enforce fairness constraints as well as a desired group-dependent reject rates. The idea of relying on a reject mechanism to enforce fairness has only been explored once, in Madras, Pitassi, and Zemel (2018). The authors introduce

"learning to defer" framework – an extension of classification with abstention – where the cost of rejection is allowed to depend on the prediction of an external decision-maker (*e.g.,* a human expert). The authors argue that by making the automated model aware of the potential biases and weaknesses of the external decision-maker, it can globally optimize for accuracy and fairness. The authors enforce Equalized Odds (Hardt, Price, and Srebro 2016) through regularization of the risk and thus cannot control explicitly the reject rate, which might potentially lead to a huge external decision-maker costs. While the authors provide empirical evidences of their claims, theoretical justification of their results remains open.

Our work offers a completely theory-driven way to enforce both fairness and rejection constraints while optimizing for accuracy, leading to a computationally efficient post-processing algorithm. We derive in Theorem 9.3.2 the *optimal form of a reject classifier*, which minimizes the misclassification risk under the discussed constraints. Our explicit characterization of the optimal reject classifier provides a better understating of the interplay between, on one side, the fairness and rejection constraints and, on the other side, the accuracy. We propose a data-driven post-processing algorithm which enjoys generic plug-and-play *finite sample guarantees* regarding fairness and risk. An appealing feature of our post-processing algorithm is that it can be used on top of *any* pre-trained classifier, thus avoiding the – potentially high – cost of re-fitting a classifier from scratch. From numerical perspective, the proposed method reduces to a solution of a sparse linear program, allowing us to leverage efficient LP solvers. Numerical experiments validate our theoretical result demonstrating that the proposed method successfully enforces fairness and rejection constraints in practice, while achieving a high level of accuracy.

CHAPTER 3

Une introduction à l'apprentissage équitable[1]

Les algorithmes entraînés sur nos données personnelles prennent des décisions cruciales qui influencent notre vie au quotidien. Des études récentes montrent qu'une utilisation naïve de ces algorithmes dans des domaines sensibles peut conduire à des décisions injustes et discriminantes, héritant souvent ou même amplifiant les biais présents dans les données (Barocas and Selbst 2016). Prenons un exemple tiré d'une enquête récente sur le sujet (Barocas, Hardt, and Narayanan 2017) : "Amazon utilise un système basé sur des données pour déterminer les quartiers dans lesquels proposer une livraison gratuite le jour même. Une étude de 2016 a révélé de fortes disparités dans la composition démographique de ces quartiers : dans de nombreuses villes américaines, les résidents blancs étaient plus de deux fois plus susceptibles que les résidents noirs de vivre dans l'un des quartiers admissibles." Cet exemple met en lumière une tendance inquiétante, à savoir que les algorithmes basés sur des données peuvent conduire à des décisions injustes dans des domaines beaucoup plus sensibles tels que, par exemple, les décisions de justice[2], les admissions dans les écoles/universités, les approbations de prêts bancaires, etc. Il est donc de plus en plus nécessaire de s'assurer que les algorithmes utilisés en pratique ne contredisent pas les fondements moraux et juridiques de nos sociétés tout en restant utiles. Cette question se situe à l'intersection de plusieurs domaines tels que la philosophie politique, la sociologie, les statistiques, l'informatique, etc. Ainsi, sa pleine compréhension doit en fin de compte impliquer des discussions interdisciplinaires. Suivant la sagesse de l'expression latine "*Sutor, ne ultra crepidam*"[3], nous nous concentrerons principalement sur les aspects de cette question qui se situent dans notre domaine d'expertise. En particulier, nous nous appuierons sur les avancées récentes de la théorie de l'apprentissage statistique qui permettent d'aborder certains aspects de ce problème dans un cadre statistique rigoureux. Nous renvoyons le lecteur à Mehrabi et al. (2019) and Barocas, Hardt, and Narayanan (2019)

[1] Ce chapitre est une traduction en français (par Google Traductions) du chapitre précédent.

[2] Voir L'étude de Propublica sur le logiciel d'évaluation des risques COMPAS qui est utilisé dans les tribunaux américains pour évaluer la probabilité qu'un criminel récidive.

[3] L'expression signifie littéralement "Cordonnier, pas plus haut que la chaussure" et était utilisée pour avertir les gens d'éviter de porter un jugement au-delà de leur expertise (Wikipedia).

pour une introduction générale sur l'équité algorithmique et à Oneto and Chiappa (2020) and Barrio, Gordaliza, and Loubes (2020) pour une revue des avancées théoriques les plus récentes.

Dans ce chapitre du manuscrit, nous nous concentrerons sur le problème de l'apprentissage équitable (Menon and Williamson 2018a). Encore une fois, nous n'avons pas la prétention ni l'objectif de bâtir une théorie générale de l'équité, ni de débattre de ce qui est juste et de ce qui ne l'est pas. Il s'agit là d'un débat politique, qui doit être mené au niveau de la société. Notre objectif sera beaucoup plus modeste. Comme présenté dans la section suivante, inspirés par des concepts juridiques ou philosophiques reconnus, des informaticiens et mathématiciens ont proposé plusieurs formalisations mathématiques de la discrimination des règles de décision. Notre objectif, en tant que statisticiens, sera d'évaluer, compte tenu d'une mesure de performance et de critère(s) d'équité, quel est le meilleur résultat que l'on puisse espérer obtenir dans une perspective d'apprentissage en matière d'équité et de performance. En particulier, nous verrons qu'un compromis naturel apparaît entre la satisfaction des contraintes d'équité et l'obtention de bonnes performances (prédictives). Notre approche des questions liées à l'équité ne discute pas de la pertinence d'un choix donné (tel que le choix d'un critère d'équité et d'une mesure du risque), qui est finalement laissé au décideur, mais permet de mieux comprendre les conséquences de ce choix. Une telle approche est en phase avec ce que Weber (1992) a défini comme l'un des objectifs de la science en général – non pas un substitut au jugement de l'homme mais un outil pour une prise de décision éclairée.

Contents

3.1 Formalisation du problème et définitions

Dans ce qui suit, nous nous plaçons dans le cadre de l'apprentissage supervisé : un statisticien, qui reçoit des couples de variables de caractéristiques et d'étiquettes, cherche à exprimer l'étiquette en fonction des variables de caractéristiques afin de prédire correctement l'étiquette associée à de nouvelles variables de caractéristiques. Le cadre d'apprentissage (supervisé) équitable diffère légèrement du cadre habituel en ce sens que nous ne traitons pas toutes les caractéristiques de la même manière. En particulier, nous distinguons deux types de caractéristiques : un ensemble de caractéristiques (nominalement) *insensibles* $\mathcal{X}$ et un ensemble de caractéristiques *sensibles* $\mathcal{S}$. Le lecteur peut considérer ces dernières comme étant, par exemple, le sexe, l'origine ethnique et/ou l'âge. Il est important de noter que l'ensemble des caractéristiques sensibles $\mathcal{S}$ contient les caractéristiques contre lesquelles nous voulons contrôler les discriminations potentielles. L'ensemble $\mathcal{S}$ sera généralement un ensemble fini dans ce manuscrit. Nous tenons à souligner que dans le cadre considéré, le choix des attributs sensibles appartient au décideur, potentiellement incité par des motifs juridiques ou éthiques. Ce n'est pas la tâche du statisticien de déterminer quelle caractéristique doit être considérée comme sensible.

Formellement, le statisticien observe des copies indépendantes d'un triplet $(\boldsymbol{X}, S, Y) \sim P$ où $\boldsymbol{X}$ est le vecteur de caractéristiques, Y la variable d'étiquette et S est une caractéristique

sensible (*par exemple* sexe, ethnicité ou âge). Ces variables aléatoires prennent leurs valeurs dans les ensembles $\mathcal{X}, \mathcal{Y}$, et $\mathcal{S}$, respectivement. Nous considérerons des prédicteurs de la forme $f : \mathcal{Z} \to \mathcal{Y}$ où, suivant la notation de Donini et al. (2018), l'ensemble $\mathcal{Z}$ est soit l'ensemble des caractéristiques non sensibles $\mathcal{X}$, soit l'ensemble des caractéristiques $\mathcal{X} \times \mathcal{S}$, selon que le statisticien est autorisé ou non à accéder à l'attribut sensible pour la prédiction. De manière analogue, nous définissons $\boldsymbol{Z}$ comme $\boldsymbol{X}$ ou $(\boldsymbol{X}, S)$ selon le type de prédiction en question. Notons que chaque prédicteur f induit des distributions par groupe des résultats prédits $\text{Law}(f(\boldsymbol{Z}) \mid S{=}s)$ pour $s \in \mathcal{S}$. L'objectif général du statisticien est double : *maximiser les performances de prédiction* en minimisant un risque donné tout en *satisfaisant une ou plusieurs contraintes d'équité données*. Rentrons maintenant dans les définitions mathématiques de l'équité pour préciser cet objectif général.

Equité individuelle et équité collective. Fondamentalement, les définitions mathématiques de l'équité peuvent être divisées en deux groupes (Dwork et al. 2012) : *équité individuelle* et *équité de groupe*. La première notion reflète le principe selon lequel des individus similaires doivent être traités de manière similaire, ce qui se traduit par des contraintes de type Lipschitz sur les règles de prédiction possibles (souvent aléatoires). La seconde définit l'équité au niveau de la population via l'indépendance statistique (conditionnelle) d'une prédiction par rapport à un attribut sensible S (*e.g.,* sexe, ethnicité). L'idée générale des notions de *équité de groupe* peut être vue comme la limitation ou la diminution d'une éventuelle divergence entre les distributions par groupe des résultats prédits. Dans cette thèse, nous nous concentrerons sur la *équité de groupe*. Nous renvoyons le lecteur à l'article fondateur de Dwork et al. (2012) pour une introduction à l'équité individuelle et à Jung et al. (2019), Dwork, Ilvento, and Jagadeesan (2020), and Mukherjee et al. (2020) pour des exemples de travaux récents dans ce contexte.

Traitement différencié. Avant d'introduire les principales définitions de l'équité de groupe, définissons une première notion naturelle (et naïve) d'équité qui restreint essentiellement la classe des prédicteurs à ceux qui ne prennent pas en entrée l'attribut sensible S.

Definition 3.1.1 (Traitement différencié)**.** *Toute fonction $f : \mathcal{X} \to \mathcal{Y}$ qui ne peut pas recevoir l'attribut sensible S dans sa forme fonctionnelle ne produit pas de* Traitement différencié*.*

Gajane and Pechenizkiy (2017) font référence à cette définition comme *fairness through unawareness*, par opposition à *fairness through awareness*, introduite par Dwork et al. (2012). Ce dernier type de prédiction permet de construire un modèle distinct pour chaque attribut sensible, tandis que le premier oblige à fixer un modèle unique qui est ensuite appliqué à tous les groupes. Cette propriété peut être souhaitable pour des raisons juridiques et/ou de confidentialité évidentes : (Primus 2003; Barocas and Selbst 2016; Gajane and Pechenizkiy 2017; Lipton, McAuley, and Chouldechova 2018). Par exemple, en France, il est interdit d'utiliser l'appartenance ethnique comme attribut sensible pour tout traitement statistique [4] donc notamment pour la prédiction.

Remark 3.1.2. *Pour certains auteurs, la disparité de traitement englobe un phénomène plus large. Par exemple, Lipton, McAuley, and Chouldechova (2018) considère que la disparité de*

[4] Voir "Décision n° 2007-557 DC du 15 novembre 2007 du Conseil Constitutionnel, Loi relative à la maîtrise de l'immigration, à l'intégration et à l'asile" (lien)

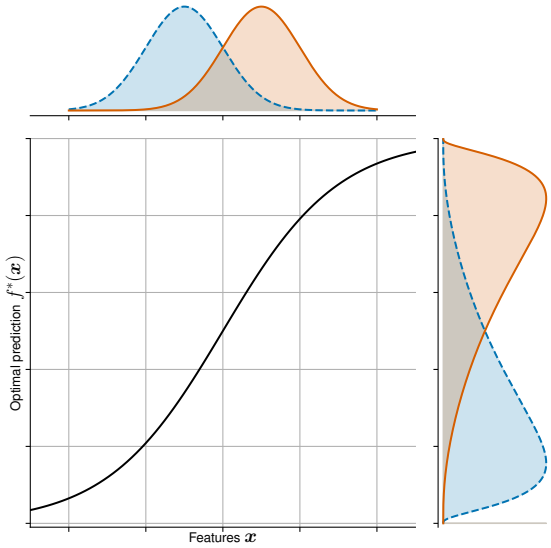


Figure 3.1: Un exemple simple d'un prédicteur (optimal) qui ne produit pas de traitement différencié mais dont les prédictions diffèrent selon les groupes sensibles. L'attribut sensible S peut prendre deux valeurs $S = 1$ et $S = 2$ avec une probabilité égale. La caractéristique X est distribuée selon un modèle de mélange gaussien : $X|S = 1 \sim \mathcal{N}(-1, 1), X|S = 2 \sim \mathcal{N}(1, 1)$. Les distributions des caractéristiques par groupe sont représentées en haut en bleu pointillé $(S = 1)$ et en orange plein $(S = 2)$. L'étiquette Y est obtenue par $Y = f^*(X)$ où $f^*(x) = \frac{1}{1+e^{-x}}$. La distribution des prédictions de f^* entre les groupes est représentée sur la droite en bleu pointillé $(S = 1)$ et en orange plein $(S = 2)$.

traitement concerne la discrimination intentionnelle qui inclut les décisions explicitement basées sur des caractéristiques protégées, comme dans la Définition 3.1.1, mais aussi la discrimination intentionnelle via des variables de substitution (par exemple les tests d'alphabétisation pour l'éligibilité au vote). La formalisation mathématique de ce phénomène n'étant pas évidente, nous préférons nous en tenir à la définition que nous avons introduite.

Il est important de noter que l'absence de traitement différencié ne garantit pas que la prédiction soit *statistiquement indépendante* de l'attribut sensible S en raison des corrélations et, plus généralement, des dépendances entre l'attribut sensible S et le vecteur de caractéristiques $\boldsymbol{X}$ (Pedreshi, Ruggieri, and Turini 2008). En effet, considérons le prédicteur optimal de Bayes $\boldsymbol{x} \mapsto f^*(\boldsymbol{x})$ défini comme suit

$$f^*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}].$$

Par définition, il ne prend pas en entrée l'attribut sensible et atteint le risque quadratique le plus faible possible parmi les prédictions évitant le traitement différencié. Pourtant, le prédicteur f^* peut encore favoriser la disparité entre les groupes sensibles si les distributions des caractéristiques $\boldsymbol{X}$ diffèrent entre les groupes. Un exemple d'un tel scénario est donné dans la Figure 3.1 : même si le prédicteur ne produit pas de DT (il est le même pour les deux groupes), la distribution des prédictions pour le premier groupe diffère significativement de celle du second groupe en raison des distributions des caractéristiques.

De plus, dans le cadre de la classification, Lipton, McAuley, and Chouldechova (2018) a montré qu'éviter le traitement différencié (TD) n'est pas nécessairement souhaitable, même lorsqu'il est combiné à d'autres critères d'équité comme dans les processus d'apprentissage différencié, une classe de procédures d'apprentissage qui peuvent accéder à un attribut sensible pendant la formation mais ne peuvent pas l'utiliser pour la prédiction. Ils prouvent qu'une règle de décision qui ne produit pas de TD ne peut pas atteindre une meilleure précision qu'une règle optimale qui utilise cette information ; ce qui les amène à la conjecture que toute règle qui évite le TD produit un compromis sous-optimal entre l'équité et la performance. En outre, les auteurs fournissent des preuves empiriques que l'évitement du TD conduit à une disparité de traitement indirecte, via des variables de substitution, ou à une discrimination au sein de la classe. Plusieurs questions découlent de ces observations. Est-il possible de fournir des justifications théoriques satisfaisantes de leurs résultats ? Ces phénomènes sont-ils limités au cadre de la classification ? Dans quelle mesure peut-on espérer être juste et précis en permettant de traiter des sous-groupes différemment ? Les chapitres 7 et 8 fourniront une réponse théorique provisoire à ces questions dans le cadre de la régression.

Disparate Impact. Comme nous l'avons vu, étant donné un prédicteur f, les distributions de prédiction $(\text{Law}(f(\boldsymbol{Z})|S = s))_{s\in\mathcal{S}}$ peuvent différer entre les sous-groupes sensibles et, fait important, ce phénomène peut se produire *involontairement*, *i.e.,* non pas à cause d'un statisticien pernicieux, mais comme *corollaire* du processus d'apprentissage. Le décideur peut estimer qu'une telle différence n'est pas souhaitable ; par exemple, si un prédicteur f est utilisé pour déterminer les salaires des employés d'une entreprise, le conseil d'administration peut vouloir (ou doit) choisir un prédicteur qui donne des distributions de salaires similaires entre les sous-groupes (par exemple les hommes et les femmes). Par conséquent, la plupart des notions d'équité se concentrent sur le contrôle de l'écart entre les distributions des prédictions de chaque groupe. Nous allons maintenant fournir certaines des définitions d'équité les plus populaires concernant les distributions des prédictions. Puisque la plupart de la littérature se concentre sur le problème de la classification équitable (voir Calders, Kamiran, and Pechenizkiy (2009) et les références précédentes), les définitions d'équité suivantes ont été initialement données dans ce cadre. Pour plus de clarté, nous fournirons ces définitions dans le cadre de la classification binaire avec attribut sensible à valeur binaire. Nous expliquerons, si nécessaire, comment étendre ces définitions aux problèmes de régression et aux attributs sensibles non binaires. Ces extensions seront basées sur la propriété d'indépendance entre les variables aléatoires. Nous utiliserons la notation $A \perp\!\!\!\perp B$ pour exprimer l'indépendance entre les variables aléatoires A et B.

Tout d'abord, il se peut que le risque d'un prédicteur soit faible en moyenne sur l'ensemble de la population, mais que l'écart de risque entre les groupes soit élevé. Une telle situation pourrait être considérée comme discriminatoire pour les groupes à haut niveau de risque. Pour éviter de tels problèmes, nous introduisons une première définition de l'équité dans la classification binaire, qui demande l'égalité des risques par groupe.

Definition 3.1.3 (Égalité des risques par groupe)**.** *Un classificateur* $f : \mathcal{Z} \to \{0, 1\}$ *atteint* l'galité des risques par groupe *par rapport à la distribution* $\mathbb{P}$ *de* $(\boldsymbol{X}, S, Y)$ *si*

$$\mathbb{P}(f(\boldsymbol{Z}) \neq Y \mid S = 0) = \mathbb{P}(f(\boldsymbol{Z}) \neq Y \mid S = 1).$$

Buolamwini and Gebru (2018) soutient qu'elle correspond à l'exigence que tous les groupes

reçoivent un bon service et Zafar et al. (2017) défend cette notion pour éviter les mauvais traitements disparates. Notons que cette définition se généralise immédiatement à toute notion de risque en espérance. Une formulation relaxée de cette notion d'équité a été considérée dans le contexte de la régression par Agarwal, Dudík, and Wu (2019).

Ensuite, nous présentons la principale notion d'équité sur laquelle cette thèse va se concentrer. Elle a été formellement introduite par Calders, Kamiran, and Pechenizkiy (2009) et demande essentiellement que la prédiction soit (statistiquement) indépendante de l'attribut sensible.

Definition 3.1.4 (Parité démographique)**.** *Un classificateur* $f : \mathcal{Z} \to \{0, 1\}$ *atteint la parité démographique par rapport à la distribution* $\mathbb{P}$ *de* $(\boldsymbol{X}, S, Y)$ *si*

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1).$$

En termes simples, cette définition attribue des chances égales de décision positive aux deux groupes. Elle agit comme une incitation à promouvoir la diversité par le biais de la discrimination positive (Mouzannar, Ohannessian, and Srebro 2019). Nous pouvons exprimer de manière équivalente la parité démographique comme l'indépendance (probabiliste) entre la distribution de la prédiction $f(\boldsymbol{Z})$ et celle de l'attribut sensible S, que nous désignons par

$$f(\boldsymbol{Z}) \perp\!\!\!\perp S. \tag{DP}$$

Une telle caractérisation généralise immédiatement la définition initiale à tout problème d'apprentissage supervisé et à tout attribut sensible.

Comme le fait valoir Hardt, Price, and Srebro (2016), si elle est déployée dans la réalité, une telle notion peut avoir des effets plus négatifs que positifs dans certains cas. Un exemple qu'ils donnent est lié à l'octroi de crédit bancaire, où $Y = 1$ signifie qu'un individu $(\boldsymbol{X}, S)$ est capable de rembourser le prêt et $f(\boldsymbol{Z}) = 1$ signifie que la banque approuve le crédit. Dans le cas où les capacités de remboursement de deux groupes sont radicalement différentes, le fait d'offrir des chances égales d'obtenir un crédit sans tenir compte de la capacité de remboursement Y met les individus moins privilégiés sur la voie de la défaillance. Pour contourner le problème ci-dessus, Hardt, Price, and Srebro (2016) a proposé les deux définitions suivantes.

Definition 3.1.5 (Egalité des erreurs)**.** *Un classificateur* $f : \mathcal{Z} \to \{0, 1\}$ *atteint* l'égalité des erreurs *par rapport à la distribution* $\mathbb{P}$ *de* $(\boldsymbol{X}, S, Y)$ *si*

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0, Y = y) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1, Y = y), \quad \forall\, y \in \{0, 1\}.$$

Cette notion d'équité demande la prédiction $f : \mathcal{Z} \to \{0, 1\}$ pour égaliser les taux de vrais positifs et de vrais négatifs dans les deux groupes. On voit immédiatement que cette définition peut être exprimée sous une forme plus générale comme suit

$$(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y. \tag{EOdd}$$

Cela signifie que, étant donné la vraie étiquette d'un individu, connaître la valeur de l'attribut sensible n'apporte aucune information sur la distribution de la prédiction. Typiquement, l'égalisation des vrais négatifs n'est pas nécessaire en pratique, puisque $f(\boldsymbol{Z}) = 1$ est interprété comme une décision positive et qu'il peut être naturel de se concentrer sur de telles décisions. De cette façon, nous arrivons à la définition de l'égalité des chances.

Definition 3.1.6 (Egalité des chances)**.** *Un classificateur* $f : \mathcal{Z} \to \{0, 1\}$ *atteint* l'égalité des chances *par rapport à la distribution* $\mathbb{P}$ *de* $(\boldsymbol{X}, S, Y)$ *si*

$$\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0, Y = 1) = \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1, Y = 1).$$

L'égalité des chances est moins contraignante que l'égalité des erreurs car elle demande simplement que les taux de vrais positifs soient les mêmes pour tous les groupes. Dans l'exemple du palier de crédit, l'égalité des chances signifie que parmi les clients qui sont en mesure de rembourser leur prêt, la proportion de prêts attribués doit être la même dans tous les groupes. Il est facile de voir que cette définition est équivalente à la contrainte d'indépendance conditionnelle plus générale

$$(f(\boldsymbol{Z}) \perp\!\!\!\perp S) \mid Y = 1. \qquad \textbf{(EOpp)}$$

Notons cependant que l'extension de cette notion au cadre de la régression n'est pas du tout évidente. En effet, la définition de l'égalité des chances repose sur le fait que $Y = 1$ peut être interprété comme une décision positive (*e.g.,* attribue un prêt) et il n'existe pas d'équivalent immédiat d'une "décision positive" dans le cadre de la régression. La question de trouver une transposition de la notion d'égalité des chances en dehors du cadre de la classification reste ouverte.

Toutes les définitions de l'équité présentées jusqu'à présent sont exprimées comme la probabilité conditionnelle d'une prédiction positive $f(\boldsymbol{Z}) = 1$ étant donné la valeur sensible de l'attribut S et éventuellement la véritable étiquette Y. Contrairement aux définitions précédentes, la suivante traite des probabilités conditionnelles par rapport à l'événement d'une prédiction positive $(f(\boldsymbol{Z}) = 1)$.

Definition 3.1.7 (Test d'équité)**.** *Un classificateur* $f : \mathcal{Z} \to \{0, 1\}$ *satisfait au* test d'équité *par rapport à la distribution* $\mathbb{P}$ *de* $(\boldsymbol{X}, S, Y)$ *si*

$$\mathbb{P}(Y = 1 \mid S = 0, f(\boldsymbol{Z}) = 1) = \mathbb{P}(Y = 1 \mid S = 1, f(\boldsymbol{Z}) = 1).$$

Comme pour les définitions précédentes, la dernière peut naturellement être exprimée comme la condition d'indépendance conditionnelle

$$(Y = 1 \perp\!\!\!\perp S) \mid f(\boldsymbol{Z}) = 1. \qquad \textbf{(Test fairness)}$$

L'équité du test est également appelée *parité prédictive*. Elle impose l'égalité entre les groupes sensibles des taux de résultats positifs $Y = 1$ parmi ceux qui ont reçu une prédiction positive $f(\boldsymbol{Z}) = 1$. Elle est étroitement liée au concept de calibrage (Barocas, Hardt, and Narayanan 2017, Chapitre 2).

Nous avons donné cinq définitions de l'équité qui englobent différentes conceptions de la discrimination. En considérant ces définitions comme des contraintes, nous pouvons maintenant clarifier l'objectif de l'apprentissage sensible à l'équité : minimiser un risque donné sur la classe de prédicteurs qui satisfont (un sous-ensemble de) ces définitions. Notons que toutes les contraintes sont *distribution-dépendantes* c'est-à-dire qu'elles dépendent de la distribution jointe $(\boldsymbol{X}, S, Y)$. Par conséquent, les problèmes d'optimisation sous contrainte qui en résultent sont très différents des problèmes habituels sous contrainte de forme dans lesquels des contraintes

indépendantes de la distribution sont imposées aux prédicteurs, comme les contraintes de régularité.

Exposons maintenant brièvement une étude de cas éclairante et bien connue sur l'équité afin de mettre en évidence certains problèmes que nous pouvons peut rencontrer. COMPAS (qui signifie *Correctional Offender Management Profiling for Alternative Sanctions*) est un logiciel d'évaluation des risques développé par Northpointe Group., une société de conseil en technologie et en gestion, qui est utilisé par les tribunaux américains pour prédire le risque de récidive. Une étude réalisée par ProPublica (Angwin et al. 2016) a montré que COMPAS donne lieu à un écart important en ce qui concerne les taux de faux positifs et de faux négatifs entre les groupes ethniques, et qu'il ne satisfait donc pas au principe d'égalité des erreurs. Cependant, la réfutation de Northpointe (Dieterich, Mendoza, and Brennan 2016) a affirmé que COMPAS satisfait (une notion généralisée de) la parité prédictive. Cet exemple souligne que le choix du critère d'équité est crucial lors de la discussion du comportement discriminatoire d'une règle de décision. En outre, il soulève la question de savoir si l'on doit s'en tenir à un critère d'équité particulier ou essayer de satisfaire plusieurs critères simultanément. Bien sûr, dans l'idéal (et naïvement), on aimerait obtenir un prédicteur qui satisfasse autant de contraintes d'équité que possible. Cependant, cela n'est pas nécessairement souhaitable puisque la classe de prédicteurs peut être petite ou même vide. En effet, Chouldechova (2017) a montré un résultat d'impossibilité qui stipule qu'à moins que $\mathbb{P}(Y = 1|S = s)$ soit le même dans tous les groupes, aucun classifieur $f : \mathcal{Z} \to \{0, 1\}$ ne peut satisfaire simultanément l'équité du test et l'égalité des erreurs. Quelles leçons pouvons-nous tirer de tout cela ? Tout d'abord, il faut effectivement faire attention aux critères d'équité choisis pour une tâche donnée ; ensuite, cela pourrait indiquer que les notions d'équité que nous considérons sont trop rigides et qu'il serait intéressant d'explorer des moyens de relâcher ces contraintes. Ce sera le sujet de la prochaine section.

Un mot sur le biais dans les données et la causalité. Avant de passer à la section suivante, discutons d'un point important que nous jugeons important. Un lecteur attentif aura peut-être remarqué que toutes les définitions que nous avons données se concentrent sur la discrimination potentielle au niveau de la prédiction alors que rien n'a été dit sur les biais potentiels dans les données. Pourquoi avons-nous fait ce choix ? En bref, nous nous intéressons au résultat des algorithmes, et non à l'identification des biais dans les données, car les raisons pour lesquelles une procédure ou un algorithme peut avoir un comportement discriminatoire sont, la plupart du temps, loin d'être évidentes et potentiellement impossibles à dévoiler à moins que nous soyons prêts à faire des hypothèses fortes sur nos données. Par exemple, si les caractéristiques sont des images à haute dimension, il peut s'avérer extrêmement difficile de fournir ou de tester un modèle causal pour les données en raison de leur complexité (Bühlmann 2013). Néanmoins, nous aimerions mentionner une sous-branche complémentaire en pleine expansion de la littérature sur l'équité algorithmique qui tente d'incorporer le raisonnement causal dans l'apprentissage équitable : (see, *e.g.,* Kilbertus et al. 2017; Kusner et al. 2017; Loftus et al. 2018; Makhlouf, Zhioua, and Palamidessi 2020). Ils fournissent de nouvelles formalisations des notions d'équité et de nouvelles procédures pour surmonter les biais et discriminations potentiels en utilisant les mécanismes de l'apprentissage causal. Le cadre qui nous intéresse est celui dans lequel fournir un modèle causal pour le problème d'intérêt est trop difficile, trop long ou trop coûteux. Par conséquent, nous nous concentrons sur le contrôle du biais dans la prédiction, qui est quelque chose que nous pouvons observer et tester.

3.2 Relaxation et compromis

Dans la section précédente, nous avons fourni quelques moyens populaires de formaliser ce que signifie pour un prédicteur d'être équitable, c'est-à-dire d'éviter des discriminations particulières. Nous avons énoncé de manière générale le problème d'apprentissage équitable comme étant celui de la recherche d'un prédicteur f qui présente un faible risque $\mathcal{R}(f)$ tout en satisfaisant un ou plusieurs critères d'équité, tels que ceux que nous avons définis. En suivant ce paradigme, le problème d'apprentissage sensible à l'équité peut être exprimé comme un problème d'optimisation sous contrainte :

$$\arg\min_{f:\mathcal{Z}\to\mathcal{Y}} \{\mathcal{R}(f) : f \text{ est “équitable”} \},$$

où la contrainte pourrait être, par exemple, d'imposer que les prédicteurs satisfassent une ou plusieurs des définitions que nous avons introduites telles que les définitions 3.1.4 et 3.1.5.

Étant donné cette formulation générale du problème d'apprentissage sensible à l'équité, une question naturelle se pose : quel est l'impact de l'introduction de la contrainte d'équité sur le risque des meilleurs prédicteurs équitables ? Afin d'apporter une réponse pertinente, nous allons expliquer pourquoi un assouplissement des contraintes d'équité est nécessaire et proposer différentes manières d'y parvenir. Pour l'instant, il n'est pas évident de savoir pourquoi on veut assouplir quoi que ce soit ? Cela nous permettra de dériver un cadre général pour étudier le compromis entre performance et équité.

Relaxation des contraintes d'équité. Il y a quatre raisons principales pour lesquelles la relaxation est nécessaire. Premièrement, les définitions mathématiques de l'équité sont une transposition imparfaite d'idées qualitatives, il n'est donc pas forcément souhaitable de viser une satisfaction exacte des contraintes qui en résultent. Deuxièmement, deux sources d'aléa peuvent brouiller la satisfaction de nos définitions : nous avons introduit des définitions d'équité pour des prédicteurs fixes alors qu'en pratique, nous travaillerons avec des estimateurs *c'est-à-dire* des prédicteurs qui dépendent d'un nombre fini d'échantillons de la distribution conjointe $(\boldsymbol{X}, S, Y)$; en outre, les quantités d'intérêt (comme le taux de faux positifs par exemple) doivent également être estimées à partir d'échantillons finis. Ces incertitudes inévitables et les erreurs qui en résultent doivent être prises en compte dans les contraintes. Troisièmement, un changement drastique d'une politique de prédiction peut avoir un impact négatif : par exemple, si la tâche consiste à fixer les salaires dans une entreprise, les employés ne toléreront probablement pas un changement brusque (négatif) de leur salaire. Nous pourrions vouloir introduire l'équité de manière continue, d'où la nécessité d'un moyen d'interpoler en douceur entre les situations justes et injustes. Enfin, les définitions introduites sont trop rigides car elles ne permettent pas de comparer le comportement discriminatoire de deux prédicteurs qui ne satisfont pas la ou les contraintes d'équité : un prédicteur est considéré soit comme équitable, soit comme injuste. En ce qui concerne le risque, nous aimerions avoir une sorte de relation d'ordre sur les prédicteurs concernant leur équité.

D'où la nécessité de trouver un moyen d'assouplir les définitions initiales et de fournir des notions significatives d'iniquité, définie comme la violation des contraintes d'équité. Dans ce qui suit, nous nous concentrerons sur les relaxations de la contrainte de parité démographique car tous les travaux de cette thèse sont basés sur cette contrainte. Rappelons que la DP requiert l'égalité de deux quantités d'intérêt, $\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0)$ et $\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1)$.

Une première façon naturelle de relâcher les contraintes d'équité exacte est de demander que la différence des quantités d'intérêt soit petite. Dans le cas de la parité démographique, cela revient à

$$|\mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 0) - \mathbb{P}(f(\boldsymbol{Z}) = 1 \mid S = 1)| \leq \varepsilon,$$

pour un certain seuil $\varepsilon \geq 0$ prescrit. On peut également considérer d'autres types de relaxations, par exemple, multiplicatives au lieu d'additives. On peut également demander que les rapports de ces quantités soient proches de 1 ou trouver un autre critère basé sur toute autre combinaison des quantités d'intérêt jugée pertinente et/ou pratique. Notons que ces idées se généralisent facilement aux autres notions d'équité que nous avons introduites dans le cadre de la classification avec attribut sensible à valeur binaire, $\mathcal{S} \times \mathcal{Y} = \{0, 1\}^2$. Cependant, la manière de procéder pour d'autres paramètres tels que la régression ou l'attribut sensible général n'est pas claire.

Une autre façon d'assouplir les critères d'équité est liée aux définitions alternatives de l'indépendance (conditionnelle) que nous avons fournies. Comme nous l'avons vu, trois contraintes d'équité populaires peuvent être exprimées comme des conditions d'indépendance (conditionnelle) dépendant de la distribution conjointe du triplet $(f(\boldsymbol{Z}), Y, S)$. Afin de formaliser cette idée, toute métrique/divergence d sur l'espace des distributions de probabilité (telle que la distance de Kolmogorov-Smirnov, la distance de variation totale, la divergence de Kullback-Leibler, etc.) peut être utilisée pour mesurer l'écart entre les prédictions par groupe, comme suit

$$d\big(\operatorname{Law}(f(\boldsymbol{Z})|S = s), \operatorname{Law}(f(\boldsymbol{Z})|S = s')\big),$$

pour toute valeur s, s' $dans\mathcal{S}$ de l'attribut sensible. En particulier, si d satisfait à la propriété de séparation (à savoir que pour toute distribution de probabilité P et Q, $d(P, Q) = 0 \implies P = Q$), la parité démographique est satisfaite si et seulement si

$$d\big(\operatorname{Law}(f(\boldsymbol{Z})|S = s), \operatorname{Law}(f(\boldsymbol{Z})|S = s')\big) = 0, \quad \forall s, s' \in \mathcal{S}.$$

Une idée naturelle est de définir une fonctionnelle $\mathcal{U}$ sur l'ensemble des prédicteurs qui quantifie la violation de la contrainte DP en utilisant une métrique choisie et de déclarer une prédiction approximativement juste si cette fonctionnelle ne dépasse pas un seuil pré-spécifié par l'utilisateur. Ces dernières années, une grande variété de relaxations de ce type a été proposée : basée sur la corrélation (Baharlouei et al. 2019; Mary, Calauzènes, and El Karoui 2019; Komiyama et al. 2018) ; distance de Kolmogorov-Smirnov (Agarwal, Dudík, and Wu 2019) ; Information mutuelle (Steinberg et al. 2020; Steinberg, Reid, and O'Callaghan 2020) ; Distance de variation totale (Oneto, Donini, and Pontil 2019b; Oneto et al. 2019) ; Égalité des moyennes et correspondance des moments supérieurs (Raff, Sylvester, and Mills 2018; Fitzsimons et al. 2019; Calders et al. 2013; Berk et al. 2017; Olfat et al. 2020; Donini et al. 2018) ; Écart moyen maximal (Quadrianto and Sharmanska 2017; Madras et al. 2018) ; Distance de Wasserstein (Chiappa et al. 2020; Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c; Gordaliza et al. 2019).

Les relaxations les plus courantes de la contrainte de parité démographique sont basées sur la variation totale (TV) et les distances de Kolmogorov-Smirnov (KS) (Agarwal, Dudík, and Wu 2019; Oneto, Donini, and Pontil 2019a; Agarwal et al. 2018; Chzhen et al. 2020a). Il existe

plusieurs façons d'utiliser la TV ou la KS afin de construire une $\mathcal{U}$ fonctionnelle, qui quantifie la violation de la contrainte DP. Pour comparer ces mesures de divergence avec celle que nous introduisons dans notre travail, nous définissons $\mathcal{U}_{\mathrm{TV}}$ et $\mathcal{U}_{\mathrm{KS}}$ comme suit

$$\textbf{TV unfairness:} \qquad \mathcal{U}_{\mathrm{TV}}(f) := \sum_{s\in[K]} \mathrm{TV}\left(\mathrm{Law}(f(\boldsymbol{X}, S) \mid S = s), \mathrm{Law}(f(\boldsymbol{X}, S))\right),$$

$$\textbf{KS unfairness:} \qquad \mathcal{U}_{\mathrm{KS}}(f) := \sum_{s\in[K]} \mathrm{KS}\left(\mathrm{Law}(f(\boldsymbol{X}, S) \mid S = s), \mathrm{Law}(f(\boldsymbol{X}, S))\right).$$

Ces mesures sont légèrement différentes des comparaisons par paire que nous avons introduites, mais on peut facilement observer qu'elles reviennent finalement au même type de relaxation.

En utilisant ces notions, on souhaite étudier les prédicteurs f qui satisfont à la contrainte d'équité relaxée $\mathcal{U}_{\square}(f) \leq \varepsilon$, où $\square$ est KS ou TV et $\varepsilon \geq 0$ est un paramètre spécifié par l'utilisateur. Notons que puisque KS et TV sont des métriques, définir $\varepsilon = 0$ est équivalent à la contrainte DP. En revanche, pour $\varepsilon > 0$, ces formulations permettent un certain relâchement. Il est connu que la distance TV est plutôt forte et extrêmement sensible aux petits changements dans les distributions, ce qui est le principal inconvénient de l'inéquité TV. Cette limitation peut être résolue par l'inéquité KS grâce à une relation évidente : $\mathcal{U}_{\mathrm{KS}}(f) \leq \mathcal{U}_{\mathrm{TV}}(f)$.

Le prix de l'équité et le compromis risque-équité. Maintenant que nous avons défini une notion quantitative de l'inéquité, nous pouvons étudier l'impact de l'incorporation d'une contrainte d'équité dans le problème de minimisation du risque. Comme d'habitude, nous définissons le prédicteur optimal de Bayes f^* comme un minimiseur non contraint du risque

$$f^* \in \arg\min_{f:\mathcal{Z}\to\mathcal{Y}} \mathcal{R}(f).$$

Pour la clarté de l'exposé, nous supposerons dans ce qui suit que le minimum est effectivement atteint. Le prédicteur de Bayes et le niveau de risque qui lui est associé nous donnent une idée de la meilleure performance (risque) que l'on peut espérer obtenir pour un problème donné. Notons que le niveau d'inéquité du prédicteur optimal de Bayes dépend du problème en question, et non des choix du statisticien. En particulier, si l'on veut optimiser simultanément le risque et l'inéquité, il faut définir de nouvelles notions d'optimalité. Pour une métrique donnée $\square$ (*e.g.,* TV ou KS) et un nombre réel positif $\varepsilon > 0$, nous définissons un prédicteur optimal ε-équitable comme un minimiseur f^*_ε du risque dont l'inéquité, mesurée par la fonctionnelle $\mathcal{U}_{\square}$, ne dépasse pas le seuil ε :

$$f^*_\varepsilon \in \arg\min\{\mathcal{R}(f) : \mathcal{U}_{\square}(f) \leq \varepsilon\}.$$

Puisque le prédicteur optimal de Bayes présente le risque le plus faible parmi tous les prédicteurs, nous devons avoir

$$\mathcal{R}(f^*_\varepsilon) - \mathcal{R}(f^*) \geq 0,$$

pour tout prédicteur ε-équitable f^*_ε et pour tout $\varepsilon > 0$, ce qui montre que l'incorporation de la contrainte d'équité ne peut qu'augmenter le risque. En outre, puisque nous avons un ordre naturel sur les contraintes, la différence de risque dans l'équation (2.1) diminue par rapport

au seuil ε. Ainsi, un compromis naturel apparaît entre la condition d'équité et le risque et il peut être entièrement mesuré par les excès de risque

$$\mathcal{R}(f_\varepsilon^*) - \mathcal{R}(f^*), \quad \varepsilon \geq 0.$$

L'un de nos objectifs sera de comprendre et de quantifier précisément ce compromis dans le cadre de la régression afin de répondre à des questions telles que : pour un niveau de risque donné, quel est le meilleur niveau d'équité réalisable ? Ou quel est le coût en risque du passage du seuil d'équité ε à ε' ?

3.3 Régression équitable et transport optimal

Contrairement à son homologue en classification, le problème de la régression équitable a reçu beaucoup moins d'attention dans la littérature. Cependant, comme le soutient Agarwal, Dudík, and Wu (2019), les classificateurs ne fournissent que des décisions binaires, alors qu'en pratique, les décisions finales sont prises par les humains sur la base des prédictions de la machine. Dans ce cas, une prédiction continue est plus informative qu'une prédiction binaire et justifie la nécessité d'étudier l'équité dans le cadre de la régression. Jusqu'à très récemment, les contributions sur la régression équitable étaient presque exclusivement axées sur l'incorporation pratique de contraintes d'équité "proxy" dans les méthodes d'apprentissage classiques, telles que la forêt aléatoire, la régression ridge, les méthodes à noyau pour n'en citer que quelques-unes (Calders et al. 2013; Komiyama and Shimao 2017; Berk et al. 2017; Pérez-Suay et al. 2017; Raff, Sylvester, and Mills 2018; Fitzsimons et al. 2018). Plusieurs travaux étudient empiriquement l'impact des contraintes d'équité (relaxées) sur le risque (Bertsimas, Farias, and Trichakis 2012; Zliobaite 2015; Haas 2019; Wick, Panda, and Tristan 2019; Zafar et al. 2017). Pourtant, le problème de la quantification précise de l'effet de ces contraintes sur le risque n'a pas été abordé.

Plus récemment, des garanties statistiques et d'apprentissage pour la régression équitable ont été obtenues (Agarwal, Dudík, and Wu 2019; Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c; Chiappa et al. 2020; Fitzsimons et al. 2019; Plečko and Meinshausen 2019; Chzhen et al. 2020a). Les travaux les plus proches de notre contribution sont ceux de (Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c; Chiappa et al. 2020), qui établissent une connexion entre le problème de la régression exactement équitable (pour la parité démographique) et la formulation du transport optimal multi-marginal (Gangbo and Święch 1998; Agueh and Carlier 2011). En particulier, (Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c) ont obtenu la forme de la prédiction équitable optimale, fournissent des garanties statistiques sur les estimateurs de type plug-in, et établissent la valeur exacte du risque de la prédiction équitable optimale. Nous citons intégralement leur résultat qui a servi de point de départ à notre travail.

Theorem 3.3.1 (Le Gouic, Loubes, and Rigollet (2020) and Chzhen et al. (2020c))**.** *Supposons que pour tout $s \in [K]$, la variable aléatoire $(f^*(\mathbf{X}, S)|S = s)$ a un second moment fini et est non-atomique. Alors,*

$$\min\left\{\mathcal{R}(f) \,:\, (f(\boldsymbol{X}, S) \mid S = s) \overset{d}{=} (f(\boldsymbol{X}, S) \mid S = s') \;\forall s, s' \in [K]\right\} = \mathcal{U}(f^*).$$

De plus, la distribution du minimiseur du problème est donnée par

$$\operatorname*{arg\,min}_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{K} w_s \mathsf{W}_2^2 \left(\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S{=}s), \nu \right).$$

Ce résultat est important pour deux raisons : il met sur la même échelle une mesure de la performance, le risque $\mathcal{R}$, et une mesure de l'inéquité $\mathcal{U}$. De plus, il exprime une connexion profonde entre le problème du barycentre de Wasserstein-2 et le problème de l'apprentissage équitable du risque au carré sous la parité démographique.

Nous notons que l'utilisation des outils de transport optimal dans l'étude de l'équité est relativement récente. Initialement, les contributions dans cette direction portaient principalement sur le problème de la classification binaire parencitegordaliza2019obtaining,jiang2019wasserstein. Plus tard, les outils de la théorie du transport optimal ont migré vers la configuration de la régression équitable (Chiappa et al. 2020; Chzhen et al. 2020c; Le Gouic, Loubes, and Rigollet 2020). Comme nous le verrons, nous avons fait un usage intensif de la métrique de Wasserstein-2 pour obtenir davantage de résultats dans le cadre de la régression équitable.

CHAPTER 4

A nonasymptotic law of iterated logarithm for general M-estimators

M-estimators are ubiquitous in machine learning and statistical learning theory. They are used both for defining prediction strategies and for evaluating their precision. In this chapter, we propose the first non-asymptotic "any-time" deviation bounds for general M-estimators, where "any-time" means that the bound holds with a prescribed probability for *every* sample size. These bounds are non-asymptotic versions of the law of iterated logarithm. They are established under general assumptions such as Lipschitz continuity of the loss function and (local) curvature of the population risk. These conditions are satisfied for most examples used in machine learning, including those ensuring robustness to outliers and to heavy-tailed distributions. As an example of application, we consider the problem of best arm identification in a stochastic multi-armed bandit setting. We show that the established bound can be converted into a new algorithm, with provably optimal theoretical guarantees. Numerical experiments illustrating the validity of the algorithm are reported.

Based on Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak S. Dalalyan (2020). "A nonasymptotic law of iterated logarithm for general M-estimators". In: *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 1331–1341.

Contents

4.1 Introduction

Perhaps the most fundamental theorems in statistics are the law of large numbers (LLN) and the central limit theorem (CLT). Morally, they state that a sample average converges almost surely or in probability to the population average, and if one zooms in by multiplying by a square root factor, a much weaker form of stochastic convergence still holds, namely, convergence in distribution towards a Gaussian law. A fine intermediate result shows what happens in between the two scales: the law of iterated logarithm (LIL). By zooming in slightly less than in the CLT, *i.e.*, by rescaling the sample average with a slightly smaller factor than in the CLT, it is possible to gain a guarantee for infinitely many sample sizes, almost surely. In practice, however, the LIL has limited applicability, since it does not specify for which sample sizes the guarantee holds. The goals of the present work are to lift this limitation, by proving a LIL valid for every sample size, and holding for general M-estimators, rather than for the sample mean only.

The precise statement of the LIL, discovered by Khintchine (1924) and Kolmogoroff (1929) almost a century ago, is as follows: for a sequence of i.i.d. random variables $\{Y_i\}_{i\in\mathbb{N}}$ with mean θ and variance $\sigma^2 < \infty$, the sample averages $\bar{Y}_n = (Y_1 + \ldots + Y_n)/n$ satisfy the relations

$$\liminf_{n\to\infty} \frac{\sqrt{n}\,(\bar{Y}_n - \theta)}{\sigma\sqrt{2\ln\ln n}} = -1,$$
$$\limsup_{n\to\infty} \frac{\sqrt{n}\,(\bar{Y}_n - \theta)}{\sigma\sqrt{2\ln\ln n}} = 1,$$

almost surely. This provides a guarantee on the deviations of the sample average as an estimator of the mean θ since it yields that, with probability one, for any constant $c > 1$, there exists an integer $n_0 \in \mathbb{N}$ such that $|\bar{Y}_n - \theta| \leq c\sigma(2\ln\ln n/n)^{1/2}$ for every $n \geq n_0$. As compared to the deviation guarantees provided by the central limit theorem, the one of the last sentence has the advantage of being valid for any sample size large enough. This advantage is gained at the expense of a factor $(\ln\ln n)^{1/2}$. Akin for the classic version of the CLT, the applicability of the LIL is limited by the fact that it is hard to get any workable expression of n_0.

In the case of the CLT and its use in statistical learning, the drawback related to n_0 was lifted by exploiting concentration inequalities, such as the Hoeffding or the Bernstein inequalities, that can be seen as non-asymptotic versions of the CLT. For bounded random variables, the aforementioned concentration inequalities imply that for a prescribed tolerance level $\delta \in (0,1)$, for every $n \in \mathbb{N}$, the event[1] $\mathcal{A}_n = \{|\bar{Y}_n - \theta| \leq C(\ln(1/\delta)/n)^{1/2}\}$ holds with probability at least $1-\delta$. Such a deviation bound is satisfactory in a batch setting, when all the data are available in advance. In contrast, when data points are observed sequentially, as in online learning, or when the number of acquired data points depends on the actual values of the data points, the event of interest is $\bar{\mathcal{A}}_N = \mathcal{A}_1 \cap \ldots \cap \mathcal{A}_N$ or even a version of it in which N can be replaced by ∞. One can use the union bound to ensure that $\bar{\mathcal{A}}_N$ has a probability at least $1 - N\delta$ but this is too crude. Furthermore, replacing in $\mathcal{A}_n$ the confidence δ by δ/n^2, we get coverage $1 - \pi^2\delta/6$, valid for any sample size n, for an interval of length $O((\ln n/n)^{1/2})$. This result, obtained by a straightforward application of the union bound, is sub-optimal. A remedy to such a sub-optimality—in the form of a nonasymptotic version of the LIL—was proposed by Jamieson et al. (2014) and further used by Kaufmann, Cappé, and Garivier (2016), Kaufmann and Koolen (2018), and Howard et al. (2018). In addition, its relevance for online learning was demonstrated by deriving guarantees for the best arm selection in a multi-armed bandit setting. Note that these recent results apply exclusively to the sample mean in the one-dimensional setting; there is no equivalent of these bounds for other types of (possibly multivariate) estimators.

In this work, we establish a non-asymptotic LIL in a general setting encompassing many estimators, far beyond the sample mean. More precisely, we focus on the class of (penalized) M-estimators comprising the sample mean but also the sample median, the quantiles, the least-squares estimator, etc. Of particular interest to us are estimators that are robust to outliers and/or to heavy-tailed distributions. This is the case of the median, the quantiles, the Huber estimator, etc. (Huber 1964; Huber and Ronchetti 2009). It is well known that under mild assumptions, M-estimators are both consistent and asymptotically normal, *i.e.*, suitably adapted versions of the LLN and the CLT apply to them (Vaart 1998; Portnoy 1984; Collins 1977). Moreover, some versions of the LIL were also shown for M-estimators (Arcones 1994; He and Wang 1995). They suffer, however, from the same limitations as those explained above for the standard LIL. Our contributions allow to circumvent these limitations by providing a general non-asymptotic LIL for M-estimators both in one dimensional and in multivariate cases.

We apply the developed methodology to the problem of multi-armed bandits when the rewards are heavy-tailed or contaminated by outliers. In such a context, Altschuler, Brunel, and Malek 2018 tackled the problem of best median arm identification; this corresponds to replacing the average regret by the median regret. The relevance of this approach relies on the fact that even a small number of contaminated samples obtained from each arm may make the corresponding means arbitrarily large. In that setup, would it be possible to improve the upper bounds on the sample complexity of their algorithm—similarly to Jamieson et al. 2014—by using some finite-sample any-time version of the LIL for empirical medians or, more generally, for robust estimators? Our main results yield a positive answer to this question and establish rate-optimality of the proposed algorithm.

The rest of the paper is organized as follows. The next section contains the statement of the

[1] Here C is a universal constant.

LIL in a univariate setting and provides some examples satisfying the required conditions. A multivariate version of the LIL for penalized M-estimators is presented in Section 4.3.3. An application to online learning is carried out in Section 4.4.2, while a summary of the main contributions and some future directions of research are outlined in Section 4.6. Detailed proofs are deferred to the supplementary material.

4.2 Uniform LIL for univariate M-estimators

In this section, we focus on the case of univariate M-estimators. This is a vast family that contains the sample mean, the sample median and many other estimators. The relevance of M-estimators in contaminated models has been highlighted by several studies (see Huber 1964; Maronna 1976 as well as the recent work Loh (2015) and references therein).

4.2.1 Assumptions and main result

The precise setting considered in this section is the following. Random variables $Y, Y_1, Y_2, \dots$ are independently drawn from a probability distribution $\mathbb{P}_Y$ on some space $\mathcal{Y}$. Let $\phi : \mathcal{Y} \times \Theta \to \mathbb{R}$ be a given loss function, where Θ is an open interval in $\mathbb{R}$. Throughout this work, we make the tacit assumption that the random variable $\phi(Y, \theta)$ has a finite expectation for all $\theta \in \Theta$. The population and the empirical risks are then defined, respectively, by the formulas

$$\Phi(\theta) = \mathbb{E}\left[\,\phi(Y, \theta)\,\right], \quad \widehat{\Phi}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, \theta),$$

where $n \geq 1$ is an integer. We denote by θ^* a minimizer of Φ on Θ, and by $\widehat{\theta}_n$ a minimizer of $\widehat{\Phi}_n$ on Θ.

Assumption 4.2.1. *The function $\phi(Y, \cdot)$ is convex $\mathbb{P}_Y$-almost surely and $\phi(Y, \theta) \to \infty$ as θ approaches the boundary of Θ, $\mathbb{P}_Y$-almost surely (we say that the function $\phi(Y, \cdot)$ is convex and coercive).*

Assumption 4.2.1 requires from the loss ϕ to be approximately U-shaped and guarantees that θ^* and $\widehat{\theta}_n$ are well defined. To show that $\widehat{\theta}_n$ converges fast enough (with high probability) to θ^*, we will impose a local positive-curvature assumption on the population risk.

Assumption 4.2.2. *There exist two positive constants r and α such that for all $\theta \in \Theta$ with $|\theta - \theta^*| \leq r$, $\Phi(\theta) \geq \Phi(\theta^*) + (\alpha/2)(\theta - \theta^*)^2$.*

It is worth emphasizing here that this "local positive-curvature" assumption needs to hold for the population risk only. Clearly, a sufficient condition for Assumption 4.2.2 to hold is that Φ is strongly convex in a neighborhood of θ^*. Finally, to be able to obtain non-asymptotic guarantees that take the form of anytime Gaussian concentration, we require from the process $\theta \mapsto \phi(Y - \theta)$ to be smooth and to have sub-Gaussian tails (see 1.3 for a definition of sub-Gaussian random variables).

Assumption 4.2.3. *There exists a positive constant σ such that the random variables $\phi(Y, \theta) - \phi(Y, \theta^*)$ are $\sigma^2(\theta - \theta^*)^2$-sub-Gaussian for all $\theta \in \Theta$.*

One checks that Assumption 4.2.3 is fulfilled if $\phi(Y, \cdot)$ is η-Lipschitz with a sub-Gaussian variable η. We stress that the function ϕ is not assumed differentiable and, more importantly, that Y is not necessarily sub-Gaussian. We are now ready to state our first theorem on the uniform concentration of M-estimators.

Theorem 4.2.4. *Let Assumptions 4.2.1 to 4.2.3 hold. For any $\delta \in (0,1)$, set*

$$t_{n,\delta}^{\mathrm{LIL}} := \frac{3.3\sigma}{\alpha}\sqrt{\frac{1.1\ln\ln n + \ln(15/\delta) + 2.6}{n}}.$$

Let $n_0 = n_0(\alpha, r, \delta)$ be the smallest integer $n \geq 12$ for which $t_{n,\delta}^{\mathrm{LIL}} \leq r$. Then,

$$\mathbb{P}\left(\forall n \geq n_0, \quad |\widehat{\theta}_n - \theta^*| \leq t_{n,\delta}^{\mathrm{LIL}}\right) \geq 1 - \delta. \tag{4.1}$$

While the complete proof of Theorem 4.2.4 is postponed to the supplementary material, let us make a quick comment. In our proof, we show that it is enough to establish any-time concentration inequalities for sums of sub-Gaussian random variables. For partial sums of a sequence of sub-Gaussian random variables, sharp any-time concentration inequalities were recently proved in Jamieson et al. (2014), Maillard (2019), and Howard et al. (2018). However, these bounds do not apply in our case, since the terms in the sums arising in our proof change with the size of the sum. In other words, our sums are not partial sums of a given sequence of sub-Gaussian random variables.

Our proof is based on the peeling trick which, as noted in Garivier and Leonardi (2011, Proposition A.1), seemed to have first appeared in proofs for the Law of Iterated Logarithm (see, *e.g.*, Neveu (1972)).

The setting described in the beginning of this section might seem disconnected from any application, since it builds on an infinite set of independent random variables. However, the validity of the bound for an infinity of values of the sample size n makes it suitable for using in situations where the sample size is random and data-dependent. More precisely, the last theorem implies that for any $\delta \in (0,1)$ and for any random variable N taking values in the set of natural numbers $\mathbb{N}$, we have, with probability larger than $1-\delta$,

$$|\widehat{\theta}_N - \theta^*| \leq \frac{3.3\sigma}{\alpha}\sqrt{\frac{1.1\ln\ln N + \ln(15/\delta) + 2.6}{N}}$$

For instance, if we assume that the acquisition of each data point y has a cost $\psi(y)$, the number N might be given by $N = \max\{n : \psi(Y_1) + \ldots + \psi(Y_n) \leq B\}$, where B is a given available budget.

4.2.2 Examples

We now present three common examples for which all the assumptions presented above are satisfied. In all these examples, $\mathcal{Y} = \Theta = \mathbb{R}$.

Mean estimation Let $\phi(x, \theta) = (x-\theta)^2$. Assume that Y is s^2-sub-Gaussian. Then, one can check that Assumptions 4.2.1 to 4.2.3 are all satisfied with $r = +\infty$, $\alpha = 2$ and $\sigma = 2s$. For an in-depth analysis of this particular case we refer to (Howard et al. 2018).

Median and quantile estimation Let $\phi(x,\theta) = |x-\theta| - |x|$. Assume that Y has a unique median θ^* and that its cumulative distribution function F satisfies $|F(\theta) - 1/2| \geq (\alpha/2)|\theta - \theta^*|$, for all $\theta \in [\theta^* - r, \theta^* + r]$, where $\alpha, r > 0$ are fixed numbers. Then, θ^* is the unique minimizer of Φ and for all $\theta \in [\theta^* - r, \theta^* + r]$, the increment $\Phi(\theta) - \Phi(\theta^*)$ is equal to

$$\begin{aligned} 2\int_{\theta^*}^{\theta} x \,\mathrm{d}F(x) - (\theta - \theta^*) + 2(\theta F(\theta) - \theta^* F(\theta^*)) \\ \overset{(a)}{=} 2\int_{\theta^*}^{\theta} F(x)\,\mathrm{d}x - (\theta - \theta^*), \end{aligned}$$

where (a) is obtained by integration by parts. Hence, $\Phi(\theta) - \Phi(\theta^*) = \int_{\theta^*}^{\theta}(2F(x)-1)\,dx \geq \alpha/2(\theta - \theta^*)^2$, yielding Assumption 4.2.2. Moreover, since $\phi(Y,\theta)$ is bounded almost surely and 1-Lipschitz, for all $\theta \in \mathbb{R}$, Assumption 4.2.3 is automatically true (with $\sigma = 1$).

The same arguments hold true if $\phi(x,\theta) = \tau_\beta(x-\theta) - \tau_\beta(x)$, where $\tau_\beta(x) = \beta x - x_-$ with $x_- = \min(x, 0)$ the negative part. For this function ϕ, θ^* is the β-quantile of $\mathbb{P}_Y$, for $\beta \in (0,1)$.

Huber's M-estimators For $c > 0$, we define by $g_c(x) = x^2$ if $|x| \leq c$ and $g_c(x) = c(2|x| - c)$ if $|x| > c$. Let $\phi(x,\theta) = g_c(x-\theta) - g_c(x)$. This function g_c being $2c$-Lipschitz, Assumption 4.2.3 is satisfied with $\sigma = 2c$. Assume that Y has a positive density f on $\mathbb{R}$. Then, it is easy to check that Φ is twice differentiable, with $\Phi''(\theta) = 2\,(\,F(\theta + c) - F(\theta - c)\,) > 0$, for all $\theta \in \mathbb{R}$, where F is the cumulative distribution function of Y. Hence, θ^* is well-defined and unique, and if there exists $m > 0$ such that $f(x) \geq m$ for $x \in [\theta^* - 2c, \theta^* + 2c]$, then Assumption 4.2.2 is satisfied with $r = 2c$ and $\alpha = 4cm$.

4.2.3 Comparison with union bound

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables and let $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a loss such that assumptions of Theorem 4.2.4 are satisfied. Let $\widehat{\theta}_n$ be the M-estimator associated with the samples $Y_1, \ldots, Y_n$ and the loss ϕ. Using the same trick we developed for the proof of Theorem 4.2.4 (see supplementary material) we obtain the following tail bound : $\forall n \geq 1$, $\mathbb{P}(|\widehat{\theta}_n - \theta^*| > \frac{2\sigma}{\alpha}\sqrt{2\ln(2/\delta)/n}) \leq \delta$. Setting

$$t^{\mathrm{UB}}_{n,\delta} := \frac{2\sigma}{\alpha}\sqrt{\frac{2\ln(2n^{1+\varepsilon}/\delta)}{n}},$$

the union bound leads to

$$\mathbb{P}\left(\forall\, n \geq 12 \quad |\widehat{\theta}_n - \theta^*| \leq t^{\mathrm{UB}}_{n,\delta}\right) \geq 1 - \sum_{n=12}^{\infty} \frac{\delta}{n^{1+\varepsilon}}. \tag{4.2}$$

Figure 4.1 shows the ratio of the sub-Gaussian upper bound $t^{\mathrm{UB}}_{n,\delta'}$ over the LIL upper bound $t^{\mathrm{LIL}}_{n,\delta}$ provided by Theorem 4.2.4 for different levels of global confidence. The parameters δ and δ' are chosen to guarantee that the right hand sides in both (4.1) and (4.2) are equal to the prescribed confidence level $1 - \nu$. For $t^{\mathrm{UB}}_{n,\delta'}$, we chose $\varepsilon = 0.1$, the results for other values of ε being very similar. We observe that for most sample sizes n, the LIL bound is tighter than the one obtained by the union bound. In addition, the gap between the bounds widens as the sample size grows.

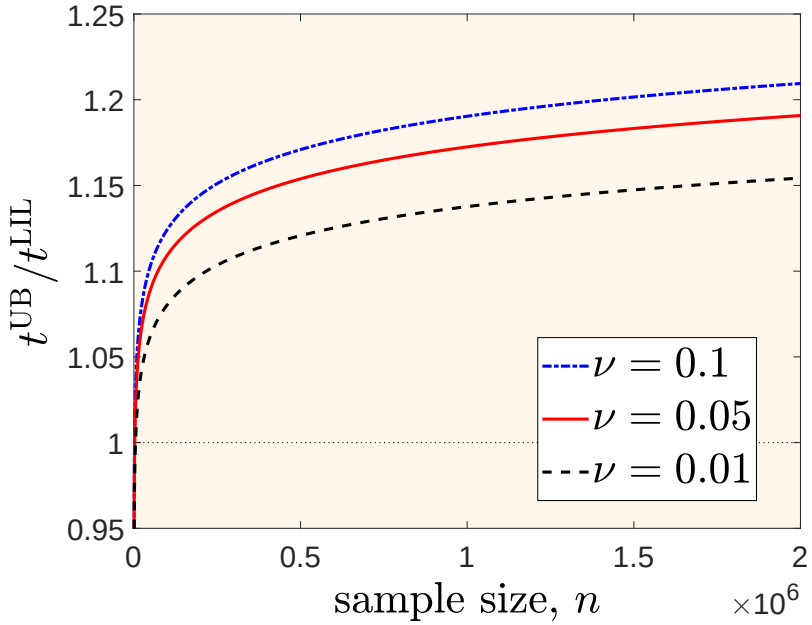


Figure 4.1: Ratio $t^{\mathrm{UB}}_{n,\delta'}/t^{\mathrm{LIL}}_{n,\delta}$ for different sample sizes n and confidence levels ν.

4.3 Uniform LIL for M-estimators of a multivariate parameter

We consider here the multivariate analog of the previous problem. The goal is to predict a real-valued label using a d-dimensional feature.

4.3.1 Assumptions and main result

We are given n independent label-feature pairs $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, with labels $Y_i \in \mathbb{R}$ and features $\boldsymbol{X}_i \in \mathbb{R}^d$, drawn from a common probability distribution $\mathbb{P}$. Let $\phi_n : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a given loss function and $\rho_n : \mathbb{R}^d \to \mathbb{R}$ be a given penalty. We assume throughout that the random variable $\phi_n(Y_1, \boldsymbol{\theta}^\top \boldsymbol{X}_1)$ has a finite expectation, for every $\boldsymbol{\theta}$, with respect to the probability distribution $\mathbb{P}$.

For a sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, we define the penalized empirical and population risks

$$\widehat{\Phi}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \phi_n(Y_i, \boldsymbol{\theta}^\top \boldsymbol{X}_i) + \rho_n(\boldsymbol{\theta}),$$
$$\Phi_n(\boldsymbol{\theta}) = \mathbb{E}\big[\phi_n(Y_1, \boldsymbol{\theta}^\top \boldsymbol{X}_1)\big] + \rho_n(\boldsymbol{\theta}).$$

Note that both the loss function ϕ_n and the penalty ρ_n are allowed to depend on the sample size n. Since our results are non-asymptotic, this dependence will be reflected in the constants appearing in the law of iterated logarithm stated below. We also define the penalized M-estimator $\widehat{\boldsymbol{\theta}}_n$ and its population counterpart $\boldsymbol{\theta}^*_n$ by

$$\widehat{\boldsymbol{\theta}}_n \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \widehat{\Phi}_n(\boldsymbol{\theta}) \quad \text{and} \quad \boldsymbol{\theta}^*_n \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \Phi_n(\boldsymbol{\theta}). \tag{4.3}$$

Typical examples where such a formalism is applicable are the maximum a posteriori approach and penalized empirical risk minimization. Our goal is to establish a tight non-asymptotic bound on the error of $\widehat{\boldsymbol{\theta}}_n$, that is, with high probability, valid for every $n \in \mathbb{N}$.

The main result of this section is valid under the assumptions listed below. We will present later on some common examples in which all these assumptions are satisfied.

Assumption 4.3.1. (Lipschitz loss) *The function $\theta \mapsto \phi_n(y, \theta)$ is L_n-Lipschitz, for every fixed $y \in \mathbb{R}$.*

Assumption 4.3.2. (Convex penalty) *The function $\theta \mapsto \widehat{\Phi}_n(\theta)$ is convex almost surely.*

Assumption 4.3.3. (Curvature of the population risk) *There exists a positive non-increasing sequence (α_n) such that, for any $n \in \mathbb{N}^*$, for any $\boldsymbol{w} \in \mathbb{R}^d$, $\Phi_n(\boldsymbol{\theta}_n^* + \boldsymbol{w}) - \Phi_n(\boldsymbol{\theta}_n^*) \geq (\alpha_n/2)\|\boldsymbol{w}\|_2^2$.*

Assumption 4.3.4. (Boundedness of features) *There exists a positive constant B such that $\|\boldsymbol{X}_1\|_2 \leq B$ almost surely.*

We will use the notation $\kappa_n = L_n/\alpha_n$ and refer to this quantity as the condition number. Note that all the foregoing assumptions are common in statistical learning, see for instance Sridharan, Shalev-shwartz, and Srebro (2009) and Rakhlin, Shamir, and Sridharan (2012). They are helpful not only for proving statistical guarantees but also for designing efficient computational methods for approximating $\widehat{\boldsymbol{\theta}}_n$.

Theorem 4.3.5. *Let Assumptions 4.3.1 to 4.3.4 be satisfied for every $n \in \mathbb{N}$. Assume, in addition, that starting from some integer $n_0 \geq 6$, the sequence $\kappa_n^2 \ln\ln n/n$ is decreasing. Define for any $\delta \in (0,1)$, $n \geq n_0$,*

$$t_{n,\delta}^{\mathrm{MVLIL}} = 3.6\kappa_n B \frac{\sqrt{\ln\ln n + \ln(50/\delta)} + 1}{\sqrt{n}}.$$

Then, for any $q \geq 2$ and $\delta \in (0,1)$, it holds that

$$\mathbb{P}\left(\forall n \geq n_0, \quad \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*\|_q \leq t_{n,\delta}^{\mathrm{MVLIL}}\right) \geq 1 - \delta.$$

4.3.2 Discussion

As an immediate consequence of Theorem 4.3.5 we get the following result. Let N be a randomly chosen integer that can depend on the infinite sequence $\{(\boldsymbol{X}_i, Y_i), i \in \mathbb{N}^*\}$ of random feature-label pairs drawn from $\mathbb{P}$. We observe only the first N elements of this sequence and wish to make a prediction of the label Y at a point $\boldsymbol{x} \in \mathbb{R}^d$. Assume that the best linear prediction is of the form $g(\boldsymbol{x}^\top \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is the minimizer of the expected loss and g is a known, L-Lipschitz, link function. Then, we can predict the label at $\boldsymbol{x}$ by $g(\boldsymbol{x}^\top \widehat{\boldsymbol{\theta}}_N)$, where $\widehat{\boldsymbol{\theta}}_N$ is the empirical risk minimizer. According to the last theorem, this predicted value satisfies

$$|g(\boldsymbol{x}^\top \widehat{\boldsymbol{\theta}}_N) - g(\boldsymbol{x}^\top \boldsymbol{\theta})| \leq L\|\boldsymbol{x}\|_2 \, t_{N,\delta}^{\mathrm{MVLIL}},$$

with probability at least $1 - \delta$.

A bound for the case $q \in [1, 2)$ can be obtained by using the fact that the ℓ_q norm is upper bounded by $d^{(2-q)/(2q)}$ times the ℓ_2 norm (Hölder's inequality). The resulting bound

corresponds to the ℓ_2 bound multiplied by this factor. Note this dependence on d is optimal, even in batch setting.

As noted above, all the foregoing assumptions are common in statistical learning. For instance, if $\rho_n(\boldsymbol{\theta}) = \lambda_n \|\boldsymbol{\theta}\|_2^2$ is the ridge penalty (Hoerl and Kennard 2000) and ϕ_n is either the absolute deviation ($\phi_{abs}(y, y') = |y - y'|$, see for instance Wang, Wonka, and Ye (2014)), the hinge ($\phi_{abs}(y, y') = (1 - yy')_+$ with $y \in [-1, 1]$) or the logistic ($\phi_{log}(y, y') = \ln(1 + e^{-yy'})$ with $y \in [-1, 1]$) loss, the aforementioned assumptions are satisfied with $L_n = 1$ and $\alpha_n = \lambda_n$. One can also consider the usual squared loss $\phi(y, y') = (y - y')^2$ under the additional assumption that Y is bounded by a known constant B_y. Under this condition, if the minimization problems in (4.3) are constrained to the ball of radius R, Assumptions 4.3.1 and 4.3.3 are satisfied with $\alpha_n = 1$ and $L_n = 2B_y + BR$. It should be noted that Assumption 4.3.3 is satisfied, for instance, when Φ_n is strongly convex. Remarkably, as opposed to some other papers (Shalev-Shwartz, Srebro, and Zhang 2010; Hsu and Sabato 2016), Theorem 4.3.5 requires this assumption for the population risk only.

Assumption 4.3.4 can be replaced, with some extra work, by sub-Gaussianity of $\|\boldsymbol{X}\|_2$. The statement of this extension and its proof can be found in Section 4.7.3.

4.3.3 Possible extensions

The conditions under which Theorem 4.3.5 holds can be further relaxed. We have in mind the following two extensions. First, the curvature condition can be restricted to a neighborhood of $\boldsymbol{\theta}_n^*$ only, by letting Φ_n grow linearly outside the neighborhood. Second, the Lipschitz assumption on ϕ_n can be replaced by the following one: for a constant β and a sub-Gaussian random variable η, the function $u \mapsto \phi_n(Y, u) - \beta u^2$ is η-Lipschitz. This last extension will allow us to cover the case of squared loss without restriction to a bounded domain. All these extensions are fairly easy to implement, but they significantly increase the complexity of the statement of the theorem. In this work, we opted for sacrificing generality in order to get better readability of the result.

Another interesting avenue for future research is the extension of the presented results to the high-dimensional online setting, *i.e.*, when the dimension might be larger than the sample size. In the batch setting, an in-depth analysis of M-estimators can be found in Negahban et al. (2012). It is also important in such a high-dimensional setting to avoid the factor B in the expression of $t_{n,\delta}^{\mathrm{MVLIL}}$, since it might scale as $\sqrt{d}$.

Finally, we can consider a more general setting in which the terms $\phi(Y_i, \boldsymbol{\theta}^\top \boldsymbol{X}_i)$ are replaced by $\psi(Z_i, \theta)$, where Z_i are i.i.d. random variables. The only change to be made is in replacing Assumptions 4.3.1 and 4.3.4 by a new assumption, that requires the function $[\psi(Z_i, \boldsymbol{\theta}) - \psi(Z_i, \boldsymbol{\theta'})]$ to be bounded by $|\boldsymbol{V}_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta'})|$, for all $\boldsymbol{\theta}$, $\boldsymbol{\theta'}$, with a random vector $\boldsymbol{V}_i$ which has a bounded (or sub-Gaussian) norm. This setting has the advantage of being more general than the one adopted in Section 3. However, the relevant examples we have in mind at correspond all to partial linear models.

4.4 Application to Bandits

In this section, we apply the univariate uniform law of iterated logarithm established in Section 4.2 to the multi-armed bandit problem. More precisely, we study the Best Arm Identification (BAI) problem in the fixed confidence setting. It consists in identifying, for a given confidence level and as fast as possible, which arm produces the highest expected outcome, (see Audibert and Bubeck (2010), Gabillon, Ghavamzadeh, and Lazaric (2012), and Kaufmann, Cappé, and Garivier (2016)). This means that we are able to collect data by sampling from K unknown distributions $\mathbb{P}_1, \ldots, \mathbb{P}_K$ and our goal is to identify the distribution having the largest expectation. Naturally, the same problem can be formulated for finding the distribution with the largest median, or the largest quantile of a given order. In particular, such a formulation of the problem might be of interest in cases where the expectations of the outcomes of each arm may not be defined (rewards are heavy-tailed) or are not meaningful (rewards are subject to some arbitrary contamination), see Altschuler, Brunel, and Malek 2018. We show in this section that theoretical results of previous sections provide an extension of the lil'UCB algorithm of Jamieson et al. (2014) to this framework.

4.4.1 Robust Best Arm Identification (RBAI)

We consider a robust version of BAI, which we call Robust BAI (RBAI). Suppose there are K arms, each arm $k \in [K]$ producing i.i.d. rewards

$$Y_{1,k}, Y_{2,k}, Y_{3,k}, \ldots \overset{\text{i.i.d.}}{\sim} \mathbb{P}_k.$$

At each round $n = 1, 2, \ldots$, the player chooses an arm $I_n \in [K]$ and receives the corresponding reward $Y_{T_{I_n}(n-1),I_n}$, where $T_k(n-1) = \mathbb{1}(I_1 = k) + \ldots + \mathbb{1}(I_{n-1} = k)$ is the number of times the arm k was pulled during the rounds $1, \ldots, n-1$.

For a given loss function $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, convex with respect to its second argument, we define

$$\theta_k \in \arg\min_{\theta \in \mathbb{R}} \mathbb{E}_{\mathbb{P}_k}[\phi(Y, \theta)].$$

From a statistical perspective, the problem under consideration encompasses that of finding the maximum point (by active learning) in a quantile regression problem (Chernozhukov 2005). For instance, consider the case of median regression. The aim is to maximize a function $f : [0,1] \to \mathbb{R}$ over a grid of points $x_1, \ldots, x_K \in [0,1]$, using noisy evaluations of f. At each round n, we can choose one x_k and observe the value

$$Y_n = f(x_k) + \xi_n,$$

where $\{\xi_n\}$ is a sequence of i.i.d. random variables with median equal to zero. Clearly, this enters into the framework described in the previous paragraph with $\theta_k = f(x_k)$ and each $\mathbb{P}_k$ is just a shifted-by-θ_k version of the distribution of ξ_n.

We use the rewards of the k-th arm for estimating θ_k by empirical risk minimisation: for every arm $k \in [K]$ and every sample size $n \geq 1$, we let

$$\widehat{\theta}_{k,n} \in \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \phi(Y_{i,k}, \theta).$$

With this notation, after n rounds, we are able to compute the quantities $\widehat{\theta}_{k,T_k(n)}$ for $k \in [K]$. These quantities, combined with the confidence bounds furnished by the LIL of Theorem 4.2.4, lead to M-estimator lil'UCB algorithm described in Algorithm 1[2].

Procedure 1 M-estimator lil'UCB.

Input: ν, λ, $\gamma > 0$ and $n_0 \in \mathbb{N}$

Sample each arm n_0 times
Set $\delta = ((\sqrt{16\nu + 9} - 3)/16)^2$
for k in $1 : K$ **do**
  $T_k \leftarrow n_0$
  Sample kth arm n_0 times
  Compute $\widehat{\theta}_{k,T_k}$
  Set $s(k) \leftarrow \widehat{\theta}_{k,T_k} + \gamma\sqrt{\frac{\ln \ln T_k + \ln(1/\delta)}{T_k}}$
end for
$n \leftarrow Kn_0$
while $(1+\lambda)\max_{k\in[K]} T_k < 1 + \lambda n$ **do**
  $I \leftarrow \arg\max_{k\in[K]} s(k)$
  Sample arm I
  Update $T_I \leftarrow T_I + 1$, $n \leftarrow n + 1$
  Compute $\widehat{\theta}_{I,T_I}$
  Set $s(I) \leftarrow \widehat{\theta}_{I,T_I} + \gamma\sqrt{\frac{\ln \ln T_I + \ln(1/\delta)}{T_I}}$
end while

Output: $\arg\max_{k\in[K]} T_k$

4.4.2 Main results

To state the theoretical results, let $k^* = \text{argmax}_{k\in[K]}\, \theta_k$ be the subscript corresponding to the best arm. We assume k^* to be unique, and we define, for $k \neq k^*$, the sub-optimality gaps $\Delta_k = \theta_{k^*} - \theta_k$. We introduce the quantities

$$\mathbf{H}_1 = \sum_{k\neq k^*} \frac{1}{\Delta_k^2}, \quad \mathbf{H}_2 = \sum_{k\neq k^*} \frac{\ln\left(2 + \ln_+(1/\Delta_k^2)\right)}{\Delta_k^2}.$$

Those quantities play a key role in characterizing the complexity of the BAI problem.

Theorem 4.4.1. *Let $\theta \mapsto \phi(y, \theta)$ be a convex function for every $y \in \mathbb{R}$ and let the distributions $\mathbb{P}_k$ satisfy Assumptions 4.2.2 and 4.2.3 with parameters $\alpha, \sigma > 0$. For any $\nu \in (0, 0.2)$ and $\beta \in (0, 4.8)$, there exist positive constants[3] λ and C such that with probability at least $1 - \nu$, Algorithm 1, used with parameters ν, λ, $\gamma = 4.4(1+\beta)\sigma/\alpha$ and $n_0 \geq 12$, stops after at most*

$$Kn_0 + C(\mathbf{H}_1 \ln(1/\nu) + \mathbf{H}_2)$$

steps and returns the best arm.

[2] λ, γ and n_0 should be seen as tuning parameters for which our theoretical results give some guidance.
[3] λ and C depend only on β and σ/α.

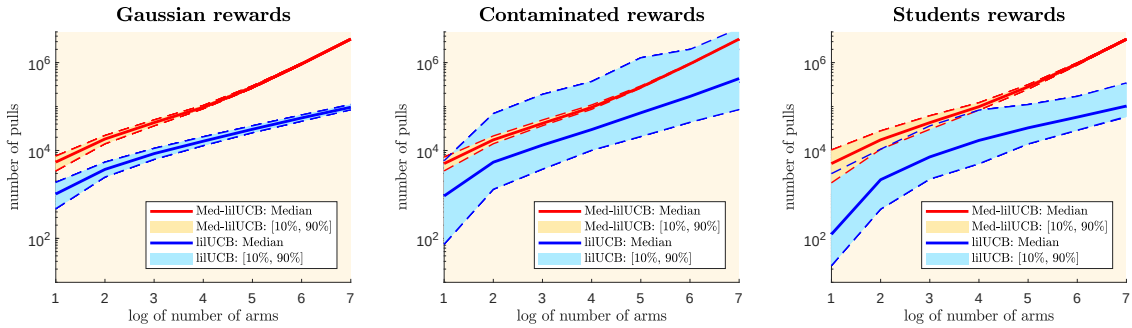


Figure 4.2: Total number of pulls done by the median lil'UCB and the lil'UCB algorithms for $K \in \{2, 4, 8, 16, 32, 64, 128\}$.

Note that $(\ln 2)\mathbf{H}_1 \leq \mathbf{H}_2$. Therefore, for a fixed confidence level ν, the number of pulls provided by Theorem 4.4.1 is $O(\mathbf{H}_2)$. The next result shows that this order of magnitude is optimal.

Theorem 4.4.2. *Consider the RBAI framework with fixed confidence $\delta \in (0, 1/2)$ described above and assume $K = 2$. Let $\theta_1, \theta_2 \in \mathbb{R}$ be such that $\Delta = |\theta_1 - \theta_2| > 0$. Let $\phi(y, \theta) = \phi_0(|y - \theta|)$ for some function ϕ_0 and the arm distributions be $\mathcal{N}(\theta_1, 1)$ and $\mathcal{N}(\theta_2, 1)$. Then, any algorithm that finds after T rounds the best arm with probability at least $1 - \delta$, for all values of $\Delta > 0$, must satisfy*

$$\limsup_{\Delta \to 0} \frac{\mathbb{E}[T]}{\Delta^{-2} \ln \ln(\Delta^{-2})} \geq 2 - 4\delta.$$

Proofs of these two theorems are provided in the supplementary material.

4.5 Numerical experiments

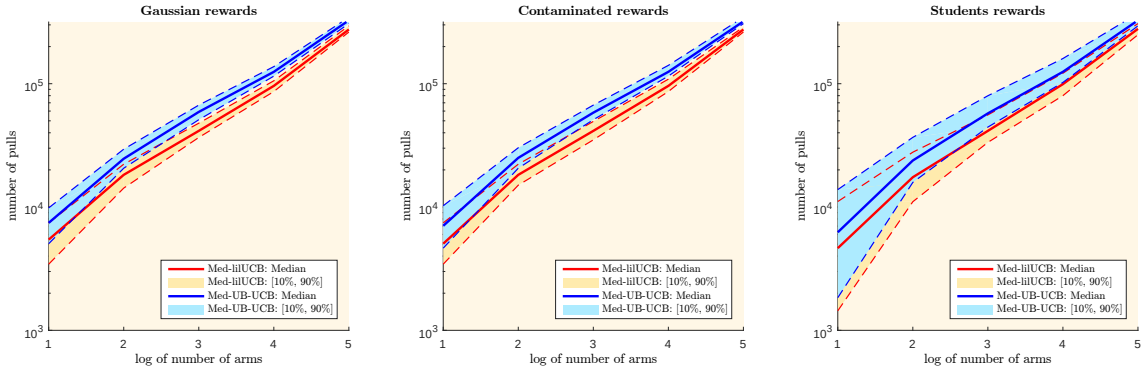


Figure 4.3: Total number of pulls done by the median lil'UCB and the median UCB based on the union bound.

To illustrate the results of the previous section, we conducted the following experiment. We chose the values of θ_k's according to the "α-model" from Jamieson et al. (2014) with $\alpha = 0.3$. It imposes an exponential decay on the parameters, that is $\theta_k = 1 - (k/K)^\alpha$. Along with these parameters, we consider three reward generating processes:

- *Gaussian rewards*, where $Y_{i,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_k, \sigma^2)$,

- *Gaussian rewards subject to Cauchy contamination*, where $Y_{i,k} \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)\mathcal{N}(\theta_k, \sigma^2) + \varepsilon\,\text{Cauchy}(\theta_k)$ for $\varepsilon = 5\%$,

- *Student rewards*, where $Y_{i,k} \overset{\text{i.i.d.}}{\sim} t_2(\theta_k)$ (*i.e.*, Student distribution with 2 degrees of freedom).

Note that all these processes are median centered at θ_k's. In the case of Gaussian and Student rewards, they are also mean-centered at θ_k, while in the case of contaminated Gaussian rewards the mean is not defined. To test the robustness of the compared algorithms, we tuned their parameters to fit the Gaussian reward scenario.

In this set-up, we compared the original lil'UCB algorithm from Jamieson et al. (2014)—see also Jamieson and Nowak (2014) for a more comprehensive experimental evaluation—and the M-estimator lil'UCB described in Algorithm 1, where $\widehat{\theta}_{k,n}$ is the empirical median of rewards from arm k up to time n. This corresponds to the M-estimator associated with the absolute deviation loss. This version of the M-estimator lil'UCB is hereafter referred to as median lil'UCB or med-lil'UCB.

In order to conduct a fair comparison, we assigned the same values to parameters shared by both procedures and set the values as in Jamieson et al. (2014): $\beta = 1$, $\lambda = (1 + 2/\beta)^2$, $\sigma = 0.5$, $\varepsilon = 0.01$ and $\nu = 0.1$. Note that, as underlined in Jamieson et al. 2014, the choice of λ does not fit their theoretical result. This choice is justified by the fact that λ should theoretically be proportional to $(1 + 2/\beta)^2$ with a constant converging to 1 when the confidence approaches 0. For our algorithm we chose $\gamma = 2$ and $n_0 = 20$.

The results, for several values of K (the total number of arms), obtained by 200 independent runs of each algorithm in all the three settings, are summarized in Figure 4.2 and in Table 4.1. Numbers reported in Table 4.1 represent the proportion of times each algorithm succeeded to find the best arm, while Figure 4.2 displays the number of pulls for each algorithm. Table 4.1 shows that lil'UCB performed poorly on the non-Gaussian models. For contaminated Gaussian rewards, the performance of lil'UCB deteriorates as the number of arms grows, while it does not seem to be affected by the number of arms in the case of Student rewards : it identifies correctly the best arm for only around 60% of the runs in this last case. In contrast, median lil'UCB performs well in all the three scenarios, giving perfect identification over all runs.

Table 4.1: Proportion of correct best arm identification (over 200 runs per scenario/algorithm).

K	Algorithm	Gauss	Contam.	Student
2	lil'UCB	1.00	0.81	0.61
	med-lil'UCB	1.00	1.00	1.00
4	lil'UCB	1.00	0.75	0.61
	med-lil'UCB	1.00	1.00	1.00
8	lil'UCB	1.00	0.69	0.63
	med-lil'UCB	1.00	1.00	1.00
16	lil'UCB	1.00	0.66	0.60
	med-lil'UCB	1.00	1.00	1.00
32	lil'UCB	1.00	0.57	0.61
	med-lil'UCB	1.00	1.00	1.00
64	lil'UCB	1.00	0.54	0.62
	med-lil'UCB	1.00	1.00	1.00
128	lil'UCB	1.00	0.44	0.60
	med-lil'UCB	1.00	1.00	1.00

The curves in Figure 4.2 represent the median number of pulls over the 200 runs while the colored areas around the curves are delimited by the 10% and 90% quantiles of the number of pulls over these 200 runs. We observe that the spread of the number of pulls of lil'UCB is large for non-Gaussian models, while the curves for median lil'UCB are almost identical in the three models of rewards. The number of pulls for median lil'UCB is higher than the number of pulls for lil'UCB in the Gaussian and Student models. However, in the contaminated Gaussian model, lil'UCB might require more pulls when the number of arms is large.

Moreover, we noticed that the performance of our procedure is not sensitive to the level of contamination : we conducted the same experiment with $\varepsilon \in \{5, 10, 20, 40, 60\}$ and in all cases our procedure is 100% successful in finding the best arm. Furthermore, the number of pulls does not increase when ε increases. In contrast, the performance of the original lil'UCB procedure drops down to 35% of correct identification when $\varepsilon = 60\%$ and there are 4 arms. Finally, we observed that if we replace the LIL by the naive union bound in our algorithm, the detection accuracy remains the same, but the running time increases (between 10% and 30%), see Figure 4.3.

These experiments illustrate the lack of robustness of lil'UCB to heavy tailed rewards and the effective robustness of median lil'UCB. Since this robustness comes with a higher number of pulls, median lil'UCB should be preferred to vanilla lil'UCB only if one suspects non-Gaussian or heavy-tailed rewards.

4.6 Conclusion and further work

We have proved a nonasymptotic law of iterated logarithm for general M-estimators both in univariate and in multivariate settings. These results can be seen as off-the-shelf deviation

bounds that are uniform in the sample size and, therefore, suitable for online learning problems and problems in which the sample size may depend on the observations. There are several avenues for future work. For simplicity, in the multivariate case, the population risk was assumed to be above an elliptic paraboloid on the whole space. First in our agenda is to replace this condition by a local curvature one. A second interesting line of research is to establish an any-time deviation bound for sequential estimators such as the online gradient descent. It would also be of interest to obtain "in-expectation" bounds of the same type as those in (Shin, Ramdas, and Rinaldo 2019).

4.7 Proofs

This section contains the proofs of the theorems stated and discussed in the main body of the paper. Some technical lemmas used in the proofs of this section are postponed to Section 4.7.6.

4.7.1 Notations

We begin by introducing notations we will use throughout the proofs. For any positive real number x, we define $\lceil x \rceil$ as the smallest integer greater than or equal to x. We denote by $\mathbb{S}^{d-1}$ the d-dimensional unit Euclidean sphere, *i.e.*, $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \|v\|_2 = 1\}$. The natural logarithm function (*i.e.*, base e logarithm) is denoted by ln. We write vectors and matrices in bold font, we use lower-case symbols for the former and upper-case symbols for the latter. When we write $\boldsymbol{X}_i$ we mean the j-th column of the matrix $\boldsymbol{X}$. We denote by $\|\mathbf{X}\|_{\mathsf{F}}$ the Frobenius norm of the matrix $\mathbf{X}$, $\|\mathbf{X}\|_{\mathsf{F}}^2 = \mathrm{Tr}(\mathbf{X}\mathbf{X}^\top)$. For a vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|_\infty$ stands for $\max_j |v_j|$. For any integer $K \in \mathbb{N}$, we set $[K] = \{1, \dots, K\}$.

4.7.2 Proof of Theorem 4.2.4

For any integer $n \geq 12$, define the sequence

$$t(n) = \frac{3.3\sigma}{\alpha\sqrt{n}} \sqrt{1.1 \ln \ln n + \ln(15/\delta) + 2.6}.$$

Note that it is a non-increasing sequence and converges to 0. Denote by n_0 the smallest positive integer $n \geq 12$ such that $t(n) \leq r$. For $k \geq 1$ and $\beta = 1.1$, let $n_k = \lceil \beta n_{k-1} \rceil$. To ease notation, we set $t_k = t(n_k)$. We also define the integer intervals $I_k = [n_k, n_{k+1}) \cap \mathbb{N}$. We wish to upper bound the probability of the event

$$\mathcal{A} = \bigcup_{n=n_0}^{\infty} \mathcal{A}_n, \quad \text{where} \quad \mathcal{A}_n = \{\widehat{\theta}_n - \theta^* > t(n)\}.$$

For $n \geq 1$ and $t \in (0, r]$, define the random variables

$$S_n(t) = n\left(\widehat{\Phi}_n(\theta^*) - \Phi(\theta^*)\right) - n\left(\widehat{\Phi}_n(\theta^* + t) - \Phi(\theta^* + t)\right).$$

For any integers $k \geq 0$, $n \in I_k$, the event $\mathcal{A}_n$ is included in the event $\mathcal{B}_n = \{S_n(t_{k+1}) \geq (\alpha/2) n_k t_{k+1}^2\}$. Indeed, the fact that the sequence $(t(n))$ is non-increasing, the convexity of the

function $\widehat{\Phi}_n$ (see Figure 4.4 for an illustration of the second implication) and Assumption 4.2.2 yield, for integers $k \geq 0$, $n \in I_k$,

$$\begin{aligned}\widehat{\theta}_n > \theta^* + t(n) &\implies \widehat{\theta}_n > \theta^* + t_{k+1} \\ &\implies \widehat{\Phi}_n(\theta^*) \geq \widehat{\Phi}_n(\theta^* + t_{k+1}) \\ &\implies S_n(t_{k+1})/n \geq \Phi(\theta^* + t_{k+1}) - \Phi(\theta^*) \\ &\implies S_n(t_{k+1}) \geq \frac{\alpha}{2} n_k t_{k+1}^2.\end{aligned}$$

Combining the previous observation with a union bound yield

$$\mathbb{P}\left(\bigcup_{n=n_0}^{\infty} \mathcal{A}_n\right) \leq \sum_{k=0}^{\infty} \mathbb{P}\left(\bigcup_{n\in I_k} \mathcal{A}_n\right) \leq \sum_{k=0}^{\infty} \mathbb{P}\left(\bigcup_{n\in I_k} \mathcal{B}_n\right).$$

Furthermore, letting $x_k = (\alpha/2) n_k t_{k+1}^2$, we get, for any positive λ,

$$\mathbb{P}\left(\bigcup_{n\in I_k} \mathcal{B}_n\right) \leq \mathbb{P}\left(\sup_{n\in I_k} S_n(t_{k+1}) \geq x_k\right) \leq \mathbb{P}\left(\sup_{n\in I_k} \exp\left\{\lambda S_n(t_{k+1})\right\} \geq e^{\lambda x_k}\right).$$

The stochastic process $(S_n(t_{k+1}))_{n\in I_k}$ is a discrete martingale. Hence, by Jensen's inequality, for all $\lambda > 0$, $(\exp(\lambda S_n(t_{k+1})))_{n\in I_k}$ is a discrete submartingale. Therefore, Markov's inequality followed by Doob's maximal inequality implies

$$\mathbb{P}\left(\bigcup_{n\in I_k} \mathcal{B}_n\right) \leq e^{-\lambda x_k} \mathbb{E}\left[\sup_{n\in I_k} \exp\left\{\lambda S_n(t_{k+1})\right\}\right] \leq e^{-\lambda x_k} \mathbb{E}\left[\exp\left\{\lambda S_{n_{k+1}}(t_{k+1})\right\}\right].$$

Since the random variable $S_{n_{k+1}}(t_{k+1})$ is the sum of n_{k+1} independent $\sigma^2 t_{k+1}^2$-sub-Gaussian random variables (Assumption 4.2.3), we have a simple upper bound on the moment generating function of $S_{n_{k+1}}(t_{k+1})$ which yields

$$\mathbb{P}\left(\bigcup_{n\in I_k} \mathcal{B}_n\right) \leq \exp\left\{-\lambda x_k + (\lambda^2\sigma^2/2) n_{k+1} t_{k+1}^2\right\}.$$

Choosing $\lambda = x_k/(\sigma^2 n_{k+1} t_{k+1}^2)$ and recalling that $\beta = 1.1$ (which ensures $\beta n_k/n_{k+1} \geq \sqrt{0.88}$), we obtain

$$\mathbb{P}\left(\bigcup_{n\in I_k} \mathcal{B}_n\right) \leq \exp\left\{-\frac{x_k^2}{2\sigma^2 n_{k+1} t_{k+1}^2}\right\} \leq \exp\left\{-\frac{\alpha^2 n_k^2 t_{k+1}^2}{8\sigma^2 n_{k+1}}\right\} \leq \exp\left\{-\frac{0.88\alpha^2 n_{k+1} t_{k+1}^2}{8\beta^2\sigma^2}\right\}.$$

Replacing t_{k+1} by its expression,

$$t_{k+1}^2 = \left(\frac{3.3\sigma}{\alpha}\right)^2 \frac{1.1 \ln\ln n_{k+1} + \ln(15/\delta) + 2.6}{n_{k+1}},$$

and using the inequality $\ln n_{k+1} \geq \ln(\beta^{k+1} n_0) \geq (k+27)\ln\beta$ we arrive at

$$\begin{aligned}\mathbb{P}\left(\bigcup_{n\in I_k} \mathcal{B}_n\right) &\leq \exp\left\{-\frac{0.88 \times 3.3^2 \times (1.1 \ln\ln n_{k+1} + \ln(15/\delta) + 2.6)}{8\beta^2}\right\} \\ &\leq \exp\left\{-\left(1.1\ln(k+27) + 1.1\ln\ln\beta + \ln(15/\delta) + 2.6\right)\right\} \\ &\leq \frac{\delta}{15}\exp\left\{-1.1\ln(k+27)\right\} = \frac{\delta}{15(k+27)^{1.1}}.\end{aligned}$$

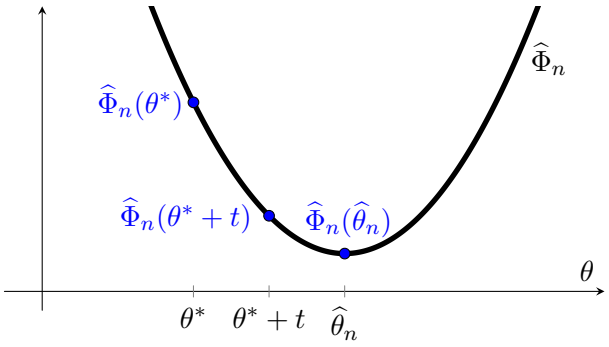


Figure 4.4: Illustration of the shape of the function $\widehat{\Phi}_n$.

Finally, using the fact that

$$\sum_{k=0}^{\infty} \frac{1}{(k+27)^{1.1}} \leq \int_{26}^{\infty} x^{-1.1}\, dx = \frac{26^{-0.1}}{0.1} \leq 7.5,$$

we get

$$\mathbb{P}\left(\exists n \geq n_0, \widehat{\theta}_n > \theta^* + t(n)\right) \leq \delta/2.$$

The exact same reasoning yields

$$\mathbb{P}\left(\exists n \geq n_0, \widehat{\theta}_n < \theta^* - t(n)\right) \leq \delta/2,$$

and Theorem 4.2.4 follows from a union bound combined with the two previous inequalities.

Remark 4.7.1. *Several high probability uniform bounds on the sum of sub-Gaussian random variables have been proved (see, e.g., Jamieson et al. (2014), Maillard (2019), and Howard et al. (2018)). However, those bound do not apply in our case since the elements of the sum change with the size of the sum.*

4.7.3 Proof of Theorem 4.3.5

Let $\beta = 1.1$. Throughout the proof, we consider the sequence of integers $\{n_k : k \in \mathbb{N}\}$ defined by $n_{k+1} = \lceil \beta n_k \rceil$ (recall that $n_0 \geq 6$). We also introduce the sequence of integer intervals $I_k = [n_k, n_{k+1}) \cap \mathbb{N}$ and the sequence $(t(n))_{n\in\mathbb{N}}$ defined by

$$t(n) = (39/11)\kappa_n B \frac{\sqrt{\ln\ln n + \ln(50/\delta)} + 1}{\sqrt{n}}, \quad \text{for } n \geq 1.$$

To avoid double subscripts, we write $t(n_k) = t_k$ for any integer k. We wish to upper bound the probability of the event

$$\mathcal{A}^q = \bigcup_{n=n_0}^{\infty} \mathcal{A}_n^q, \quad \text{where} \quad \mathcal{A}_n^q = \{\|\boldsymbol{\theta}_n^* - \widehat{\boldsymbol{\theta}}_n\|_q > t(n)\}.$$

n_0	β	C_{mult}	C_{add}
6	1.1	3.35	5.4
	1.05	3.3	6
	1.01	3.3	7.6
12	1.1	3.3	5.3
	1.05	3.1	6
	1.01	3.1	7.6
20	1.1	3.25	5.3
	1.05	3.1	6
	1.01	3	7.6
50	1.1	3.2	5.3
	1.05	3.1	6
	1.01	2.9	7.6
80	1.1	3.2	5.3
	1.05	3	6
	1.01	2.9	7.5

Table 4.2: Effect of the choice of n_0 and β on the constants in the sequence $t_n^{\text{LIL}} = C_{\text{mult}} \frac{\sigma}{\alpha} \sqrt{\frac{1.1 \ln\ln n + \ln(1/\delta) + C_{\text{add}}}{n}}$

.

Reduction to the case $q = 2$ Since for any real number $q \geq 2$ and vector $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\|_q \leq \|\boldsymbol{x}\|_2$, an upper bound on the probability of $\mathcal{A}^2$ implies an upper bound on the probability pf $\mathcal{A}^q$ for any $q \geq 2$. Therefore it is sufficient to obtain an upper bound for the case $q = 2$. For simplicity we write $\mathcal{A} := \mathcal{A}^2$ and $\mathcal{A}_n := \mathcal{A}_n^2$ from now on.

For every $\boldsymbol{w} \in \mathbb{R}^p$, we define the random variables

$$S_n(\boldsymbol{w}) = n\left(\widehat{\Phi}_n(\boldsymbol{\theta}_n^*) - \Phi_n(\boldsymbol{\theta}_n^*)\right) - n\left(\widehat{\Phi}_n(\boldsymbol{\theta}_n^* - \boldsymbol{w}) - \Phi_n(\boldsymbol{\theta}_n^* - \boldsymbol{w})\right), n \geq 1.$$

The following result is a consequence of the convexity of $\widehat{\Phi}_n$.

Lemma 4.7.2. *Under Assumptions 4.3.1 to 4.3.3, for any $k \in \mathbb{N}$ and $n \in I_k$, the event $\mathcal{A}_n$ is included in the event*

$$\mathcal{B}_n := \left\{ \sup_{\boldsymbol{w} \in t_{k+1}\mathbb{S}^{d-1}} \left[S_n(\boldsymbol{w}) - (\alpha_n/2) n t_{k+1}^2 \right] \geq 0 \right\}.$$

Combining Lemma 4.7.2 with the union bound gives

$$\mathbb{P}(\mathcal{A}) \leq \mathbb{P}\Big(\bigcup_{k \geq 0} \bigcup_{n \in I_k} \mathcal{B}_n \Big) \leq \sum_{k \geq 0} \mathbb{P}\Big(\bigcup_{n \in I_k} \mathcal{B}_n \Big).$$

Let k be an integer. Since the sequence $\{\alpha_n\}_n$ is non-increasing we have, for any integer $n \in I_k$, $\alpha_n \geq \alpha_{n_{k+1}}$. Setting $\beta = 1.1$ we have $n_k/n_{k+1} \geq 11/13$ for $n \geq 6$. Thus, for any

positive real λ,

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{n\in I_k}\mathcal{B}_n\Big) &\leq \mathbb{P}\Big(\sup_{n\in I_k}\sup_{\boldsymbol{w}\in t_{k+1}\mathbb{S}^{d-1}}\Big[S_n(\boldsymbol{w})-\frac{\alpha_n}{2}n_k t_{k+1}^2\Big]\geq 0\Big)\\ &\leq \mathbb{P}\Big(\sup_{n\in I_k}\sup_{\boldsymbol{w}\in t_{k+1}\mathbb{S}^{d-1}}\Big[S_n(\boldsymbol{w})-\frac{11\alpha_{n_{k+1}}}{26}n_{k+1}t_{k+1}^2\Big]\geq 0\Big)\\ &\leq \mathbb{P}\Big(\sup_{n\in I_k}\sup_{\boldsymbol{w}\in t_{k+1}\mathbb{S}^{d-1}}\exp\Big\{\lambda\Big(S_n(\boldsymbol{w})-\frac{11\alpha_{n_{k+1}}}{26}n_{k+1}t_{k+1}^2\Big)\Big\}\geq 1\Big).\end{aligned}$$

The stochastic process $\big(\sup_{\boldsymbol{w}\in t_{k+1}\mathbb{S}^{d-1}}\exp\big\{\lambda\big(S_n(\boldsymbol{w})-11\alpha_{n_{k+1}}n_{k+1}t_{k+1}^2/26\big)\big\}\big)$, $n\in\mathbb{N}^*$, is a submartingale with respect to its natural filtration. Therefore, Doob's maximal inequality for submartingales yields,

$$\mathbb{P}\Big(\bigcup_{n\in I_k}\mathcal{B}_n\Big)\leq\inf_{\lambda\geq 0}\mathbb{E}\left[\sup_{\boldsymbol{w}\in t_{k+1}\mathbb{S}^{d-1}}\exp\left\{\lambda\left(S_{n_{k+1}}(\boldsymbol{w})-\frac{11\alpha_{n_{k+1}}}{26}n_{k+1}t_{k+1}^2\right)\right\}\right]. \tag{4.4}$$

The next lemma uses classic tools from empirical processes theory such as the symmetrization trick and the contraction principle to bound the expectation from (4.4).

Lemma 4.7.3. *Under Assumption 4.3.1, given a positive integer m and three positive real numbers t, α and λ, letting $t'=(2m\alpha/L)t$, we have,*

$$\inf_{\lambda\geq 0}\mathbb{E}\left[\sup_{\boldsymbol{w}\in t\mathbb{S}^{d-1}}\exp\left\{\lambda\left(S_m(\boldsymbol{w})-\alpha m t^2\right)\right\}\right]\leq\inf_{\lambda\geq 0}\mathbb{E}\left[\sup_{\boldsymbol{w}\in t'\mathbb{S}^{d-1}}\exp\left\{\lambda\left(\boldsymbol{w}^\top\mathbf{X}\boldsymbol{\varepsilon}-(t')^2/2\right)\right\}\right],$$

where $\boldsymbol{\varepsilon}$ is a n-dimensional vector of i.i.d. Rademacher random variables independent of the matrix $\mathbf{X}\in\mathbb{R}^{d\times n}$ whose columns are the observations vectors $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$.

Let us introduce the additional notation

$$s_{k+1}=\frac{11n_{k+1}}{13\kappa_{n_{k+1}}}t_{k+1}.$$

Applying Lemma 4.7.3 with $m=n_{k+1},\alpha=11\alpha_{n_{k+1}}/26$ and $t=t_{k+1}$ gives

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{n\in I_k}\mathcal{B}_n\Big)&\leq\inf_{\lambda\geq 0}\mathbb{E}\left[\sup_{\boldsymbol{w}\in s_{k+1}\mathbb{S}^{d-1}}\exp\left\{\lambda(\boldsymbol{w}^\top\mathbf{X}\boldsymbol{\varepsilon}-s_{k+1}^2/2)\right\}\right] &(4.5)\\ &=\inf_{\lambda\geq 0}\mathbb{E}\left[\exp\left\{\lambda s_{k+1}\|\mathbf{X}\boldsymbol{\varepsilon}\|_2\right\}\right]e^{-\lambda s_{k+1}^2/2}. &(4.6)\end{aligned}$$

The last line follows from the simple identity $\|\boldsymbol{x}\|_2=\sup_{\|\boldsymbol{y}\|_2=1}\boldsymbol{y}^\top\boldsymbol{x}$ valid for any vector $\boldsymbol{x}\in\mathbb{R}^d$. We now state a lemma to bound the quantity $\mathbb{E}e^{\|\mathbf{X}\boldsymbol{\varepsilon}\|_2}$ for deterministic matrix $\mathbf{X}$.

Lemma 4.7.4. *Let $\mathbf{X}$ be a $d\times n$ deterministic matrix and $\boldsymbol{\varepsilon}$ be an n-dimensional vector with i.i.d. Rademacher entries. Then the following inequality holds*

$$\mathbb{E}e^{\|\mathbf{X}\boldsymbol{\varepsilon}\|_2}\leq 2e^{(3\|\mathbf{X}\|_{\mathsf{F}}+\|\mathbf{X}\|_{\mathsf{F}}^2)/2}.$$

Combining Lemma 4.7.4 and (4.6), we get

$$\mathbb{P}\Big(\bigcup_{n\in I_k}\mathcal{B}_n\Big) \le \inf_{\lambda\ge 0}\mathbb{E}\left[e^{\lambda(\|\mathbf{X}\boldsymbol{\varepsilon}\|_2 - s_{k+1}^2/2)}\right] \tag{4.7}$$

$$\le 2\mathbb{E}\exp\left\{\frac{1}{2}\inf_{\lambda\ge 0}\left[(\lambda s_{k+1})^2\|\mathbf{X}\|_{\mathsf{F}}^2 - (\lambda s_{k+1})(s_{k+1} - 3\|\mathbf{X}\|_{\mathsf{F}})\right]\right\} \tag{4.8}$$

$$= 2\mathbb{E}\exp\left\{\frac{1}{2}\inf_{\lambda\ge 0}\left[\lambda^2\|\mathbf{X}\|_{\mathsf{F}}^2 - \lambda(s_{k+1} - 3\|\mathbf{X}\|_{\mathsf{F}})\right]\right\}. \tag{4.9}$$

Choosing $\lambda^* = (s_{k+1} - 3\|\mathbf{X}\|_{\mathsf{F}})_+/(2\|\mathbf{X}\|_{\mathsf{F}}^2)$ and upper bounding the infimum over all positive λ's by the value at λ^*, we arrive at the inequality

$$\mathbb{P}\Big(\bigcup_{n\in I_k}\mathcal{B}_n\Big) \le 2\exp\left\{-\frac{1}{8}\Big(\frac{s_{k+1}}{\|\mathbf{X}\|_{\mathsf{F}}} - 3\Big)_+^2\right\} \le 2\exp\left\{-\frac{1}{8}\Big(\frac{s_{k+1}}{B\sqrt{n_{k+1}}} - 3\Big)_+^2\right\}.$$

In the last step above we have used the inequality $\|\mathbf{X}\|_{\mathsf{F}} \le \sqrt{n_{k+1}}B$. Replacing s_{k+1} and t_{k+1} by their expressions, we get

$$\begin{aligned}\frac{1}{8}\Big(\frac{s_{k+1}}{B\sqrt{n_{k+1}}} - 3\Big)_+^2 &= \frac{1}{8}\Big(\frac{11 n_{k+1}t_{k+1}}{13B\kappa_{n_{k+1}}\sqrt{n_{k+1}}} - 3\Big)_+^2 \\ &= \frac{1}{8}\Big(\frac{11\sqrt{n_{k+1}}t_{k+1}}{13B\kappa_{n_{k+1}}} - 3\Big)_+^2 \\ &= (9/8)\big(\ln\ln n_{k+1} + \ln(50/\delta)\big).\end{aligned}$$

Using the inequality $\ln n_{k+1} \ge \ln(\beta^{k+1}n_0) \ge (k+19)\ln\beta$ we arrive at

$$\mathbb{P}\left(\bigcup_{n\in I_k}\mathcal{B}_n\right) \le \exp\left\{-(9/8)\big(\ln(k+19) + \ln\ln\beta + \ln(50/\delta)\big)\right\} \le \frac{0.21\delta}{(k+19)^{9/8}}.$$

Using the fact that

$$\sum_{k=0}^{\infty}\frac{1}{(k+19)^{9/8}} \le \int_{18}^{\infty} x^{-9/8}\,dx = 8\times 18^{-1/8} \le 5.58,$$

we get $\mathbb{P}\left(\exists n \ge n_0, \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*\|_2 \ge t(n)\right) \le \delta$. To conclude, it suffices to note that $t_{n,\delta}^{\mathrm{MVLIL}} \ge t(n)$ for every $n \in \mathbb{N}^*$.

Extension to sub-Gaussian norm

The boundedness of the features (Assumption 4.3.4) can be relaxed to a sub-Gaussian assumption[4] on the norm of the features, stated as follows :

Assumption 4.7.5. *$\|\boldsymbol{X}_1\|_2$ is σ-sub-Gaussian for some $\sigma > 0$: $\mathbb{E}e^{\|\boldsymbol{X}_1\|_2^2/\sigma^2} \le 2$.*

We now restate Theorem 4.3.5 with this new assumption.

[4] We refer the reader to Vershynin 2018, Section 2.5 for equivalent definitions and details on sub-Gaussian random variables.

Theorem 4.7.6. *Let Assumptions 4.3.1 to 4.3.3 and 4.7.5 be satisfied for every integer $n \in \mathbb{N}^*$. Assume, in addition, that starting from some integer $n_0 \geq 1$, the sequence $\kappa_n^2 \ln\ln n/n$ is decreasing. Then, for any $q \geq 2$ and $\delta \in (0,1)$, there exist positive absolute constants c, c', c'', such that*

$$\mathbb{P}\left(\forall n \geq n_0, \quad \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*\|_q \leq c\kappa_n \sigma \frac{\sqrt{\ln\ln n + \ln(c'/\delta)} + c''}{\sqrt{n}}\right) \geq 1 - \delta.$$

Proof. In the proof we denote by $c, c', c''...$ positive absolute constants, their values may change from one line to another. We use the symbol $\lesssim$ to mean "less than or equal to, up to an absolute constant". Under this new assumption, the proof is exactly the same as for Theorem 4.3.5 up to (4.7). Let k be an integer. Applying Cauchy-Schwarz inequality at this point yields

$$\mathbb{E}\left[\inf_{\lambda>0} e^{2\lambda^2\|\boldsymbol{X}\|_F^2 - \lambda(s_{k+1} - 3\|\boldsymbol{X}\|_F)}\right] \leq \inf_{\lambda>0} e^{-\lambda s_{k+1}} \left[\mathbb{E}e^{4\lambda^2\|\boldsymbol{X}\|_F^2}\right]^{1/2} \left[\mathbb{E}e^{6\lambda\|\boldsymbol{X}\|_F}\right]^{1/2}$$

For $i = 1, \dots, n_{k+1}$, let $z_i := \|\boldsymbol{X}_i\|_2$ and define the random vector $\boldsymbol{z} := (z_1, \dots, z_{n_{k+1}})$. Note that we have $\|\boldsymbol{X}\|_F = \|\boldsymbol{z}\|_2$ and that the coordinates of the random vector $\boldsymbol{z}$ are independent σ-sub-Gaussian random variables. In particular it implies that $\|\boldsymbol{z}\|_2$ is a $\sqrt{n_{k+1}}\sigma$-sub-Gaussian random variable. Indeed, Jensen's inequality applied to the concave map $(x \mapsto x^{1/n_{k+1}})$ gives

$$\mathbb{E}e^{\|\boldsymbol{z}\|_2^2/(\sqrt{n_{k+1}}\sigma)^2} \leq \left(\mathbb{E}\prod_{i=1}^{n_{k+1}} e^{z_i^2/\sigma^2}\right)^{1/n_{k+1}},$$

then the independence of the coordinates assumption allows us to switch the product and the expectation and the resulting quantity is upper bounded by 2 thanks to the sub-Gaussian assumption on the coordinates.

Bounding the first expectation Vershynin 2018, Proposition 2.5.2 (iii) gives a bound on the moment-generating function of a squared sub-Gaussian random variable. Using the independence assumption to switch the product and the expectatione before applying this proposition to each squared coordinate z_i^2 independently, we get, for any $|\lambda| \lesssim \sigma^{-1}$,

$$\mathbb{E}e^{2\lambda^2\|\boldsymbol{z}\|_2^2} = \prod_{i=1}^{n_{k+1}} \left(\mathbb{E}e^{2\lambda^2 z_i^2}\right)^{n_{k+1}} \leq e^{cn_{k+1}\sigma^2\lambda^2}.$$

where c is an absolute constant.

Bounding the second expectation The centering lemma Vershynin 2018, Lemma 2.6.8 states that if a random variable is sub-Gaussian, then its centered version is sub-Gaussian with same sub-Gaussian variance proxy, up to an absolute constant. Therefore the centered random variable $\|\boldsymbol{z}\|_2 - \mathbb{E}\|\boldsymbol{z}\|_2$ is a $c\sqrt{n_{k+1}}\sigma$-sub-Gaussian random variable with c an absolute constant. Applying Vershynin 2018, Proposition 2.5.2 (v) to control the moment-generating function of a centered sub-Gaussian random variable, we get

$$\mathbb{E}e^{6\lambda(\|\boldsymbol{z}\|_2 - \mathbb{E}\|\boldsymbol{z}\|_2)} \leq e^{c\lambda^2 n_{k+1}\sigma^2} \tag{4.10}$$

Finally, Jensen's inequality followed by a control on the L_2 norm of a sub-Gaussian random variable (see, *e.g.*, Vershynin 2018, Proposition 2.5.2 (ii)) yield

$$\mathbb{E}\|\boldsymbol{z}\|_2 \leq \sqrt{n_{k+1}}\sqrt{\mathbb{E}z_1^2} \lesssim \sqrt{n_{k+1}}\sigma. \tag{4.11}$$

Combining (4.10) and (4.11), we get

$$\mathbb{E}e^{6\lambda\|\boldsymbol{z}\|_2} \leq e^{6\lambda\mathbb{E}\|\boldsymbol{z}\|_2}\mathbb{E}e^{6\lambda(\|\boldsymbol{z}\|_2-\mathbb{E}\|\boldsymbol{z}\|_2)} \leq e^{c\lambda^2 n_{k+1}\sigma^2+c'\lambda\sqrt{n_{k+1}}\sigma}.$$

Combining the bounds on expectations

$$\mathbb{E}\left[\inf_{\lambda>0} e^{2\lambda^2\|\boldsymbol{X}\|_F^2-\lambda(s_{k+1}-3\|\boldsymbol{X}\|_F)}\right] \leq \inf_{0<\lambda\lesssim\sigma^{-1}} \exp\{cn_{k+1}\sigma^2\lambda^2 + c'\lambda\sqrt{n_{k+1}}\sigma - \lambda s_{k+1}\}.$$

The non-negative minimizer of the polynomial inside the exponential is given by $\lambda^* = \frac{(s_{k+1}-c'\sqrt{n_{k+1}}\sigma)_+}{cn\sigma^2}$. It satisfies the upper bound constraint (up to taking bigger constants) : since s_{k+1} is of order $\sqrt{n_{k+1}}\sigma$, λ^* is of order $(\sqrt{n_{k+1}}\sigma)^{-1}$. This yields the following upper bound, valid for any integer k,

$$\mathbb{P}\left(\bigcup_{n\in I_k} B_n\right) \leq 2\exp\left\{-\left(\frac{s_{k+1}}{c\sqrt{n_{k+1}}\sigma} - c'\right)_+^2\right\}.$$

Note that, replacing σ by B, it is the same upper bound we get in the bounded case, up to absolute constants. Recalling that $s_{k+1} = c\frac{n_{k+1}}{\kappa_{n_{k+1}}}t_{k+1}$, we have

$$\mathbb{P}\left(\bigcup_{n\in I_k} B_n\right) \leq 2\exp\left\{-\left(c\frac{\sqrt{n_{k+1}}t_{k+1}}{\sigma\kappa_{n_{k+1}}} - c'\right)_+^2\right\}.$$

Letting

$$t(n) = c\kappa_n\sigma\frac{\sqrt{\ln\ln n + \ln(c'/\delta)} + c''}{n},$$

for some suitable absolute positive constants c, c' and c'', we get

$$\mathbb{P}\left(\bigcup_{n\in I_k} B_n\right) \lesssim \frac{\delta}{(k+c')^\gamma},$$

with $\gamma > 1$ an absolute constant. If the absolute constants are set suitably we finally obtain

$$\mathbb{P}\left(\exists n \geq n_0, \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*\|_2 \geq t(n)\right) \leq \delta.$$

□

4.7.4 Proof of Theorem 4.4.1

In this section, we provide the proof of the upper bound established for the proposed algorithm in the problem of best arm identification for multi-armed bandit. We start with two technical lemmas, then we provide two other lemmas that constitute the core technical part of the proof of Theorem 4.4.1. Finally, in Section 4.7.4, we put all the pieces together and present the proof of the theorem.

Preliminary lemmas

We state and prove two elementary lemmas which we will need for the proof of Theorem 4.4.1.

Lemma 4.7.7. *For $t \geq 2, c > 0$ and $0 < \omega \leq 0.15$, we have*

$$\frac{1}{t}\ln\left(\frac{\ln t}{\omega}\right) \geq c \implies t \leq \frac{1}{c}\ln\left(\frac{2\ln(1/(2c\omega))}{\omega}\right).$$

Proof. Let $f(t) = \frac{1}{t}\ln\left(\frac{\ln t}{\omega}\right)$, defined for any $t \geq 2$ and $t_* = \frac{1}{c}\ln\left(\frac{2\ln(1/(2c\omega))}{\omega}\right)$. It suffices to show that $f(t_*) \leq c$. Indeed, since the function f is decreasing, it implies that $f(t) < c$ for any $t > t_*$ which is the contrapositive of the claimed implication. Using the definition of f and t_* we have,

$$\begin{aligned} f(t_*) \leq c &\iff \ln\left(\frac{\ln(t_*)}{\omega}\right) \leq t_* c \\ &\iff t_* \leq \frac{1}{(2c\omega)^2} \\ &\iff \ln\left(\frac{2\ln(1/(2c\omega))}{\omega}\right) \leq \frac{1}{4c\omega^2}. \end{aligned}$$

The last inequality is clearly true since $\ln(x) \leq \frac{x}{2}$ on $(0, \infty)$ and this proves our claim. □

Lemma 4.7.8. *For $t \geq 2$, $s \geq e$, $c \in (0, 1]$, $0 < \omega \leq \delta \leq e^{-e}/2$, we have,*

$$\frac{1}{t}\ln\left(\frac{\ln t}{\omega}\right) \geq \frac{c}{s}\ln\left(\frac{\ln s}{\delta}\right) \implies t \leq \frac{s}{c}\frac{\ln(2/\omega) + \ln\ln(1/2c\omega)}{\ln(1/\delta)}.$$

Proof. Lemma 4.7.7 immediately implies that

$$\frac{ct}{s} \leq \frac{\ln(2/\omega) + \ln\left[\ln s + \ln(1/2c\omega) - \ln\ln(\ln s/\delta)\right]}{\ln(1/\delta) + \ln\ln(s)}.$$

Using the fact that $\ln\ln(\ln s/\delta) \geq 1$ and the following fact

$$\begin{aligned} s \geq e &\implies \ln s - 1 \geq 0 \\ &\implies \ln s - 1 \leq e(\ln s - 1) \\ &\implies \ln s - 1 \leq (\ln s - 1)\ln(1/2c\omega) \\ &\implies \ln s + \ln(1/2c\omega) - 1 \leq \ln s \ln(1/c\omega) \\ &\implies \ln s + \ln(1/2c\omega) - \ln\ln(\ln s/\delta) \leq \ln s \ln(1/2c\omega), \end{aligned}$$

we have

$$\frac{ct}{s} \leq \frac{\ln(2/\omega) + \ln\ln(1/2c\omega) + \ln\ln s}{\ln(1/\delta) + \ln\ln s}.$$

We conclude by applying the inequality $a \geq b, x > 0 \implies \frac{x+a}{x+b} \leq a/b$ with $a = \ln(2/\omega) + \ln\ln(1/2c\omega)$, $b = \ln(1/\delta)$ and $x = \ln\ln s$. □

Main lemmas

Without loss of generality, we assume hereafter that the arms' parameters are ranked in decreasing order: $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_K$. We define the function

$$U(n,\omega) = \frac{4.4\sigma}{\alpha}\sqrt{\frac{1}{n}\ln\left(1 \vee \frac{\ln n}{\omega}\right)}, \quad n \in \mathbb{N}^*,\ \omega \in (0,1),$$

and the events

$$\mathcal{E}_k(\omega) = \{\forall n \geq n_0(\omega) \text{ it holds that } |\widehat{\theta}_{k,n} - \theta_k| \leq U(n,\omega)\},$$

where $n_0(\omega)$ is the smallest integer $n \geq 15$ for which $U(n,\omega) \leq r$. According to Theorem 4.2.4, we have $\mathbb{P}(\mathcal{E}_k(\omega)^\complement) \leq 15\omega$ for every $k \in [K]$, $w \in (0, 0.001)$. The proof of Theorem 4.4.1 is essentially the combination of two lemmas. The first lemma states that with high probability the number of times each sub-optimal arm is pulled is not too large. The second lemma shows that the algorithm indeed stops at some time and returns the best arm with high probability.

Lemma 4.7.9. *Let $\beta \in (0, 4.8)$, $\delta \in (0, 0.001)$ and $\varkappa = (2+\beta)^2(4.4\sigma/\alpha)^2$. For every $n \geq 1$, with probability at least $1 - 16\delta$,*

$$\sum_{k=2}^{K} T_k(n) \leq n_0(\delta)(K-1) + 150\varkappa \mathbf{H}_1 \ln(1/\delta) + \sum_{k=2}^{K} \varkappa \frac{\ln(2\max\{1, \ln(\varkappa/(2\Delta_k^2\delta))\})}{\Delta_k^2}$$

Proof. The proof is carried out in two steps. In the first step, we upper bound the number of pulls on events for which the rewards are well behaved. In the second step we resort to standard concentration arguments to show that the events considered in the first step happen with high probability.

Step 1. Let $k > 1$, $\omega \in (0,1)$ and $\mathcal{E}(n,k,\delta,\omega) = \mathcal{E}_1(\delta) \cap \mathcal{E}_k(\omega) \cap \{I_n = k\}$. Throughout this step, we assume that $\mathcal{E}(n,k,\delta,\omega)$ holds true and that $n \geq Kn_0(\delta)$ (*i.e.*, the warm-up stage is over). This yields

$$\begin{aligned}
\theta_k + U(T_k(n),\omega) + (1+\beta)U(T_k(n),\delta) &\geq \widehat{\theta}_{k,T_k(n)} + (1+\beta)U(T_k(n),\delta) && (\mathcal{E}_k(\omega) \text{ holds})\\
&\geq \widehat{\theta}_{1,T_1(n)} + (1+\beta)U(T_1(n),\delta) && (I_n = k)\\
&\geq \theta_1. && (\mathcal{E}_1(\delta) \text{ holds})
\end{aligned}$$

Since the function U is decreasing in its second argument, we deduce from the last inequality that

$$\begin{aligned}
\Delta_k \coloneqq \theta_1 - \theta_k &\leq (2+\beta)\max\{U(T_k(n),\omega), U(T_k(n),\delta)\}\\
&\leq (2+\beta)U(T_k(n), \min(\omega,\delta)).
\end{aligned}$$

For $\varkappa = (2+\beta)^2(4.4\sigma/\alpha)^2$ and $c = \Delta_k^2/\varkappa$, Lemma 4.7.7 implies that

$$\begin{aligned}
T_k(n) &\le \frac{\varkappa}{\Delta_k^2}\ln\left(\frac{2\ln\left(\varkappa/\left(2\Delta_k^2\min(\omega,\delta)\right)\right)}{\min(\omega,\delta)}\right)\\
&= \frac{\varkappa}{\Delta_k^2}\left(\ln\left(\frac{2}{\delta}\right)+\ln\left(\frac{\ln(\varkappa/2\Delta_k^2\delta)+\ln(1/\omega)-\ln\min(1/\delta,1/\omega)}{(1+\ln(1/\omega))\min(1/\delta,1/\omega)}\right)+\ln\left(\frac{1+\ln(1/\omega)}{\omega}\right)\right)\\
&\le \frac{\varkappa}{\Delta_k^2}\ln\left(\frac{2}{\delta}\right)+\frac{\varkappa}{\Delta_k^2}\ln\left(\frac{\ln(\varkappa/2\Delta_k^2\delta)+\ln(1/\omega)}{1+\ln(1/\omega)}\right)+\frac{\varkappa}{\Delta_k^2}\ln\left(\frac{1+\ln(1/\omega)}{\omega}\right)\\
&\le \tau_k+\frac{\varkappa}{\Delta_k^2}\ln\left(\frac{1+\ln(1/\omega)}{\omega}\right)\\
&\le \tau_k+\frac{2\varkappa}{\Delta_k^2}\ln(1/\omega)\,.
\end{aligned}$$

with $\tau_k = \frac{\varkappa}{\Delta_k^2}\ln\left((2/\delta)\max\{1,\ln(\varkappa/2\Delta_k^2\delta)\}\right)$. Since $T_k(n)$ increases only when k is pulled, the above argument shows that the following inequality is true for any time $n \ge 1$:

$$T_k(n)\mathbb{1}\{\mathcal{E}_1(\delta)\cap\mathcal{E}_k(\omega)\} \le n_0(\delta)+\tau_k+\frac{2\varkappa}{\Delta_k^2}\ln(1/\omega)\,. \tag{4.12}$$

Indeed, let $m_n = \max\{m \le n : I_m = k\}$ be the last time the arm k is pulled among first n rounds. If $m_n > Kn_0(\delta)$ then

$$T_k(n)\mathbb{1}\{\mathcal{E}_1(\delta)\cap\mathcal{E}_k(\omega)\} = T_k(m_n)\mathbb{1}\{\mathcal{E}(m_n,k,\delta,\omega)\} \le \tau_k+\frac{2\varkappa}{\Delta_k^2}\ln(1/\omega).$$

Otherwise, $m_n \le Kn_0(\delta)$, which means that the arm k has not been pulled after the warm-up stage. Therefore,

$$T_k(n)\mathbb{1}\{\mathcal{E}_1(\delta)\cap\mathcal{E}_k(\omega)\} = T_k(Kn_0)\mathbb{1}\{\mathcal{E}_1(\delta)\cap\mathcal{E}_k(\omega)\} \le n_0(\delta) \le n_0(\delta)+\tau_k+\frac{2\varkappa}{\Delta_k^2}\ln(1/\omega)\,.$$

Step 2. We define the random variable $\Omega_k := \max\{\omega \in [0, 0.001] : \mathcal{E}_k(\omega) \text{ holds true}\}$. Theorem 4.2.4 guarantees that it is well defined and that $\mathbb{P}(\Omega_k < \omega) = \mathbb{P}(\mathcal{E}_k(\omega) \text{ is wrong}) \le c\omega$ with $c = 15$. Furthermore, one can rewrite Eq. (4.12) as

$$T_k(n)\mathbb{1}\{\mathcal{E}_1(\delta)\} \le n_0(\delta)+\tau_k+\frac{2\varkappa}{\Delta_k^2}\ln(1/\Omega_k)\,.$$

Therefore, for any $x > 0$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=2}^K T_k(n) > x+\sum_{k=2}^K(\tau_k+n_0(\delta))\right) &\le \mathbb{P}\left(\mathcal{E}_1(\delta)^\complement\right)\\
&\quad+\mathbb{P}\left(\left\{\sum_{k=2}^K T_k(n) > x+\sum_{k=2}^K(\tau_k+n_0(\delta))\right\}\bigcap\mathcal{E}_1(\delta)\right)\\
&\le c\delta+\mathbb{P}\left(\sum_{k=2}^K\frac{2\varkappa}{\Delta_k^2}\ln(1/\Omega_k) > x\right).
\end{aligned}$$

Define the random variables $Z_k = \frac{2\varkappa}{\Delta_k^2}\ln(1/\Omega_k)$, for $k \in [K]\backslash\{1\}$. Observe that these are independent non-negative random variables and since $\mathbb{P}(\Omega_k < \omega) \le c\omega$, it holds that

$$\mathbb{P}(Z_k > x) = \mathbb{P}(\Omega_k < \exp\{-x\Delta_k^2/(2\varkappa)\}) \le c\exp(-x/a_k),$$

with $a_k = 2\varkappa/\Delta_k^2$ for every $x \ge 3a_k \ln 10$. Observing that

$$\mathbb{E}Z_k = \int_0^{+\infty} \mathbb{P}\left(Z_k > x\right) dx \le 3a_k \ln 10 + c\int_{3a_k \ln 10}^{+\infty} e^{-x/a_k}\, dx \le 0.5ca_k$$

and applying a basic concentration inequality for the sum of sub-exponential random variables (see Lemma 4.7.12), we have,

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=2}^{K}(Z_k - 0.5ca_k) > z\right) &\le \mathbb{P}\left(\sum_{k=2}^{K}(Z_k - \mathbb{E}Z_k) > z\right) \\
&\le \exp\left(-\min\left\{\frac{z^2}{8c\|a\|_2^2}, \frac{z}{4\|a\|_\infty}\right\}\right) \\
&\le \exp\left(-\min\left\{\frac{z^2}{8c\|a\|_1^2}, \frac{z}{4\|a\|_1}\right\}\right).
\end{aligned}$$

Putting everything together with $z = 4c\|a\|_1 \ln(1/\delta)$, $x = z + 0.5c\|a\|_1$ one obtains, for $n \ge 1$

$$\mathbb{P}\left(\sum_{k=2}^{K} T_k(n) > \sum_{k=2}^{K}\left(\frac{10\varkappa c\ln(1/\delta)}{\Delta_k^2} + \tau_k + n_0(\delta)\right)\right) \le 16\delta$$

and the claim of the lemma follows. □

Lemma 4.7.10. *Let $\beta \in (0, 4.8), \delta \in (0, 0.001)$ and $c_\beta = \left(\frac{2+\beta}{\beta}\right)^2$. If*

$$\lambda \ge \frac{\varrho}{1 - 15\delta - \sqrt{\delta^{1/4}\ln(1/\delta)}}, \qquad \textit{with} \qquad \varrho = c_\beta \frac{\ln\left(2\ln(c_\beta/2\delta)/\delta\right)}{\ln(1/\delta)},$$

then, for all $k = 2, \dots, K$ and $n = 1, 2, \dots$ we have $T_k(n) < n_0(\delta) + \lambda\sum_{\ell\neq k} T_\ell(n)$ with probability at least $1 - 6\sqrt{\delta}$.

Proof. Let $k > \ell$. Assuming that $\mathcal{E}_k(\omega)$ and $\mathcal{E}_\ell(\delta)$ hold true and that $I_n = k$, one has, for $n \ge Kn_0(\delta)$,

$$\begin{aligned}
\theta_k + U(T_k(n), \omega) + (1+\beta)U(T_k(n), \delta) &\ge \widehat{\theta}_{k, T_k(n)} + (1+\beta)U(T_k(n), \delta) \\
&\ge \widehat{\theta}_{\ell, T_\ell(n)} + (1+\beta)U(T_\ell(n), \delta) \\
&\ge \theta_\ell + \beta U(T_\ell(n), \delta).
\end{aligned}$$

This implies $(2+\beta)U(T_k(n), \min(\omega, \delta)) \ge \beta U(T_\ell(n), \delta)$. Applying Lemma 4.7.8 with $c = c_\beta^{-1}$ one obtains that if $\mathcal{E}_k(\omega)$ and $\mathcal{E}_\ell(\delta)$ hold true and $I_n = k$ then

$$T_k(n) \le c_\beta \frac{\ln\left(2\ln(c_\beta/2\min(\omega,\delta))/\min(\omega,\delta)\right)}{\ln(1/\delta)} T_\ell(n). \tag{4.13}$$

Since $T_k(n)$ only increases when k is played, then, for all $n \geq 1$,

$$(T_k(n) - n_0(\delta))\mathbb{1}\left(\mathcal{E}_k(\omega) \cap \mathcal{E}_\ell(\delta)\right) \leq c_\beta \frac{\ln\left(2\ln(c_\beta/2\min(\omega,\delta))/\min(\omega,\delta)\right)}{\ln(1/\delta)} T_\ell(n).$$

Using (4.13) with $\omega = \delta^{k-1}$ we see that

$$\mathbb{1}\{\mathcal{E}_k(\delta^{k-1})\}\frac{1}{k-1}\sum_{\ell=1}^{k-1}\mathbb{1}\{\mathcal{E}_\ell(\delta)\} > 1-\alpha \implies (1-\alpha)(T_k(n) - n_0(\delta)) \leq \varrho \sum_{\ell \neq k} T_\ell(n).$$

The above implication leads to the following inequalities

$$\begin{aligned}
&\mathbb{P}\Big(\exists(k,n) \in \{2,\dots,K\} \times \mathbb{N}^* : (1-\alpha)(T_k(n) - n_0(\delta)) \geq \varrho \sum_{\ell \neq k} T_\ell(n)\Big) \\
&\quad \leq \mathbb{P}\Big(\exists k \in \{2,\dots,K\} : \mathbb{1}\{\mathcal{E}_k(\delta^{k-1})\}\frac{1}{k-1}\sum_{\ell=1}^{k-1}\mathbb{1}\{\mathcal{E}_\ell(\delta)\} \leq 1-\alpha\Big) \\
&\quad \leq \sum_{k=2}^{K}\mathbb{P}\Big(\mathcal{E}_k(\delta^{k-1})^{\complement}\Big) + \sum_{k=2}^{K}\mathbb{P}\Big(\frac{1}{k-1}\sum_{\ell=1}^{k-1}\mathbb{1}\left(\mathcal{E}_\ell(\delta)\right) \leq 1 - c\delta - (\alpha - c\delta)\Big).
\end{aligned}$$

Since $\mathbb{E}\mathbb{1}\left(\mathcal{E}_\ell(\delta)\right) \geq 1 - c\delta$ with $c = 15$, using *separately* a union bound and Hoeffding's inequality, we get

$$\mathbb{P}\Big(\frac{1}{k-1}\sum_{\ell=1}^{k-1}\mathbb{1}\left(\mathcal{E}_\ell(\delta)\right) \leq 1 - c\delta - (\alpha - c\delta)\Big) \leq \min\left(c(k-1)\delta, \exp(-2(k-1)(\alpha - c\delta)^2\right).$$

Define $R = e^{-2\delta^{1/4}\ln(1/\delta)}$ and $j = \lceil \ln\{2\delta^{3/4}(1-R)\}/\ln R \rceil$. One can check that

$$1 - R = 1 - e^{2\delta^{1/4}\ln\delta} \geq 0.64\delta^{1/4}\ln(1/\delta),$$

which leads to

$$j - 1 \leq -\frac{\ln\{2\delta^{3/4}(1-R)\}}{2\delta^{1/4}\ln(1/\delta)} \leq -\frac{\ln\{1.28\delta\ln(1/\delta)\}}{2\delta^{1/8}\ln(1/\delta)} \leq (1/2)\delta^{-1/4}.$$

Setting $\alpha = c\delta + \sqrt{\delta^{1/4}\ln(1/\delta)}$, we have

$$\begin{aligned}
&\mathbb{P}\Big(\exists(k,n) \in \{2,\dots,K\} \times \mathbb{N}^* : \big(1 - c\delta - \sqrt{\delta^{1/4}\ln(1/\delta)}\big)(T_k(n) - n_0(\delta)) \geq \varrho \sum_{\ell \neq k} T_\ell(n)\Big) \\
&\leq \sum_{k=2}^{K}\Big\{c\delta^{k-1} + \min\left(c(k-1)\delta, e^{-2(k-1)\delta^{1/4}\ln(1/\delta)}\right)\Big\} \\
&\leq c\frac{\delta}{1-\delta} + \frac{c\delta}{2}j^2 + \frac{R^j}{1-R} \leq 15.2\delta + 7.5\delta j^2 + 2\delta^{3/4} \leq 6\sqrt{\delta}.
\end{aligned}$$

This completes the proof of the lemma. □

Putting all lemmas together

Let ν be the confidence level from Theorem 4.4.1 and let δ satisfy the relation $\nu = 16\delta + 6\sqrt{\delta}$. Note that this implies $\sqrt{\delta} = (\sqrt{16\nu + 9} - 3)/16$, which is the value of δ given in Algorithm 1. On the one hand, Lemma 4.7.9 states that, with probability at least $1-16\delta$, the total number of times the suboptimal arms are sampled does not exceed $(K-1)n_0(\delta) + \varkappa\,(150\mathbf{H}_1 \ln(1/\delta) + \mathbf{H}_2)$ where $\varkappa = ((2+\beta)4.4\sigma/\alpha)^2$. On the other hand, Lemma 4.7.10 states that with probability at least $1 - 6\sqrt{\delta}$, if the parameter λ is large enough, only the optimal arm will meet the stopping criterion and therefore, the number of pulls from the optimal arm is equal to $n_0(\delta) + \lambda \sum_{k\geq 2} T_k(n)$. Combining those two lemmas, we have that with probability at least $1 - 16\delta - 6\sqrt{\delta}$, the optimal arm meets the stopping criterion and the total number of pulls does not exceed $(1+\lambda)Kn_0(\delta) + (1+\lambda)\varkappa\,(150\mathbf{H}_1 \ln(1/\delta) + \mathbf{H}_2)$.

4.7.5 Proof of Theorem 4.4.2

Since ϕ_0 is symmetric, the means of the two arms θ_1 and θ_2 coincide with the parameters of interest and so, the gap Δ coincides with the difference in means, i.e., $\Delta = |\theta_1 - \theta_2|$. Therefore, finding the best arm amounts to finding the arm with the best mean and the result is equivalent to Jamieson et al. 2014, Corollary 1, which in turn is a consequence of the following result by Farrell (1964).

Theorem 4.7.11. *Farrell 1964, Theorem 1 Let $X_1, X_2, \ldots$ be i.i.d. Gaussian random variables with unknown mean $\Delta \neq 0$ and variance 1. Consider testing whether $\Delta > 0$ or $\Delta < 0$. Let $Y \in \{-1, 1\}$ be the decision of any such test based on T samples (possibly a random number) and let $\delta \in (0, 1/2)$. If $\sup_{\Delta\neq 0} \mathbb{P}\,(Y \neq \mathit{sign}(\Delta)) \leq \delta$, then*

$$\limsup_{\Delta\to 0} \frac{\mathbb{E}_\delta[T]}{\delta^{-2} \ln\ln \Delta^{-2}} \geq 2 - 4\delta.$$

4.7.6 Proofs of postponed lemmas

Proof of Lemma 4.7.2 Let $k \geq 1, n \in I_k$ and define

$$\boldsymbol{v}_n^* := \frac{\boldsymbol{\theta}_n^* - \widehat{\boldsymbol{\theta}}_n}{\|\boldsymbol{\theta}_n^* - \widehat{\boldsymbol{\theta}}_n\|_2} \in \mathbb{S}^{d-1}, \quad \bar{\boldsymbol{\theta}}_n := \boldsymbol{\theta}_n^* - t_{k+1}\boldsymbol{v}_n^* \quad \text{and} \quad p_n := \frac{t_{k+1}}{\|\boldsymbol{\theta}_n^* - \widehat{\boldsymbol{\theta}}_n\|_2}.$$

Simple algebra yields

$$\bar{\boldsymbol{\theta}}_n = p_n \widehat{\boldsymbol{\theta}}_n + (1 - p_n)\boldsymbol{\theta}_n^*.$$

Furthermore, since the sequence $(t(n))$ is non-increasing on each interval I_k,

$$\|\boldsymbol{\theta}_n^* - \widehat{\boldsymbol{\theta}}_n\|_2 > t(n) \implies p_n \in (0, 1).$$

Therefore, on the event $\mathcal{A}_n$, by convexity of $\widehat{\Phi}_n$,

$$\inf_{\boldsymbol{w}\in t_{k+1}\mathbb{S}^{d-1}} \widehat{\Phi}_n(\boldsymbol{\theta}_n^* - \boldsymbol{w}) \leq \widehat{\Phi}_n(\bar{\boldsymbol{\theta}}_n) \leq (1 - p_n)\widehat{\Phi}_n(\boldsymbol{\theta}_n^*) + p_n\widehat{\Phi}_n(\widehat{\boldsymbol{\theta}}_n) \leq \widehat{\Phi}_n(\boldsymbol{\theta}_n^*).$$

Finally, after a centering step, the curvature of the population risk yields the stated result. □

Proof of Lemma 4.7.3 A modified version[5] of the symmetrization inequality yields

$$\mathbb{E}\left[\sup_{\boldsymbol{w}\in t\mathbb{S}^{d-1}} \exp\left\{\lambda\left(S_m(\boldsymbol{w}) - \alpha m t^2\right)\right\}\right] \leq \mathbb{E}\left[\sup_{\boldsymbol{w}\in t\mathbb{S}^{d-1}} \exp\left\{2\lambda(S'_m(\boldsymbol{w}) - \alpha m t^2)\right\}\right],$$

where $S'_m(\boldsymbol{w})$ is the symmetrized version of $S_m(\boldsymbol{w})$, defined by

$$S'_m(\boldsymbol{w}) = \sum_{i=1}^{m} \varepsilon_i \left\{\phi(Y_i, \boldsymbol{X}_i^\top \boldsymbol{\theta}^*) - \phi(Y_i, \boldsymbol{X}_i^\top(\boldsymbol{\theta}^* - \boldsymbol{w}))\right\}.$$

We define the set $R = \left\{t\mathbf{X}^\top \boldsymbol{v} : \boldsymbol{v} \in \mathbb{S}^{d-1}\right\} \subset \mathbb{R}^m$ and the functions $\varphi_i : \mathbb{R} \to \mathbb{R}$ by

$$\varphi_i : r \mapsto \left[\phi(Y_i, \boldsymbol{X}_i^\top \boldsymbol{\theta}^*) - \phi(Y_i, \boldsymbol{X}_i^\top \boldsymbol{\theta}^* - r)\right]/L, \quad i = 1, \dots, m.$$

These functions φ_i are contractions (Assumption 4.3.1) such that $\varphi_i(0) = 0$. The contraction principle Koltchinskii 2011a, Theorem 2.2 gives

$$\mathbb{E}\left[\sup_{\boldsymbol{w}\in t\mathbb{S}^{d-1}} \exp\left\{2\lambda(S'_m(\boldsymbol{w}) - \alpha m t^2)\right\}\right] \leq \mathbb{E}\left[\sup_{\boldsymbol{w}\in t\mathbb{S}^{d-1}} \exp\left\{2\lambda(L\boldsymbol{w}^\top \mathbf{X}\boldsymbol{\varepsilon} - \alpha m t^2)\right\}\right].$$

Setting $t' = (2m\alpha/L)t$ and $\lambda' = (L^2/m\alpha)\lambda$, we arrive at

$$\mathbb{E}\left[\sup_{\boldsymbol{w}\in t\mathbb{S}^{d-1}} \exp\left\{2\lambda(S'_m(\boldsymbol{w}) - \alpha m t^2)\right\}\right] \leq \mathbb{E}\left[\sup_{w\in t'\mathbb{S}^{d-1}} \exp\left\{\lambda'(\boldsymbol{w}^\top \mathbf{X}\boldsymbol{\varepsilon} - (t')^2/2)\right\}\right].$$

Finally, since the positive real numbers λ and λ' are positively proportional, taking the infimum over all positive λ is exactly the same as taking the infimum over all positive λ'. □

Proof of Lemma 4.7.4 Note that the following inequality is always true,

$$\mathbb{E}e^{\|\mathbf{X}\boldsymbol{\varepsilon}\|_2} \leq e^{\|\mathbf{X}\|_F}\mathbb{E}\left[e^{(\|\mathbf{X}\boldsymbol{\varepsilon}\|_2 - \|\mathbf{X}\|_F)_+}\right].$$

Using the fact that for a non-negative random variable η, $\mathbb{E}\eta = \int_0^{+\infty} P(\eta > t)dt$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{(\|\mathbf{X}\boldsymbol{\varepsilon}\|_2 - \|\mathbf{X}\|_F)_+} - 1\right] &= \int_0^{+\infty} \mathbb{P}\left(e^{(\|\mathbf{X}\boldsymbol{\varepsilon}\|_2 - \|\mathbf{X}\|_F)_+} > t + 1\right) dt \\
&= \int_0^{+\infty} \mathbb{P}\left(\|\mathbf{X}\boldsymbol{\varepsilon}\|_2 > \|\mathbf{X}\|_F + \ln(t+1)\right) dt \\
&\leq \int_0^{+\infty} \exp\left(-\frac{\left(\ln(t+1)\right)^2}{2\|\mathbf{X}\|_F^2}\right) dt.
\end{aligned}$$

The last inequality follows from an application of the bounded difference inequality, see Boucheron, Lugosi, and Massart 2013, Theorem 6.2, Example 6.3 for more details. Using the

[5] The version we use here can be found, for instance, in Lecué and Rigollet 2014, Eq. (2.3).

change of variable $u = \ln(t+1) \iff t = e^u - 1$, we get

$$\begin{aligned}\int_0^{+\infty} \exp\left(-\frac{(\ln(t+1))^2}{2\|\mathbf{X}\|_F^2}\right) dt &= \int_0^{+\infty} \exp\left(u - \frac{u^2}{2\|\mathbf{X}\|_F^2}\right) du \\ &= \int_0^{+\infty} \exp\left(-\frac{1}{2\|\mathbf{X}\|_F^2}(u - \|\mathbf{X}\|_F^2)^2 + \frac{\|\mathbf{X}\|_F^2}{2}\right) du \\ &\leq \sqrt{2\pi}\|\mathbf{X}\|_F \exp\left(\|\mathbf{X}\|_F^2/2\right) F_{\mathcal{N}(0,1)}(\|\mathbf{X}\|_F),\end{aligned}$$

where $F_{\mathcal{N}(0,1)}$ is the cdf of the standard normal distribution. Therefore,

$$\begin{aligned}\mathbb{E}e^{\|\mathbf{X}\boldsymbol{\varepsilon}\|_2} &\leq e^{\|\mathbf{X}\|_F}\left(1 + \sqrt{2\pi}\|\mathbf{X}\|_F \exp\left(\|\mathbf{X}\|_F^2/2\right) F_{\mathcal{N}(0,1)}(\|\mathbf{X}\|_F)\right) \\ &\leq \exp\left\{(\|\mathbf{X}\|_F^2 + 3\|\mathbf{X}\|_F)/2\right\} \sup_{y\geq 0}\left(e^{-(y+y^2)/2} + \sqrt{2\pi}\, y e^{-y/2} F_{\mathcal{N}(0,1)}(y)\right) \\ &\leq 1.86 \exp\left\{(\|\mathbf{X}\|_F^2 + 3\|\mathbf{X}\|_F)/2\right\}.\end{aligned}$$

This completes the proof of the lemma. □

Bounding the sum of random variables with sub-exponential right tails

Lemma 4.7.12. *Let $X_1, \ldots, X_n$ be independent, non-negative, random variables such that there exist positive constants c and $a_1, \ldots, a_n$ satisfying*

$$\mathbb{P}(X_i > x) \leq c e^{-x/a_i}, \qquad \forall\, x > 0, \qquad i = 1, \ldots, n.$$

Then, for any real positive t,

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) > t\right) \leq \exp\left(-\min\left(\frac{t^2}{8\|\boldsymbol{a}\|_2^2}, \frac{t}{4\|\boldsymbol{a}\|_\infty}\right)\right).$$

Proof Defining $\psi_i(\lambda) \coloneqq \log \mathbb{E}e^{\lambda(X_i - \mathbb{E}X_i)}$, $i = 1, \ldots, n$, Markov inequality and the independence hypothesis give

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) > t\right) \leq \inf_{\lambda > 0} e^{-\lambda t} \prod_{i=1}^n e^{\psi_i(\lambda)}. \tag{4.14}$$

Using the inequality $\ln u \leq u - 1$ valid for any positive real u, we have

$$\psi_i(\lambda) \coloneqq \ln \mathbb{E}e^{\lambda X_i} - \lambda \mathbb{E}X_i \leq \mathbb{E}\left[e^{\lambda X_i} - \lambda X_i - 1\right].$$

Let $\phi(u) = e^u - u - 1$. The monotone convergence theorem guarantees that for any $\lambda > 0$,

$$\mathbb{E}\phi(\lambda X_i) = \sum_{p\geq 2} \frac{\lambda^p}{p!} \mathbb{E}X_i^p.$$

Since the X_i's are non-negative, we have, for any integer $p \geq 2$ and for any index $i = 1, \ldots, n$,

$$\mathbb{E}X_i^p = \int_0^{+\infty} \mathbb{P}\left(X_i > t^{1/p}\right) dt \leq cp \int_0^{+\infty} t^{p-1} e^{-t/a_i} dt = ca_i^p p!.$$

Therefore, for any $\lambda \in (0, 1/2a_i)$

$$\psi_i(\lambda) \leq \mathbb{E}\phi(\lambda X_i) \leq 2c(\lambda a_i)^2. \tag{4.15}$$

Plugging (4.15) into (4.14) yields

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) > t\right) \leq \inf_{\lambda \in (0, 1/2a_i)} \exp\left(2c\|\boldsymbol{a}\|_2^2 \lambda^2 - \lambda t\right).$$

The minimum above is attained in

$$\lambda^* = \min\left(\frac{t}{4c\|\boldsymbol{a}\|_2^2}, \frac{1}{2\|\boldsymbol{a}\|_\infty}\right).$$

This yields the stated upper bound

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) > t\right) \leq \exp\left(-\min\left(\frac{t^2}{8\|\boldsymbol{a}\|_2^2}, \frac{t}{4\|\boldsymbol{a}\|_\infty}\right)\right)$$

and the claim of the lemma follows. □

CHAPTER 5

Bounding the expectation of the supremum of empirical processes indexed by Hölder classes

In this chapter, we provide upper bounds on the expectation of the supremum of empirical processes indexed by Hölder classes of any smoothness and for any distribution supported on a bounded set in $\mathbb{R}^d$. These results can alternatively be seen as non-asymptotic risk bounds, when the unknown distribution is estimated by its empirical counterpart, based on n independent observations, and the error of estimation is quantified by integral probability metrics (IPM). In particular, IPM indexed by Hölder classes are considered and the corresponding rates are derived. These results interpolate between two well-known extreme cases: the rate $n^{-1/d}$ corresponding to the Wassertein-1 distance (the least smooth case) and the fast rate $n^{-1/2}$ corresponding to very smooth functions (for instance, functions from a RKHS defined by a bounded kernel).

Based on Nicolas Schreuder (2020). "Bounding the expectation of the supremum of empirical processes indexed by Hölder classes". In: *Mathematical Methods of Statistics* 29, pp. 76–86.

Contents

5.1 Introduction

In many problems of mathematical statistics and learning theory, a crucial step is to understand how well the empirical distribution of a sample approximates the underlying true distribution. The theory of empirical processes is devoted to this question. There are many papers and books treating this and related problems, both from asymptotic and nonasymptotic points of view; see, for instance, Vaart and Wellner (1996) and Barrio, Deheuvels, and Geer (2007). Among many remarkable achievements of the theory of empirical processes, there are two results that have been particularly often evoked and used in the recent literature in statistics and machine learning.

To quickly present these two results, let us give some details on the framework. It is assumed that n independent copies $X_1, \dots, X_n$ of a random variable X taking its values in the d-dimensional hypercube $[0,1]^d$ are observed. The aforementioned two results characterize the order of magnitude of supremum of the empirical process $\mathbb{X}_n(f) = \frac{1}{n}\sum_{i=1} f(X_i) - \mathbb{E}[f(X)]$ over some class of functions $\mathcal{F}$. More precisely, the first result established by Dudley 1968 states that $\sup_{f\in \mathsf{Lip}(1)} \mathbb{X}_n(f)$ is of order $O(n^{-1/d})$,where $\mathsf{Lip}(1)$ is the set of all the Lipschitz-continuous functions with Lipschitz constant 1. The second result Briol et al. 2019, Lemma 1, tells us that if $\mathcal{F}$ contains functions that are smooth enough, for instance functions that are in a finite ball of a RKHS defined by a bounded kernel, then $\sup_{f\in\mathcal{F}} \mathbb{X}_n(f)$ is of order $O(n^{-1/2})$, *i.e.*, the same order as in the case when $\mathcal{F}$ contains only one function.

The main result of this chapter provides an interpolation between the two aforementioned results. Roughly speaking, it shows that if $\mathcal{F}$ is the class of functions defined on $[0,1]^d$ that are Hölder-continuous for a given constant L and a given order $\alpha > 0$, then the supremum of the empirical process over $\mathcal{F}$ is of order $O(n^{-(\frac{\alpha}{d}\wedge\frac{1}{2})})$ with an additional slowly varying factor $\log n$ when $\alpha = d/2$. Clearly, when $\alpha = 1$ this coincides with the result from Dudley (1968), while for $\alpha \geq d/2$ we get the fast and dimension-free rate $n^{-1/2}$, up to a log factor.

The rest of this chapter is organized as follows. We complete this introduction by providing all the important notations used throughout this chapter. Section 5.2 is devoted to presenting and formally defining Hölder classes and Integral Probability Metrics (IPM). In Section 5.3, we expose some important concepts and results from empirical process theory needed for our proofs. We end this chapter by stating our main theorem in Section 5.4. Some extensions are mentioned in Section 5.5. The proofs are postponed to the appendix.

Notations

A multi-index $\boldsymbol{k}$ is a vector with integer coordinates $(k_1, \dots, k_d)$. We write $|\boldsymbol{k}| = \sum_{i=1}^d k_i$. For a given multi-index $\boldsymbol{k} = (k_1, \dots, k_d)$, we define the differential operator

$$D^{\boldsymbol{k}} = \frac{\partial^{|\boldsymbol{k}|}}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}.$$

For any positive real number x, $\lfloor x \rfloor$ denotes the largest integer strictly smaller than x. We let $\mathcal{X}$ be a convex bounded set in $\mathbb{R}^d$ with non-empty interior. We assume that all the functions and function classes considered in this chapter are supported on the bounded set $\mathcal{X}$. For any integer k, we denote by $C^k(\mathcal{X}, \mathbb{R})$ the class of real-valued functions with domain $\mathcal{X}$ which are k-times differentiable with continuous k-th differentials. For any real-valued bounded function f on $\mathcal{X}$, we let $\|f\|_\infty \coloneqq \sup_{x \in \mathcal{X}} |f(x)| \in [0, +\infty)$. Note that we can consider the essential supremum instead of the supremum over $\mathcal{X}$ in which case our results would hold almost surely. We let $\|\cdot\|$ denote some norm on $\mathbb{R}^d$. We denote by $\sigma_1, \dots, \sigma_n$ i.i.d. Rademacher random variables, *i.e.*, discrete random variables such that $\mathbb{P}(\sigma_1 = 1) = \mathbb{P}(\sigma_1 = -1) = 1/2$ which are independent of any other source of randomness. We use the convention $1/0 = +\infty$.

5.2 A primer on Hölder classes and integral probability metrics

In this section we define Hölder classes of functions and integral probability metrics. We then discuss some properties of these notions and highlight their role in statistics and statistical learning theory.

5.2.1 Hölder classes

A central problem in nonparametric statistics is to estimate a function belonging to an infinite-dimensional space (*e.g.*, density estimation, regression function estimation, hazard function estimation), see Tsybakov (2008) for an introduction to the topic of nonparametric estimation. To obtain nontrivial rates of convergence, some kind of regularity is assumed on the function of interest. It can be expressed as conditions on the function itself, on its derivatives, on the coefficients of the function in a given basis, etc. Hölder classes are one of the most common classes considered in the nonparametric estimation literature, they form a natural extension of Lipschitz-continuous functions and can be formalised with the following simple conditions. For any real number $\alpha > 0$, we define the Hölder norm of smoothness α of a $\lfloor \alpha \rfloor$-times differentiable function f as

$$\|f\|_{\mathcal{H}^\alpha} \coloneqq \max_{|k| \leq \lfloor \alpha \rfloor} \|D^k f\|_\infty + \max_{|k| = \lfloor \alpha \rfloor} \sup_{x \neq y} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\alpha - \lfloor \alpha \rfloor}}.$$

The Hölder ball of smoothness α and radius $L > 0$, denoted by $\mathcal{H}^\alpha(L)$, is then defined as the class of $\lfloor \alpha \rfloor$-times continuously differentiable functions with Hölder norm bounded by the radius L:

$$\mathcal{H}^\alpha(L) = \left\{ f \in C^{\lfloor \alpha \rfloor}(\mathcal{X}, \mathbb{R}) \mid \|f\|_{\mathcal{H}^\alpha} \leq L \right\}.$$

To get a grasp of why Hölder classes are convenient, let us consider the case $d = 1$. In this setting, one can easily derive an upper bound on the remainder of the best polynomial approximation of any given Hölder function. Indeed, for any positive $\alpha > 0$ with $\lfloor \alpha \rfloor = \ell$, for

any function $f \in \mathcal{H}^\alpha(L)$, Taylor's theorem yields that for any points $x, y \in \mathcal{X}$,

$$\begin{aligned}\left|f(y) - \sum_{k=0}^{\ell} \frac{f^{(k)}(x)}{k!}(y-x)^k\right| &\leq \frac{|y-x|^\ell}{(\ell-1)!}\int_0^1 |f^{(\ell)}(x+t(y-x)) - f^{(\ell)}(x)|(1-t)^\ell dt \\ &\leq L\frac{|y-x|^\alpha}{(\ell-1)!}\int_0^1 t^{\alpha-\ell}(1-t)^\ell dt \\ &\leq L\frac{|y-x|^\alpha}{\ell!}.\end{aligned}$$

Note that this bound holds uniformly over the Hölder ball $\mathcal{H}^\alpha(L)$.

5.2.2 Integral probability metrics

The class $\mathcal{H}^1(1)$ of 1-Lipschitz functions has received a lot of attention in the optimal transport literature; see (Santambrogio 2015) for an overview of the topic of mathematical optimal transport. This interest comes from the Kantorovitch duality, which implies that the Wasserstein-1 distance (also known as the earth mover's distance) can be expressed, for any probability measures P, Q, as a supremum of some functional over 1-Lipschitz functions:

$$W_1(P,Q) = \sup_{f \in \mathcal{H}^1(1)} |\mathbb{E}_{X\sim P} f(X) - \mathbb{E}_{Y\sim Q} f(Y)|.$$

More generally, recall from (IPM) that for a given class $\mathcal{F}$ of bounded functions, one can define a pseudo-metric on the space of probability measures, the integral probability metric (IPM) induced by the class $\mathcal{F}$, as

$$d_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} |\mathbb{E}_{X\sim P} f(X) - \mathbb{E}_{Y\sim Q} f(Y)|.$$

The literature on IPM has recently been boosted by the advent of adversarial generative models (Arjovsky, Chintala, and Bottou 2017; Goodfellow et al. 2014). A reason for this is that an IPM can be seen as an adversarial loss: to compare two probability distributions, it seeks for the function which discriminates the most the two distributions in expectation. Initially studied by the deep learning community, impressive empirical results obtained by adversarial generative models on several tasks such as image generation led statisticians to study it theoretically (Liang 2018; Chen et al. 2020; Briol et al. 2019) (see also Sriperumbudur et al. (2012) for statistical results on IPM in a general framework). Since, as pointed out earlier, Lipschitz functions are also Hölder, one can wonder what happens for IPM indexed by general Hölder classes. Such IPM already appeared in the literature: Scetbon et al. (2020) showed that α-Hölder IPM with smoothness $\alpha \leq 1$ correspond to the cost of a generalized optimal transport problem.

To further motivate our study, let us consider the abstract problem of minimum distance estimation: for a given probability measure P, find a distribution Q in a given set of probability measures $\mathcal{Q}$ such that Q is close to P under the metric $d_{\mathcal{F}}$:

$$\min_{Q\in\mathcal{Q}} d_{\mathcal{F}}(Q,P). \tag{5.1}$$

For example, when $\mathcal{F}$ is taken to be the class of 1-Lipschitz function, this problem is known as minimum Kantorovitch estimation (Bassetti, Bodini, and Regazzini 2006). In statistics, the

probability P is usually unknown and one is only given i.i.d. samples $X_1, \dots, X_n$ from the probability distribution P. A natural strategy is then to employ the empirical distribution $P_n = 1/n \sum_{i=1}^n \delta_{X_i}$ as a proxy for the theoretical distribution and instead of (5.1) solve the problem:

$$\min_{Q \in \mathcal{Q}} d_{\mathcal{F}}(Q, P_n). \tag{5.2}$$

Since the triangle inequality yields

$$|d_{\mathcal{F}}(Q, P) - d_{\mathcal{F}}(Q, P_n)| \leq d_{\mathcal{F}}(P, P_n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X) \right|,$$

one question of interest is to measure how fast the empirical measure approximates the true measure under the IPM $d_{\mathcal{F}}$. If the rates are fast, we do not loose much by considering the empirical problem (5.2) instead of the theoretical one of (5.1). However if the rates are slow, one cannot expect the distances of the solutions to the measure P to be close. We will see in the next section that the latter expression corresponds to the supremum of the empirical process indexed by the class $\mathcal{F}$, it will enable us to leverage the rich literature on empirical processes to obtain rates of convergence for $d_{\mathcal{F}}(P, P_n)$.

5.3 Empirical processes, metric entropy and Dudley's bounds

This section provides a short account of the notions and tools from the theory of empirical processes which are necessary for stating and establishing the main result.

5.3.1 Empirical processes

Empirical process are ubiquitous in statistical learning theory, we refer the reader to Koltchinskii 2011b; Giné and Nickl 2016 for a general presentation of results on empirical processes and their link with statistics and learning theory. For clarity, we begin by recalling the definition of an empirical process.

Definition 5.3.1. *Let $\mathcal{F}$ be a class of real-valued functions $f \colon \mathcal{X} \to \mathbb{R}$, where $(\mathcal{X}, \mathcal{A}, P)$ is a probability space. Let X be a random point in $\mathcal{X}$ distributed according to the distribution P and let $X_1, \dots, X_n$ be independent copies of X. The random process $\left(\mathbb{X}_n(f)\right)_{f \in \mathcal{F}}$ defined by*

$$\mathbb{X}_n(f) \coloneqq \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X)$$

is called an empirical process indexed by $\mathcal{F}$.

In our case, we are interested in controlling the (expectation of the) supremum of an empirical process, a common case in the literature. Most of the time, the first step to apply for achieving this goal is to "symmetrize" the empirical process as allowed by the following lemma. Let $\widehat{R}_n(\mathcal{F})$ be the empirical Rademacher complexity of function class $\mathcal{F}$, defined as

$$\widehat{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \,\middle|\, X_1, \dots, X_n \right].$$

Lemma 5.3.2 (Symmetrization). *For any class $\mathcal{F}$ of P-integrable functions,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\mathbb{X}_n(f)|\right] \leq 2\,\mathbb{E}[\widehat{R}_n(\mathcal{F})].$$

The advantage of Rademacher processes is that, regardless of the distribution of the random variable X and the function class $\mathcal{F}$, for a fixed sample $X_1,\dots,X_n$, the random variable $\sum_{i=1}^n \sigma_i f(X_i)$ has a sub-Gaussian behavior, in the following sense.

Definition 5.3.3 (Sub-Gaussian behavior). *A centered random variable Y has a sub-Gaussian behavior if there exists a positive constant σ such that*

$$\mathbb{E}e^{\lambda Y} \leq e^{\lambda^2\sigma^2/2}, \quad \forall\lambda\in\mathbb{R}.$$

In that case, we define the sub-Gaussian norm[1] *of Y as*

$$\|Y\|_{\psi_2} = \inf\left\{t>0 : \mathbb{E}e^{Y^2/t^2} \leq 2\right\}.$$

Having a sub-Gaussian behavior essentially means to be at least as concentrated as a Gaussian random variable around its mean. Our definition is equivalent to the tail inequalities

$$\mathbb{P}(|Y|>t) \leq 2e^{-t^2/(2\sigma^2)}, \quad \forall t>0.$$

This type of behavior will be crucial to obtain the main result of this note. Indeed, as we will see, the behavior of the supremum of an empirical process (and more generally a stochastic process) which has sub-Gaussian increments exclusively depends on the topology of the space by which the process is indexed.

5.3.2 Metric entropy

Let (T,d) be a totally bounded metric space, *i.e.*, for every real number $\varepsilon>0$, there exists a finite collection of open balls of radius ε whose union contains T. We give a formal definition of such finite collections, see also Figure 5.1 for an illustration.

Definition 5.3.4. *Given $\varepsilon>0$, a subset $T_\varepsilon\subset T$ is called an* ε-cover *of T if for every $t\in T$, there exists $s\in T_\varepsilon$ such that $d(s,t)\leq\varepsilon$.*

Note that adding any point to an ε-cover still yields an ε-cover. Thus we can look for ε-covers of a set with smallest cardinality, which we call covering number.

Definition 5.3.5. *The* ε-covering number *of T, denoted by $\mathcal{N}(T,d,\varepsilon)$, is the cardinality of the smallest ε-cover of T, that is*

$$\mathcal{N}(T,d,\varepsilon) \coloneqq \min\left\{|T_\varepsilon| : T_\varepsilon \text{ is an } \varepsilon\text{-cover of } T\right\}.$$

The metric entropy *of T is given by the logarithm of the ε-covering number.*

[1] See Vershynin 2018, Section 2.5 for the link between definitions of sub-Gaussian random variables (bound on moment-generating function, tail inequalities...) and the Orlicz norm ψ_2.

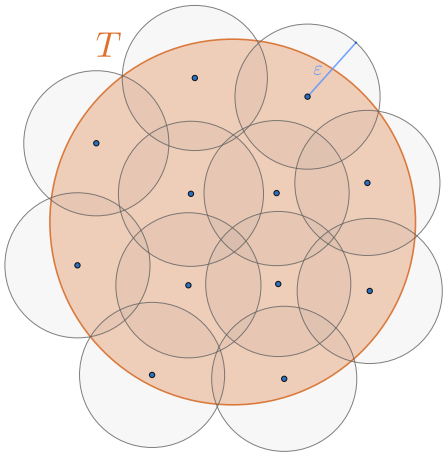


Figure 5.1: Illustration of an ε-cover for some space T.

Remark 5.3.6. *A totally bounded metric space (T, d) is pre-compact in the sense that its closure is compact. The metric entropy (or entropic numbers) of (T, d) can then be seen as some measure of compactness of the space. Indeed, $\mathcal{N}(T, d, \varepsilon)$ quantifies precisely how many balls of radius ε are needed to cover the whole space T.*

Entropic numbers for Hölder classes are known and can be found in *e.g.* Shiryayev (1993) and Vaart and Wellner (1996).

Theorem 5.3.7 (Theorem 2.7.3 in Vaart and Wellner 1996)**.** *Let $\mathcal{X}$ be a bounded, convex subset of $\mathbb{R}^d$ with nonempty interior. There exists a constant $K_{\alpha,d}$ depending only on α and d such that, for every $\varepsilon > 0$,*

$$\log \mathcal{N}(\mathcal{H}^\alpha(1), \|\cdot\|_\infty, \varepsilon) \leq K_{\alpha,d} \lambda_d(\mathcal{X}^1)\, \varepsilon^{-d/\alpha},$$

where λ_d is the d-dimensional Lebesgue measure and $\mathcal{X}^1$ is the 1-blowup of $\mathcal{X}$: $\mathcal{X}^1 = \{y : \inf_{x\in\mathcal{X}} \|y - x\| < 1\}$.

5.3.3 Dudley's bound and its refined version

We now present classic results which show the link between the topology of the indexing set and the behavior of the supremum of the corresponding empirical process. Following Vershynin 2018, Definition 8.1.1, for $K \geq 0$, we say that a random process $(X_t)_{t\in T}$ on a metric space (T, d) has K-sub-Gaussian increments if

$$\|X_t - X_s\|_{\psi_2} \leq K d(t, s), \quad \text{for all} \quad t, s \in T.$$

Theorem 5.3.8 (Dudley's inequality)**.** *Let $(X_t)_{t\in T}$ be a mean-zero random process on a metric space (T, d) with K-sub-Gaussian increments. Then*

$$\mathbb{E}\Big[\sup_{t\in T} X_t\Big] \leq CK \int_0^{+\infty} \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon,$$

for some universal constant $C > 0$.

One drawback of Dudley's bound is that the integral on the right hand side may diverge if the metric entropy of T tends to infinity at a very fast rate when $\varepsilon \to 0$. For example, when the metric entropy is upper bounded by $\varepsilon^{-\gamma}$, as it was seen to be the case with $\gamma = d/\alpha$ for α-Hölder-smooth d-variate functions, the integral converges if and only if $\gamma < 2$.

An improvement of Dudley's bound in the case where the process X_t is a Rademacher average indexed by a class of functions $\mathcal{F}$—circumventing the problem of divergence of the integral—was proposed by Srebro, Sridharan, and Tewari 2010, Lemma A.3 (see also Srebro and Sridharan (2010)). Before stating the theorem, let us recall the definition of the $L_2(P_n)$ norm of a function f:

$$\|f\|^2_{L_2(P_n)} = \int_{\mathcal{X}} f^2 \, dP_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i)^2.$$

Theorem 5.3.9. *Let $\mathcal{F} \subset \{f \colon \mathcal{X} \to \mathbb{R}\}$ be any class of measurable functions containing the uniformly zero function and let $S_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \|f\|_{L_2(P_n)}$. We have*

$$\widehat{R}_n(\mathcal{F}) \leq \inf_{\tau > 0} \left\{ 4\tau + \frac{12}{\sqrt{n}} \int_{\tau}^{S_n(\mathcal{F})} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)} \, d\varepsilon \right\}.$$

Note that the refined Dudley bound gives an upper bound on the empirical Rademacher process and depends on the metric entropy with respect to the empirical norm $L_2(P_n)$. The following simple lemma shows that the $L_2(P_n)$-norm can be replaced by the supremum-norm in the refined Dudley bound.

Lemma 5.3.10. *Let $\mathcal{F}$ be any class of bounded functions defined on $\mathcal{X}$. For any sample $X_1, \ldots, X_n$, let $\mathcal{F}_{|X_1,\ldots,X_n}$ be the subset of $\mathbb{R}^n$ defined by*

$$\mathcal{F}_{|X_1,\ldots,X_n} = \{u \in \mathbb{R}^n : \exists f \in \mathcal{F} \text{ such that } u_i = f(X_i) \text{ for all } i = 1, \ldots, n\}.$$

For any $\varepsilon > 0$, we have

$$\mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon) \leq \mathcal{N}(\mathcal{F}_{|X_1,\ldots,X_n}, \|\cdot\|_\infty, \varepsilon) \leq \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon).$$

Proof. Let $\{u_1, \ldots, u_M\}$ be a minimal ε-net for $\mathcal{F}_{|X_1,\ldots,X_n}$ with respect to the supremum norm. Let $f_1, \ldots, f_M \in \mathcal{F}$ be such that $(f_j(X_1), \ldots, f_j(X_n)) = u_j$ for every $j = 1, \ldots, M$. Then, for any $f \in \mathcal{F}$, there exists an index $j \in [M]$ such that $\max_i |f(X_i) - (u_j)_i| = \max_i |f(X_i) - f_j(X_i)| \leq \varepsilon$. Since for any function f in $\mathcal{F}$,

$$\|f - f_j\|^2_{L_2(P_n)} = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - f_j(X_i))^2 \leq \|f - f_j\|^2_\infty,$$

$\{f_1, \ldots, f_M\}$ is an ε-net for $\mathcal{F}$ with respect to the empirical L_2 norm. This proves the first inequality. Let now $f_1, \ldots, f_M$ be an ε-net of $(\mathcal{F}, \|\cdot\|_\infty)$. One readily checks that $u_1, \ldots, u_M$ defined by $u_j = (f_j(X_1), \ldots, f_j(X_n))$ is an ε-net of $\mathcal{F}_{|X_1,\ldots,X_n}$. This completes the proof. □

5.4 Main result

We are now in a position to state the main theorem which gives, for an IPM defined by a Hölder class, the rate of convergence of the empirical measure towards its theoretical counterpart.

Theorem 5.4.1. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex bounded set with non-empty interior. Let $\mathcal{H}^\alpha(L)$ be the Hölder class of α-smooth functions supported on the set $\mathcal{X}$ and with Hölder norm bounded by L. For any probability distribution P supported on $\mathcal{X}$, denoting by P_n the empirical measure associated to i.i.d. samples $X_1, \dots, X_n \sim P$, we have,*

$$\mathbb{E}\big[d_{\mathcal{H}^\alpha(L)}(P_n, P)\big] = \mathbb{E}\Big[\sup_{h \in \mathcal{H}^\alpha(L)} \big|\mathbb{X}_n(h)\big|\Big] \leq cL \begin{cases} n^{-\alpha/d} & \text{if } \alpha < d/2, \\ n^{-1/2}\ln(n) & \text{if } \alpha = d/2, \\ n^{-1/2} & \text{if } \alpha > d/2, \end{cases} \tag{5.3}$$

where c is a constant depending only on d, $\lambda_d(\mathcal{X}^1)$ and α.

We notice two different regimes: for highly smooth functions ($\alpha > d/2$), the rate of convergence does not depend on the smoothness α nor on the dimension d and corresponds to the usual parametric rate of convergence (note that it also matches the rate known for the Maximum Mean Discrepancy metric, which is an IPM indexed by the unit ball of a RKHS with bounded kernel (Briol et al. 2019)). For less regular Hölder functions ($\alpha < d/2$), the rate of convergence depends both on the smoothness and on the dimension in a typical curse of dimensionality behavior. These two regimes coincide, up to a logarithmic factor, at their smoothness boundary $\alpha = d/2$: we have a continuous transition in terms of the exponent of the sample size. Interestingly the rates we obtain interpolate between the $n^{-1/d}$ rate known for Wasserstein-1 distance (Weed and Bach 2019) when considering $\mathcal{H}^1(1)$ and the $n^{-1/2}$ rate for Maximum Mean Discrepancy when considering Hölder classes with enough smoothness. Those observations are summarised in Figure 5.2.

Finally, let us mention that the formulation of Theorem 5.4.1 given above aims at characterizing the behaviour of the expected error in the asymptotic setting of large samples. This result follows from the following finite sample upper bound (proved in Section 5.6.2):

$$\mathbb{E}\left[d_{\mathcal{H}^\alpha}(P_n, P)\right] \leq 12 \begin{cases} \left(\frac{K\lambda}{n}\right)^{\alpha/d} \left[\frac{d}{d-2\alpha} \wedge \left(1 + 0.5\log\left(\frac{n}{9K\lambda}\right)\right)\right] & \text{if } \alpha < d/2, \\ \left(\frac{K\lambda}{n}\right)^{1/2} \left[\frac{2\alpha}{2\alpha-d} \wedge \left(1 + \frac{\alpha}{d}\log\left(\frac{n}{9K\lambda}\right)\right)\right] & \text{if } \alpha \geq d/2, \end{cases} \tag{5.4}$$

where $\lambda := \lambda_d(\mathcal{X}^1)$ and $K = K_{\alpha,d}$ is the constant depending only on α and d borrowed from Theorem 5.3.7.

5.5 Some extensions

A slightly less precise but more general result can be obtained for any bounded class whose entropy grows polynomially in $1/\varepsilon$; see also Rakhlin, Sridharan, and Tsybakov 2017, Theorem 2, where this condition naturally arises. Such an extension can be stated as follows.

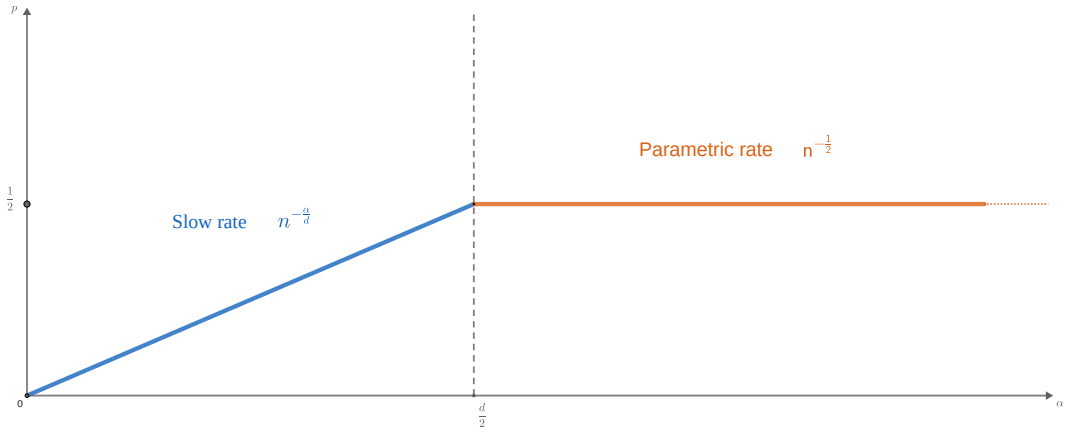


Figure 5.2: Exponent p appearing in the rates of convergence n^{-p} in Theorem 5.4.1 as a function of the smoothness α.

Theorem 5.5.1. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex bounded set with non-empty interior. Let $\mathcal{H}$ be a bounded class of functions supported on the set $\mathcal{X}$. Assume that the entropy of the class grows polynomially, i.e., there exist positive real numbers p and A such that*

$$\forall \varepsilon > 0, \quad \log \mathcal{N}(\mathcal{H}, \|\cdot\|_\infty, \varepsilon) \leq A\varepsilon^{-p}.$$

Then, for any probability distribution P supported on $\mathcal{X}$, denoting by P_n the empirical measure associated to i.i.d. samples $X_1, \ldots, X_n \sim P$, we have,

$$\mathbb{E}\big[d_{\mathcal{H}}(P_n, P)\big] = \mathbb{E}\Big[\sup_{h \in \mathcal{H}} |\mathbb{X}_n(h)|\Big] \leq c \begin{cases} n^{-1/p} & \text{if } p > 2, \\ n^{-1/2} \ln(n) & \text{if } p = 2, \\ n^{-1/2} & \text{if } p < 2, \end{cases} \tag{5.5}$$

where c is a constant.

The proof of the extension is exactly the same as the proof of Theorem 5.4.1 up to constants. In this chapter we have seen Hölder classes as examples of classes with polynomial growth of the entropy but there are many other such classes. To illustrate this we give the example of Sobolev classes which, in some cases, are more general than Hölder classes. For a positive integer s and a real number $1 \leq p \leq +\infty$, define the Sobolev space $\mathcal{W}_p^s(r)$ with radius $r > 0$ as

$$\mathcal{W}_p^s(r) := \left\{ f \in C^s(\mathcal{X}, \mathbb{R}) : \sum_{|k| \leq s} \|D^k f\|_p \leq r \right\}.$$

Note that for any positive integer s and for any positive radius L, there exist radii r and r' such that

$$\mathcal{W}_\infty^s(r) \subset \mathcal{H}^s(L) \subset \mathcal{W}_\infty^{s-1}(r').$$

Birman and Solomyak 1967 Metric entropy bounds for Sobolev function classes on bounded subsets of Euclidean space were initialy obtained by Birman and Solomyak 1967. Nickl and Pötscher 2007 extended those results to more general smoothness classes. In particular, a consequence of Nickl and Pötscher 2007, Corollary 1 is that for any positive integer $s > 0$, and real number p such that $d/s < p \leq +\infty$, the entropy of a Sobolev class grows polynomially as

$$\log \mathcal{N}(\mathcal{W}_p^s(L), \|\cdot\|_\infty, \varepsilon) \leq A\varepsilon^{-d/s},$$

for some positive constant A. Thus Theorem 5.5.1 holds for this class. Finally we point out that such bounds on the entropy hold for more general spaces such as some Besov spaces. We refer the reader to Nickl and Pötscher (2007) for more details.

5.6 Proofs

This section contains the proofs of the main results, Theorems 5.3.9 and 5.4.1, stated in the main body of the note.

5.6.1 Proof of Theorem 5.3.9

The proof of Theorem 5.3.9 can be found in Srebro and Sridharan 2010. We add it here for completeness.

Let $\gamma_0 = S_n(\mathcal{F}) = \sup_{f\in\mathcal{F}}\|f\|_{L_2(P_n)}$. Define $\gamma_j = 2^{-j}\gamma_0$, for every integer $j \in \mathbb{N}$, and let T_j be a minimal γ_j-cover of $\mathcal{F}$ with respect to $L_2(P_n)$. For any function $f \in \mathcal{F}$, we denote by $\widehat{f}_j$ an element of T_j which is an γ_j approximation of f. For any positive integer N we can decompose the function f as

$$f = f - \widehat{f}_N + \sum_{j=1}^{N}(\widehat{f}_j - \widehat{f}_{j-1})$$

where $\widehat{f}_0 = 0 \in \mathcal{F}$. Hence, for any positive integer N, we have

$$\begin{aligned}
\widehat{R}_n(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i\left(f(X_i) - \widehat{f}_N(X_i) + \sum_{j=1}^{N}(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right)\right] \\
&\le \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(f(X_i) - \widehat{f}_N(X_i))\right] + \sum_{j=1}^{N}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right] \\
&\le \frac{1}{n}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}|(f(X_i) - \widehat{f}_N(X_i))| + \sum_{j=1}^{N}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right] \\
&= \sup_{f\in\mathcal{F}}\|f - \widehat{f}_N\|_{L_2(P_n)} + \sum_{j=1}^{N}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right] \\
&\le \gamma_N + \sum_{j=1}^{N}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right].
\end{aligned}$$

For any positive integer j, the triangle inequality gives

$$\|\widehat{f}_j - \widehat{f}_{j-1}\|_{L_2(P_n)} \le \|\widehat{f}_j - f\|_{L_2(P_n)} + \|f - \widehat{f}_{j-1}\|_{L_2(P_n)} \le \gamma_j + \gamma_{j-1} = 3\gamma_j. \tag{5.6}$$

We need the following classic lemma which controls the expectation of a Rademacher average over a finite set[2].

Lemma 5.6.1 (Massart's finite class lemma)**.** *Let $\mathcal{X}$ be a finite subset of $\mathbb{R}^n$ and let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher random variables. Denote the radius of $\mathcal{X}$ by $R = \sup_{x\in\mathcal{X}}\|x\|$. Then, we have,*

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i x_i\right] \le R\frac{\sqrt{2\log|\mathcal{X}|}}{n}.$$

[2]We refer the reader to https://ttic.uchicago.edu/~tewari/lectures/lecture10.pdf for a simple proof of this lemma.

Applying this lemma to $\mathcal{X}_j = \left\{(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))_{i=1}^n \in \mathbb{R}^n : f \in \mathcal{F}\right\}$ for any $j = 1, \dots, n$ and using (5.6), we get

$$\sum_{j=1}^{N} \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i (\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i)) \right] \leq \sum_{j=1}^{N} 3\gamma_j \frac{\sqrt{2 \log(|T_j| \cdot |T_{j-1}|)}}{n}$$

Therefore we have

$$\begin{aligned}
\widehat{R}_n(\mathcal{F}) &\leq \gamma_N + \sum_{j=1}^{N} 3\gamma_j \frac{\sqrt{2 \log(|T_j| \cdot |T_{j-1}|)}}{n} \\
&\leq \gamma_N + \frac{6}{n} \sum_{j=1}^{N} \gamma_j \sqrt{\log |T_j|} \\
&= \gamma_N + \frac{12}{n} \sum_{j=1}^{N} (\gamma_j - \gamma_{j+1}) \sqrt{\log |T_j|} \\
&= \gamma_N + \frac{12}{n} \sum_{j=1}^{N} (\gamma_j - \gamma_{j+1}) \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \gamma_j)} \\
&\leq \gamma_N + \frac{12}{n} \int_{\gamma_{N+1}}^{\gamma_0} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)} d\varepsilon.
\end{aligned}$$

For any $\tau > 0$, pick $N = \sup\{j : \gamma_j > 2\tau\}$. Then $\gamma_N = 2\gamma_{N+1} \leq 4\tau$ and $\gamma_{N+1} = \gamma_N/2 \geq \tau$. Hence, we conclude that

$$\widehat{R}_n(\mathcal{F}) \leq 4\tau + \frac{12}{\sqrt{n}} \int_{\tau}^{\gamma_0} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)}\, d\varepsilon.$$

Since τ can take any positive value we can take the infimum over all positive τ and this concludes the proof.

5.6.2 Proof of Theorem 5.4.1

Without loss of generality, we prove the theorem in the case $L = 1$. The general case will follow by homogeneity. For simplicity we write $\mathcal{H}^\alpha = \mathcal{H}^\alpha(1)$, $Ph = \int_{\mathcal{X}} h\, dP$ and $P_n h = \int_{\mathcal{X}} h\, dP_n$. A symmetrization argument (Lemma 5.3.2) gives

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}^\alpha} |Ph - P_n h| \right] \leq 2\mathbb{E}[\widehat{R}_n(\mathcal{H}^\alpha)],$$

where the empirical Rademacher process $\widehat{R}_n(\mathcal{H}^\alpha)$ is given by

$$\widehat{R}_n(\mathcal{H}^\alpha) = \frac{1}{n} \mathbb{E}\left[\sup_{h \in \mathcal{H}^\alpha} \sum_{i=1}^{n} \sigma_i h(X_i) \middle| X_1, \dots, X_n \right].$$

Noting that, for any $h \in \mathcal{H}^\alpha$,

$$P_n h^2 := \frac{1}{n} \sum_{i=1}^{n} h^2(X_i) \leq \|h^2\|_\infty \leq 1,$$

the improved Dudley bound (Theorem 5.3.9) coupled with Lemma 5.3.10 yields,

$$\begin{aligned}\mathbb{E}\Big[\sup_{h\in\mathcal{H}^\alpha}|P_nh-Ph|\Big] &\le \inf_{\tau>0}\left(4\tau+\frac{12}{\sqrt{n}}\int_\tau^1\sqrt{\log\mathcal{N}(\mathcal{H}^\alpha,\|\cdot\|_\infty,\varepsilon)}d\varepsilon\right)\\ &\le \inf_{\tau>0}\left(4\tau+\frac{12\sqrt{K\lambda_d(\mathcal{X}^1)}}{\sqrt{n}}\int_\tau^1\varepsilon^{-d/2\alpha}d\varepsilon\right)\end{aligned}$$

Applying Lemma 5.6.2 with $\beta=\frac{d}{2\alpha}$ and $a=3\sqrt{\frac{K\lambda}{n}}$ where $K=K_{\alpha,d}$ is the constant depending only on α and d borrowed from Theorem 5.3.7 and $\lambda\coloneqq\lambda_d(\mathcal{X}^1)$, we get

$$\mathbb{E}\Big[\sup_{h\in\mathcal{H}^\alpha}|P_nh-Ph|\Big]\le 12\begin{cases}\left(\frac{K\lambda}{n}\right)^{\alpha/d}\left[\frac{d}{d-2\alpha}\wedge(1+0.5\log(\frac{n}{9K\lambda}))\right] & \text{if } \alpha<{}^d\!/_2,\\ \left(\frac{K\lambda}{n}\right)^{1/2}\left[\frac{2\alpha}{2\alpha-d}\wedge(1+\frac{\alpha}{d}\log(\frac{n}{9K\lambda}))\right] & \text{if } \alpha\ge{}^d\!/_2.\end{cases}\tag{5.7}$$

The proof is finished since the upper bound stated in Theorem 5.4.1 is a direct consequence of (5.7)

5.6.3 Additional lemma

The following lemma enables to obtain an upper bound on Dudley's refined bound (Theorem 5.3.9) for any bounded class whose entropy grows polynomially in $1/\varepsilon$.

Lemma 5.6.2. *For any real positive numbers a and β, it holds*

$$\min_{0\le\tau\le1}\left(\tau+a\int_\tau^1\varepsilon^{-\beta}d\varepsilon\right)\le(a^{1/\beta}\vee a)\left[\left(\frac{\beta\vee1}{|\beta-1|}\right)\wedge\left(1+\frac{\log({}^1\!/_a)}{\beta\vee1}\right)\right].$$

Proof. Let a and β be real positive numbers. Define the function

$$\begin{aligned}f\colon[0,1]&\to\mathbb{R}\\ \tau&\mapsto\tau+a\int_\tau^1\varepsilon^{-\beta}d\varepsilon.\end{aligned}$$

One can easily check that

$$f^*\coloneqq\min_{0\le\tau\le1}f(\tau)=\begin{cases}1 & \text{if } a>1,\\ a^{1/\beta}+\frac{a}{1-\beta}(1-a^{1/\beta-1}) & \text{if } a<1\ .\end{cases}$$

In the case $a<1$, using the fact that $1-x^\alpha\le\log(x^{-\alpha})$ for any $\alpha>0$ and $x\in(0,1]$, we have

$$f^*\le(a^{1/\beta}\vee a)\left[\left(\frac{\beta\vee1}{|\beta-1|}\right)\wedge\left(1+\frac{\log({}^1\!/_a)}{\beta\vee1}\right)\right].\tag{5.8}$$

Finally, since the RHS of (5.8) is greater than 1 for any $a>1$, (5.8) holds for any positive real a and this concludes the proof. □

CHAPTER 6

Statistical guarantees for generative models without domination

> To copy is merely to reflect something already there, inertly [...]. But by imitation we enlarge nature itself, we become nature or we discover in ourselves nature's active part.
>
> *William Carlos William*

In this chapter, we introduce a convenient framework for studying (adversarial) generative models from a statistical perspective. It consists in modeling the generative device as a smooth transformation of the unit hypercube of a dimension that is much smaller than that of the ambient space and measuring the quality of the generative model by means of an integral probability metric. In the particular case of integral probability metric defined through a smoothness class, we establish a risk bound quantifying the role of various parameters. In particular, it clearly shows the impact of dimension reduction on the error of the generative model.

Based on Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak S. Dalalyan (2021). "Statistical guarantees for generative models without domination". In: *Algorithmic Learning Theory*. Ed. by Vitaly Feldman, Katrina Ligett, and Sivan Sabato. Vol. 132. Proceedings of Machine Learning Research. PMLR, pp. 1051–1071.

Contents

6.1 Introduction

The problem of learning generative models has attracted a lot of attention during the last 5 years in machine learning and artificial intelligence. The most prominent example is generating artificial images that look similar to actual photographs, by means of generative adversarial networks. The more general formulation of the problem can be given as a game between the user and the learner. The user samples a set of elements (images of natural scenes, poems, pieces of music, etc.) from a hidden distribution $P^* = P_{\text{user}}$ defined on a hidden (and not so well known) space. The learner receives a noisy and possibly contaminated version of these elements and aims at generating a new set of elements, that are different from those transmitted by the user, but that could have been sampled from the hidden distribution P^*. Note that the revealed elements are usually of very high dimension. However, they may exhibit rich structures such as the harmonic and rhythmic schemes followed by a melody or a poem, or the presence of simple shapes in an image. It is therefore reasonable to assume that these elements can be represented by means of a much lower dimensional latent variable, which is unobserved.

In other words, generative models are used for accomplishing the following task. The user draws n independent samples $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ from a distribution P_{user} defined on $\mathbb{R}^D$. The learner is given a noisy and contaminated version $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of this sample. The goal of the learner is to design an algorithm that generates random samples from a distribution P_{learner} which is as close as possible to P_{user}. This can be viewed as a distribution estimation problem with two requirements:

[R1] *It should be easy to sample from* P_{learner}.

[R2] *The way we measure the closeness between* P_{learner} *and* P_{user} *for evaluating the error has to admit an interpretation as a sampling error.*

Of course, this formulation is incomplete since it allows to take the uniform distribution over the observed samples as P_{learner}, *i.e.*, $P_{\text{learner}} = \widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{X}_i}$ (the empirical distribution based on the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$). From a generative modeling perspective, $\widehat{P}_n$ is pointless since it does not yield new samples that are different from the previous ones. Hence, generative modeling requires a third distinctive feature:

[R3] *Samples drawn from* P_{learner} *should be different from those revealed by the user.*

Requirement **R3** is perhaps the hardest to translate into a statistical language. Most prior work focused on the case where both P_{user} and P_{learner}, defined on $\mathbb{R}^D$ equipped with the Borel σ-field, are absolutely continuous with respect to the Lebesgue measure (or another σ-finite

measure). This readily implies that the total variation distance between P_{learner} and $\widehat{P}_n$ is equal to 1, which can be considered as a guarantee for P_{learner} to satisfy **R3**.

Positing that P_{user} has a density with respect to the Lebesgue measure, or any other dominating σ-finite measure μ on $\mathbb{R}^D$, is, in general, incompatible with the fact that P_{user} is inherited from a low-dimensional latent variable and supported by a low-dimensional manifold. For instance, in the simple example of $P_{\text{user}} = \mathcal{U}(a\mathbb{S}^{D-1})$, the uniform measure on $a\mathbb{S}^{D-1}$ (the sphere of radius a centered at the origin), there exists no σ-finite measure dominating all the measures $\mathcal{U}(a\mathbb{S}^{D-1})$, for $a > 0$. Very importantly, as a consequence of the restriction to dominated distributions, the available statistical results fail to assess the positive impact of the reduced dimension of the latent space (as compared to the ambient dimension D) on the quality of the generative model.

We propose to circumvent this drawback by restricting the set of candidate generators to those defined as a smooth transformation of the uniform distribution on a low-dimensional hyper-cube. Obviously, the support of these candidate distributions is a path-connected set. Therefore, the empirical distribution $\widehat{P}_n$, as well as any finitely or countably supported distribution is not among these candidates.

The following notation will be used throughout this work. For every positive integer p, we denote by $\mathcal{U}_p$ the uniform distribution on the hyper-cube $[0,1]^p$. For any convex set $\mathcal{X} \subset \mathbb{R}^p$, $\text{Lip}_L(\mathcal{X})$ stands for the set of all Lipschitz-continuous functions defined on $\mathcal{X}$ with a Lipschitz constant less than or equal to L. For a distribution P defined on a measurable space $(E, \mathscr{E})$ and a measurable map $g : E \mapsto F$, where F is another space endowed with a σ-algebra $\mathscr{F}$, we denote by $g\sharp P$ the "push-forward" measure defined by $(g\sharp P)(A) = P(g^{-1}(A))$ for all $A \in \mathscr{F}$. For a function $g : \mathcal{X} \to \mathbb{R}$, $\|g\|_\infty = \max_{x\in\mathcal{X}} |g(x)|$ is the supremum norm of g.

The rest of the paper is organised as follows. A brief review of the prior work on generative models is presented in Section 6.2, while Section 6.3 provides the formal statement of the problem. In order to convey the main ideas in a simple setting, we analyse the case of noise-free and uncontaminated observations in Section 6.4. The main results are stated and discussed in Section 6.5. A summary of the contributions and some avenues for future research are included in Section 6.6, while Section 6.7 gathers the proofs of the results stated in previous sections.

6.2 Related work (and contributions)

The procedures for generative modeling can be split into two groups: prescribed and implicit probabilistic models (Mohamed and Lakshminarayanan 2016). The former requires an explicit (parametric) specification of the distribution of the observed random variables (*e.g.*, mixture of Gaussian) through a likelihood function, whereas the latter defines a stochastic procedure that directly generates data. The growing complexity of the data makes it harder to design a relevant likelihood function and thus favoured the advent of the latter models. For instance, Generative Adversarial Networks (GANs), perhaps the most well-known generative models based on implicit modeling, enabled groundbreaking advances in the generation of realistic images (Goodfellow et al. 2014; Radford, Metz, and Chintala 2015; Goodfellow 2016; Isola et al. 2017; Zhu et al. 2017; Brock, Donahue, and Simonyan 2018; Karras, Laine, and Aila 2019). In the original GAN framework (Goodfellow et al. 2014) a generator G competes against a discriminator D, both implemented as deep neural networks, in the following zero-sum game:

the generator G (resp. the discriminator D) maximizes (resp. minimizes) the objective

$$\Phi(G, D) = \frac{1}{n}\sum_{i=1}^{n} \log D(\boldsymbol{X}_i) + \mathbf{E}_{\tilde{\boldsymbol{X}} \sim G \sharp P_{\boldsymbol{U}}} \log\left(1 - D(\tilde{\boldsymbol{X}})\right), \tag{6.1}$$

where $P_{\boldsymbol{U}}$ is an easy-to-sample-from noise distribution (*e.g.*, Gaussian or uniform). The goal of the generator is to transform the (low-dimensional) latent variable into artificial data as indistinguishable as possible from the examples drawn from the target distribution. As for the discriminator, the aim is to discriminate between true examples and generated data. See Figure 6.1 for an illustration of the original GAN model. Informally, the generative model can be thought of as a counterfeiter, trying to produce fake paintings and selling it without detection, while the discriminative model is analogous to art experts, trying to detect the counterfeit paintings. Let us note that here P_{learner} would be the distribution of the generated data, *i.e.*, $G \sharp P_{\boldsymbol{U}}$.

Despite their impressive empirical performance, GANs are notoriously hard to train; Even if some fixes have been proposed (Salimans et al. 2016), several problems are yet to be fully understood and solved (*e.g.*, mode collapse, vanishing gradients, failure to converge). Goodfellow et al. (2014) showed that, when the discriminator is optimal, minimizing (6.1) with respect to the generator G amounts to minimizing the Jensen-Shannon (JS) divergence between the generated data distribution and the real sample distribution. Arguing that the topology induced by the JS divergence is rather coarse, Arjovsky, Chintala, and Bottou (2017) proposed to replace this divergence by the Wasserstein-1 distance to stabilize training, leading to the so-called *Wasserstein GAN*. More precisely, the goal of the generator G in this variant is to generate data from a distribution that is as close as possible, w.r.t. the Wasserstein-1 distance, to the empirical distribution of the original data. This leads to the objective

$$\mathsf{W}_1(G \sharp P_{\boldsymbol{U}}, \widehat{P}_n) = \sup_{f \in \mathrm{Lip}_1(\mathcal{X})} \left| \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i) - \mathbf{E}_{\tilde{\boldsymbol{X}} \sim G \sharp P_{\boldsymbol{U}}} f(\tilde{\boldsymbol{X}}) \right|. \tag{6.2}$$

In view of this relation, which follows from the Kantorovitch-Rubinstein duality theorem (Villani 2008, Theorem 5.9, Remark 6.5), the Wasserstein distance admits a nice interpretation as a sampling error. Recall from (IPM) that replacing the class of Lipschitz functions by an arbitrary functional class $\mathcal{F}$, we obtain general Integral Probability Metrics (IPM) : a class of pseudo-metrics on the space of probability measures (Müller 1997). We refer the reader to Liang (2019) and Sriperumbudur et al. (2012) for statistical results related to IPM. An IPM can naturally be interpreted as an adversarial loss: to compare two probability distributions, it seeks for the function f^* in $\mathcal{F}$ for which the expectations of $f(\boldsymbol{X})$ under the two distributions have the largest discrepancy. This formalization enables to study a family of pseudo-metrics which encompasses the Wasserstein-1 distance and generalises the Wasserstein GAN problem. In particular, in this work, we will consider IPM indexed by Sobolev-type classes of functions.

Since GANs initially emerged from the deep learning community, the first line of work primarily relied on empirical insights and general mathematical intuitions. Later on, a parallel line of work tackled the GAN problem from the statistical perspectives (Biau, Sangnier, and Tanielian 2020; Biau et al. 2018; Chen et al. 2020; Liang 2018; Singh et al. 2018; Luise, Pontil, and Ciliberto 2020; Uppal, Singh, and Poczos 2019) as well as optimization and algorithmic viewpoints (Liang and Stokes 2018; Kodali et al. 2017; Pfau and Vinyals 2016; Nie and Patel 2020; Nagarajan and Kolter 2017; Genevay, Peyré, and Cuturi 2017; Genevay et al. 2018).

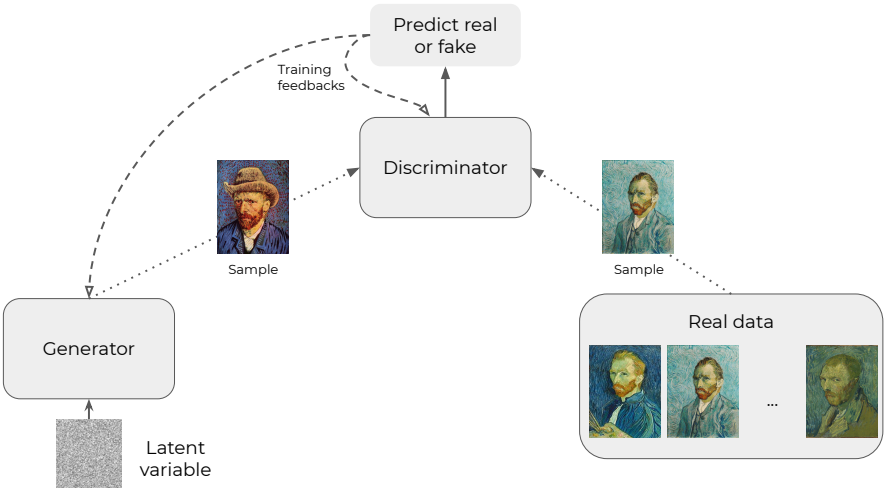


Figure 6.1: Illustration of the original GAN model on some of Vincent Van Gogh's self-portraits. During the training phase, real data and generated data are fed to the discriminator (dotted arrows) which in turn must predict which data is real and which is fake. Feedback (in the form of gradients of the loss) are then sent to the generator and the discriminator (broken arrows) based on predictions from the latter to update their parameters (through back-propagation in the case of neural networks). Note that the generator does not directly have access to real data.

From a statistical perspective, the usual goal is to obtain a bound on the discrepancy between the learned distribution P_{learner} and the true distribution of the data $P^* = P_{\text{user}}$ with respect to a given evaluation metric d. A particularly relevant task is the quantification of the rate of convergence to zero of this discrepancy as the sample size n grows to infinity. Given a family of candidate distributions $\mathcal{P}$, typical bounds are of the form

$$\mathbf{E}_{(X_1,\ldots,X_n)\sim P_{\text{obs}}}\left[\mathsf{d}(P_{\text{learner}}, P_{\text{user}})\right] - \inf_{P\in\mathcal{P}} \mathsf{d}(P, P_{\text{user}}) \lesssim n^{-r(\alpha,\beta,d,D)}.$$

for some exponent $r(\alpha, \beta, d, D) > 0$, where the parameter α characterises the *complexity* of the discriminator (*e.g.*, the smoothness of the class $\mathcal{F}$ used in the IPM), β represents the *smoothness* of the generator, d is the intrinsic dimension of the data, (*i.e.*, the dimension of the latent variable $\boldsymbol{U}$) and D is the ambient dimension (*e.g.*, the number of pixels in an image). Since D is typically much larger than d, it is suitable to avoid any dependence on D in the exponent $r(\alpha, \beta, d, D)$.

Chen et al. 2020; Liang 2018; Singh et al. 2018; Uppal, Singh, and Poczos 2019 obtained rates depending on the smoothness of the density of the target distribution and (eventually) on the smoothness of the class $\mathcal{F}$ of admissible discriminators. Their rates do depend on the ambient dimension D, leading to the curse of dimensionality phenomenon; they do not account for possible low-dimensionnality of the data. Moreover, the learner distributions proposed in those papers are not necessarily easy-to-sample-from.

Without any smoothness assumptions, Biau et al. 2018 provide large sample properties of the estimated distribution assuming that all the densities induced by the class of generators are dominated by a fixed known measure on a Borel subset of $\mathbb{R}^D$. When the admissible

discriminators are neural networks with a given architecture, Biau, Sangnier, and Tanielian 2020 obtained the parametric rate $n^{-1/2}$.

To our knowledge, Luise, Pontil, and Ciliberto (2020) is the only work which establishes statistical guarantees under the assumption that the data generating process is a smooth transformation of a low-dimensional latent distribution. Two key differences with our work is that Luise, Pontil, and Ciliberto (2020) measure quality of sampling through the Sinkhorn divergence (while we consider IPMs) and consider smoothness larger than $d/2$. The latter leads to parametric rates of convergence $n^{-1/2}$. Note also that the Sinkhorn divergence, introduced as a compelling computational alternative to the Wasserstein distance (Cuturi 2013), does not admit a straightforward interpretation as a sampling error.

In this work, we assess the impact of the smoothness of the data generating process and the low-dimensionality of the latent space on the rates of convergence. The rates in the literature either depend on the ambient dimension, which can not explain the effectiveness of GANs, or assume strong smoothness assumption leading to parametric rate. This prevents a fine-grained analysis of the interplay between dimensions and smoothness. In this work we obtain rates which, in terms of dimension, depend only on the intrinsic dimension d of the data and on the smoothness of the data generating process and the admissible discriminators.

6.3 Problem statement

We are given n points $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ in $\mathbb{R}^D$, that we assume drawn independently from an unknown joint probability distribution $P_{\text{obs}}^{(n)}$. We will make the hypothesis that the data points lie—up to a small noise—on a d-dimensional smooth manifold $\mathcal{M}$ with an intrinsic dimension d much smaller than the ambient dimension D. More precisely, we assume that the $\boldsymbol{X}_i$'s are perturbed versions of n independent copies of a point randomly sampled from a distribution P^* supported on the smooth manifold $\mathcal{M}$. The goal of generative modeling is to design a smooth function

$$g : [0,1]^d \to [0,1]^D$$

such that the image of the uniform distribution $\mathcal{U}_d := \mathcal{U}([0,1]^d)$ by g is close to the target distribution P^*. Of course, this framework requires to make precise what is meant by "smoothness" of the function g and how the closeness of two distributions is measured. Since the goal of the present work is to gain a better theoretical understanding of the problem of generative modeling, we assume that the "intrinsic dimension" d is known.

The following condition will be assumed to be true throughout this work, where $\sigma \geq 0$ and $\varepsilon \in [0,1]$ are fixed yet possibly unknown constants.

Assumption A: There exists a mapping $g^* : [0,1]^d \to [0,1]^D$ (with $d \ll D$), as well as random vectors $\boldsymbol{U}_1, \dots, \boldsymbol{U}_n \in \mathbb{R}^d$ and $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n \in \mathbb{R}^D$ such that

- $\boldsymbol{U}_i$ are iid uniformly distributed in the hypercube $[0,1]^d$ (denoted by $\boldsymbol{U}_i \overset{\text{iid}}{\sim} \mathcal{U}_d$),
- $\max_{i=1,\dots,n} \mathbf{E}[\|\boldsymbol{\xi}_i\|_2] \leq \sigma$ for some $\sigma < \infty$,
- For some $\mathcal{I} \subset \{1, \dots, n\}$ of cardinality at least $(1-\varepsilon)n$, we have $\boldsymbol{X}_i = g^*(\boldsymbol{U}_i) + \boldsymbol{\xi}_i$ for every $i \in \mathcal{I}$.

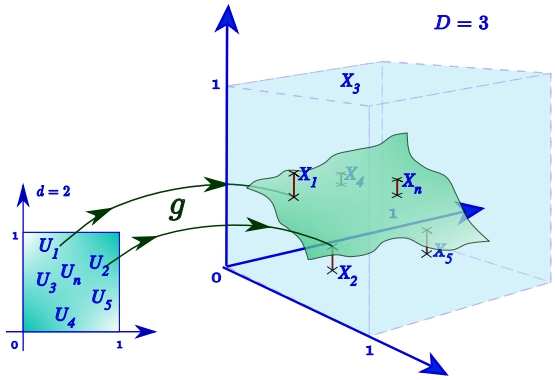


Figure 6.2: An illustration of **Assumption A**. Most $\boldsymbol{X}_i$'s are close to the manifold defined as the image of $[0,1]^d$ by the smooth map g. A small fraction of the $\boldsymbol{X}_i$'s (such as $\boldsymbol{X}_3$ in this figure) might be at a large distance from $g([0,1]^d)$.

The parameters σ and ε, referred to as the noise magnitude and the rate of contamination, are unknown but assumed to be small. The subset $\mathcal{I}$ in the last item of the assumption is the set of inliers. **Assumption A** means that up to some noise, the inliers are drawn from the uniform distribution on the hyper-cube and pushed-forward by g^*. The setting considered here is *adversarial*: the set of inliers and the values of the outliers $\{\boldsymbol{X}_i : i \notin \mathcal{I}\}$ may depend on all the random variables $\boldsymbol{U}_i, \boldsymbol{X}_i, \boldsymbol{\xi}_i$. Furthermore, $\boldsymbol{U}_i$ and $\boldsymbol{\xi}_i$ are not necessarily independent.

Note that the the mapping g^* is not identifiable: pre-composing it with any mass preserving mapping $h : [0,1]^d \to [0,1]^d$ and taking the image of $\mathcal{U}_d$ yields the same distribution as $g^* \sharp \mathcal{U}_d$. Similar identifiability issues arise in graphon estimation problems where identifiability is only possible up to some transformations (Klopp, Tsybakov, and Verzelen 2017; Klopp and Verzelen 2019).

In what follows, we set $P^* = g^* \sharp \mathcal{U}_d$ and call it the oracle generator. Let d be a pseudo-metric on the space of all probability measures on $\mathbb{R}^D$. Most relevant examples in the present context are IPMs, but one could also consider the Wasserstein q-distances with $q \geq 1$, the Hellinger distance, the maximum mean discrepancy and so on. For every candidate generator g—a measurable mapping from $[0,1]^d$ to $\mathbb{R}^D$—we define the risk

$$R_{\mathsf{d},P^*}(g) \coloneqq \mathsf{d}(g \sharp \mathcal{U}_d, P^*). \tag{6.3}$$

Our goal is to find a mapping

$$\begin{aligned} \widehat{G} : (\mathbb{R}^D)^n &\to \mathcal{G} \\ (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) &\mapsto \widehat{g}_n, \end{aligned}$$

such that $R_{\mathsf{d},P^*}(\widehat{g}_n)$ is as small as possible. Note here that $R_{\mathsf{d},P^*}(\widehat{g}_n)$ is a random variable, since $\widehat{g}_n$ is random. Let $\mathcal{G}$ be a set of smooth (at least Lipschitz continuous) functions from $[0,1]^d$ to $\mathbb{R}^D$. We define the generator minimizing the empirical risk, hereafter referred to as the ERM, by

$$\widehat{g}^{\mathrm{ERM}}_{n,\mathcal{G}} \in \arg\min_{g \in \mathcal{G}} \mathsf{d}(g \sharp \mathcal{U}_d, \widehat{P}_n). \tag{ERM}$$

We assume that the minimum is attained. Our results extend easily to the case in which it is not attained but adds some unnecessary technicalities. Our main result, presented in the next section, provides an upper bound on the risk (6.3) of the ERM.

To enforce requirement **R2**, we consider distances on the space of probability distributions that can be expressed as integral probability metrics for a class $\mathcal{F}$ of real-valued functions defined on $[0,1]^D$ (see (IPM)).

6.4 Warming up: guarantees in the noiseless setting for W_1

Let us first consider the noiseless and uncontaminated setting $\sigma = \varepsilon = 0$, corresponding to $P_{\text{obs}} = (P^*)^{\otimes n}$. To convey the main ideas of this work without diving into technicalities, we first consider the case of the Wasserstein W_1-distance. Using arguments that are now standard in learning theory, we get[1]

$$R_{\mathsf{d},P^*}(\widehat{g}_{n,\mathcal{G}}^{\mathrm{ERM}}) \leq \inf_{g\in\mathcal{G}} \mathsf{d}(g\sharp\mathcal{U}_d, P^*) + 2\,\mathsf{d}(\widehat{P}_n, P^*). \tag{6.4}$$

This inequality holds for any pseudo-metric d. It follows from the following chain of inequalities:

$$\begin{aligned} R_{\mathsf{d},P^*}(\widehat{g}_{n,\mathcal{G}}^{\mathrm{ERM}}) &= \mathsf{d}(\widehat{g}_{n,\mathcal{G}}^{\mathrm{ERM}}\sharp\mathcal{U}_d, P^*) \\ &\leq \mathsf{d}(\widehat{g}_{n,\mathcal{G}}^{\mathrm{ERM}}\sharp\mathcal{U}_d, \widehat{P}_n) + \mathsf{d}(\widehat{P}_n, P^*) \\ &\leq \inf_{g\in\mathcal{G}} \mathsf{d}(g\sharp\mathcal{U}_d, \widehat{P}_n) + \mathsf{d}(\widehat{P}_n, P^*) \\ &\leq \inf_{g\in\mathcal{G}} \mathsf{d}(g\sharp\mathcal{U}_d, P^*) + 2\,\mathsf{d}(\widehat{P}_n, P^*). \end{aligned}$$

Note that if we replace in (ERM) the empirical distribution $\widehat{P}_n$ by another estimator $\tilde{P}_n$ of P^*, then (6.4) continues to be true with $\tilde{P}_n$ instead of $\widehat{P}_n$ in the right hand side.

The inequality (6.4) provides an upper bound on the risk that is composed of the approximation error $\inf_{g\in\mathcal{G}} \mathsf{d}(g\sharp\mathcal{U}_d, P^*)$ and the stochastic error $2\,\mathsf{d}(\widehat{P}_n, P^*)$. While the former is unavoidable, it is not clear how tight the latter is. In particular, the fact that the term $2\,\mathsf{d}(\widehat{P}_n, P^*)$ measures the distance between the unknown distribution P^* and an approximation of it that does not take into account the specific structure of P^* suggests that it might be possible to get a better upper bound.

This being said, we stick here to inequality (6.4) and devote the rest of this chapter to establishing upper bounds on the stochastic error. To this end, we take advantage of the interplay between the assumptions on P_X and P^* on the one hand, and the set $\mathcal{F}$ defining the IPM $\mathsf{d} = \mathsf{d}_{\mathcal{F}}$ on the other hand. In the case when both the mapping g^* underlying P^* and the elements of $\mathcal{F}$ are Lipschitz, we get the following result.

Theorem 6.4.1. *Let Assumption A be fulfilled with $\sigma = \varepsilon = 0$ and $g^* \in \mathrm{Lip}_L([0,1]^d)$ for some $L > 0$. Let $\mathsf{d} = \mathsf{W}_1$ and set $\widehat{g}_n = \widehat{g}_{n,\mathcal{G}}^{\mathrm{ERM}}$. Then, for some universal constant $c > 0$,*

$$\mathbf{E}[R_{\mathsf{W}_1,P^*}(\widehat{g}_n)] \leq \inf_{g\in\mathcal{G}} R_{\mathsf{W}_1,P^*}(g) + \frac{cL\sqrt{d}}{n^{1/d} \wedge n^{1/2}}(1 + \mathbb{1}_{d=2}\log n). \tag{6.5}$$

[1] See (Liang 2018, Lemma 1) for a similar result.

The full proof of this result being postponed to Section 6.7.1, we provide here a sketch of it. In view of (6.4), it suffices to upper bound $\mathsf{d}(\widehat{P}_n, P) = \mathsf{W}_1(\widehat{P}_n, P^*)$. Since P^* and $\widehat{P}_n$ are the pushforward measures of $\mathcal{U}_d$ and its empirical counterpart by the same Lipschitz mapping, and the composition of two Lipschitz mappings is still Lipschitz, we can upper bound $\mathsf{W}_1(\widehat{P}_n, P^*)$ by $L\mathsf{W}_1(\widehat{P}_{U,n}, \mathcal{U}_d)$. Here, $\widehat{P}_{U,n}$ is the empirical distribution of $\boldsymbol{U}_1, \dots, \boldsymbol{U}_n$ independently sampled from $\mathcal{U}_d$. It is known that, for the Wasserstein-1 distance, there is a universal constant $c > 0$ such that $\mathbf{E}[\mathsf{W}_1(\widehat{P}_{U,n}, \mathcal{U}_d)]$ is upper bounded by the second summand of the right hand side of (6.5); this fact has been established in the seminal paper Dudley (1969) and later refined and extended by many authors; see Weed and Bach (2019), Singh and Póczos (2018), and Lei (2020) and references therein. The version we use here (with an explicit dependence of the constant on the dimension) can be found in Niles-Weed and Rigollet 2019, Prop. 1. This completes the proof.

Some remarks are in order. First, the rate of convergence to zero of the stochastic term, when the sample size goes to infinity, is characterized by the intrinsic dimension only. This rate, $n^{-1/d}$, is much smaller than the naive rate $n^{-1/D}$ provided that the intrinsic dimension is small as compared to D. To the best of our knowledge, despite the embarrassing simplicity of this result, this is the first time that this phenomenon is highlighted in the context of generative modeling.

The second remark concerns the fact that the choice of the set $\mathcal{G}$ in (ERM) impacts only the first term, the approximation error, in the risk bound given by (6.5). This indicates that inequality (6.5) might not be tight when $\mathcal{G}$ is a very narrow set. On the positive side, this bound implies that the set $\mathcal{G}$ can be chosen very large, as long as feature **R1** holds and optimisation problem (ERM) is computationally tractable. Finally, one can wonder whether the assumption that g^* is Lipschitz is realistic in some applications. We believe that it is. Indeed, the generator learned by GAN is a Lipschitz function of the input (Seddik et al. 2020) and leads to qualitatively good results. Therefore, it makes perfect sense to assume that g^* is Lipschitz.

6.5 Main result in the noisy setting for smooth classes

The rate of convergence obtained in the previous section might be overly pessimistic. Indeed, the Wasserstein distance W_1 might be very weak for many applications: it may be sufficient to take as $\mathcal{F}$ a set which is much smaller than that of the Lipschitz functions. In particular, one can consider the case where $\mathcal{F}$ is a smoothness class with a degree of smoothness strictly larger than one. The main result stated below considers this setting and answers the following three questions:

[Q1] *Can we take advantage of the further smoothness of g^* and that of the functions in $\mathcal{F}$ for improving the risk bound (6.5)?*

[Q2] *How does the noise magnitude σ impact the risk?*

[Q3] *Can we get meaningful risk bounds if some data points $\boldsymbol{X}_i$ are corrupted?*

To answer these questions, we consider the case of smoothness classes containing all the functions with bounded partial derivatives up to a given order. Let $\mathcal{X} \subset \mathbb{R}^D$ be some compact set, which will be chosen to be $[0,1]^D$ later on in this section. In what follows, for every

positive integer α, $C^\alpha(\mathcal{X}, \mathbb{R})$ denotes the set of all α-times continuously differentiable functions. In addition, for a multi-index $\boldsymbol{k} \in \mathbb{N}^D$, we write $\mathtt{D}^{\boldsymbol{k}} f$ for the $\boldsymbol{k}$-th order differential of f. Define the α-smoothness class $\mathcal{W}^\alpha(\mathcal{X}; r)$ over $\mathcal{X}$ with radius $L > 0$ by

$$\mathcal{W}^\alpha(\mathcal{X}; L) := \Big\{ f \in C^\alpha(\mathcal{X}, \mathbb{R}) : \max_{|\boldsymbol{k}| \le \alpha} \|\mathtt{D}^{\boldsymbol{k}} f\|_\infty \le L \Big\}.$$

Clearly, $\mathcal{W}^1(\mathcal{X}; L)$ is included in the set $\mathrm{Lip}_L(\mathcal{X})$ of Lipschitz-continuous functions. Furthermore, one can check that $\mathcal{W}^1(\mathcal{X}; L)$ is dense in $\mathrm{Lip}_L(\mathcal{X})$.

Theorem 6.5.1. *Let Assumption A hold and let the coordinates g_j^* of g^* belong to $\mathcal{W}^\alpha([0,1]^d, L)$ for some $L \ge 1$. Then, if $\mathcal{F} = \mathcal{W}^\alpha([0,1]^D, 1)$ in the definition of the IPM, we have*

$$\mathbf{E}[R_{\mathtt{d}_\mathcal{F}, P^*}(\widehat{g}_{n,\mathcal{G}}^{\mathrm{ERM}})] \le \inf_{g \in \mathcal{G}} R_{\mathtt{d}_\mathcal{F}, P^*}(g) + L(\sigma + 2\varepsilon) + \frac{cL^\alpha}{n^{\alpha/d} \wedge n^{1/2}} (1 + \mathbb{1}_{d=2\alpha} \log n). \tag{6.6}$$

where c is a constant which depends only on α, d, D.

Let us note that this theorem answers the three questions **Q1**-**Q3**. In particular, it shows that if the oracle generator map g^* is α-smooth with $\alpha \le d/2$, and the test function defining the distance $\mathtt{d}_\mathcal{F}$ are α-smooth as well, then the last term of the risk bound of the generator minimizing the empirical risk is of order $n^{-\alpha/d}$. This rate improves with increasing α and reaches the optimal rate $n^{-1/2}$, up to a log factor, when $\alpha = d/2$. It also follows from (6.6) that the risk of the generator $\widehat{g}_{n,\mathcal{G}}^{\mathrm{ERM}}$ decreases linearly fast in the noise magnitude σ and the contamination rate ε, when these parameters go to zero.

As mentioned earlier, (6.6) is a consequence of (6.4) and we do not know whether the latter is tight. However, we can show that the right hand side of (6.6) is a tight upper bound on the right hand side of (6.6). More precisely, as stated in the next result, the dependence on σ and ε is tight, while the dependence on n is tight when $\alpha = 1$ or $\alpha > d/2$.

Theorem 6.5.2. *Let $\mathcal{P}_{n,D}(d, \sigma, \varepsilon, g^*)$ be the set of all distributions of n points $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ in $\mathbb{R}^D$ satisfying Assumption A. Let $\mathcal{G}^*$ be a set of functions $g : [0,1]^d \to [0,1]^D$ containing the linear functions. If $\sigma \le 1/2$ and $\mathcal{F}$ contains the projection onto the first axis $\boldsymbol{x} \in [0,1]^D \mapsto x_1 \in \mathbb{R}$, then there is a universal constant $c_1 > 0$ such that*

$$\sup_{g^* \in \mathcal{G}^*} \sup_{P^{(n)} \in \mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)} \mathbf{E}_{P^{(n)}}[\mathtt{d}_\mathcal{F}(\widehat{P}_n, P^*)] \ge c_1 \Big(\sigma + \varepsilon + \frac{1}{n^{1/2}} \Big).$$

If, in addition, $\mathcal{F}$ contains the set of all 1-Lipschitz functions, then

$$\sup_{g^* \in \mathcal{G}^*} \sup_{P^{(n)} \in \mathcal{P}_{n,D}(d,\sigma,\varepsilon)} \mathbf{E}_{\mathcal{P}^{(n)}}[\mathtt{d}_\mathcal{F}(\widehat{P}_n, P^*)] \ge c_1(\sigma + \varepsilon) + \frac{c_d(1 + \mathbb{1}_{d=2} \log n)}{n^{1/d} \wedge n^{1/2}},$$

where c_1 is a universal constant and c_d is a constant depending on d.

The proof of the theorem is postponed to Section 6.7.4. Note that it does not establish the tightness of the dependence of the bound in n in the case of smoothness $\alpha \in (1, d/2)$. However, it is very likely that the rate is also optimal in this case as well.

To complete this section, we show that the dependence in ε and σ of the upper bound (6.6) is tight.

Theorem 6.5.3. *Let $\mathcal{P}_{n,D}(d, \sigma, \varepsilon, g^*)$ be the set of all distributions of n points $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ in $\mathbb{R}^D$ satisfying Assumption A. Let $\mathcal{G}^*$ be a set of functions $g : [0,1]^d \to [0,1]^D$ containing the affine functions. Assume that $\mathcal{F}$ is a set of functions $f : [0,1]^D \to \mathbb{R}$ bounded by[2] L and containing the projection onto the first axis $\boldsymbol{x} \in [0,1]^D \mapsto x_1 \in \mathbb{R}$. Then, if $\sigma \leq 1/2$ and $n \geq (6/\varepsilon)\log(20L/\varepsilon)$, we have*

$$\inf_{\widehat{g}_n} \sup_{g^* \in \mathcal{G}^*} \sup_{P^{(n)} \in \mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)} \mathbf{E}[R_{\mathtt{d}_{\mathcal{F}},P^*}(\widehat{g}_n)] \geq 0.1(\sigma + \varepsilon),$$

where the inf is taken over all possible generators $\widehat{g}_n$.

The proof of this result is postponed to Section 6.7.5. If we compare this lower bound with the upper bound of Theorem 6.5.1, we see that the linear dependence of the expected risk on the parameters σ and ε is optimal and cannot be improved. This is true for any generator, meaning that the empirical risk minimizer is minimax rate-optimal in terms of σ and ε. We are currently working on establishing similar lower bounds showing the optimality in terms of n as well.

6.6 Conclusion and outlook

In this work, we introduced a general and nonparametric framework for learning generative models. Given data in a possibly high-dimensional space, we learn their distribution in order to sample new data points that resemble the training ones, while not being identical to those. A key point in our work is to leverage the fact that the distribution of the training samples, up to some noise and adversarial contamination, is supported by a low-dimensional smooth manifold. This allows us to alleviate the curse of dimensionality. Such an assumption is very reasonable as it reflects the structural properties of the training samples. For instance, the MNIST dataset (LeCun 1998) is composed of 28×28 pixels pictures of handwritten digits while the intrinsic dimension of the data is estimated to be around 14 (Costa and Hero 2004a; Levina and Bickel 2005).

We established risk bounds for the minimizer of the distance between the empirical distribution and admissible generators, where an admissible generator is a smooth function pushing forward a low-dimensional uniform distribution into the high-dimensional sample space. We use Integral Probability Metrics for measuring the discrepancy between the target distribution and our estimate: These metrics, which include the total variation and the Wasserstein-1 distances, mimic the role of a discriminator which would try to discriminate between true samples and the simulated ones.

By proving new bounds on the distance between such distributions and their empirical counterparts, we were able to derive nonasymptotic bounds for the regret of our empirical risk minimizer, with rates of convergence that only depend on the ambient dimension through fixed multiplicative constants. Our new bounds, which are of independent interest, leverage both the smoothness of the distribution of the samples and that of the functions in the IPM class.

We were also able to take into account possible adversarial corruption of the training samples both by noise (*e.g.*, blurry images) and by a small proportion of outliers (*i.e.*, wrong samples

[2] It can be checked that the same result holds if the functions f satisfy $\max_x f(x) - \min_x f(x) \leq L$.

in the training set), inducing some error terms that are shown to be unavoidable. To the best of our knowledge, this is the first result assessing the influence of the noise and of the contamination on the error of generative modeling. This constitutes an appealing complement to the recently obtained statistical guarantees (Biau, Sangnier, and Tanielian 2020; Luise, Pontil, and Ciliberto 2020).

As a route for future work, we believe that our regret bounds are not minimax optimal in all possible regimes (depending on the smoothness of the generators). Namely, it is not clear that fitting our generator to the empirical distribution $\widehat{P}_n$ yields an optimal method, especially when the smoothness α is less than the half of the dimension d. It might be more judicious to fit the generator to a smoothed version of the empirical distribution $\widehat{P}_n$. In another direction, since the particular structure of deep neural networks might explain why they appear to avoid the curse of dimensionality (Poggio et al. 2017)), it could be worth incorporating in our procedure this feature of deep neural networks, that are used for implementing GANs in practice.

6.7 Proofs

This section contains the proofs of the main results stated in previous sections. We start by providing the proof of Theorem 6.4.1. Then, the proof of Theorem 6.5.1 is presented up to the proof of a technical lemma on the composition of smooth functions, postponed to Section 6.7.3.

6.7.1 Proof of Theorem 6.4.1

To ease notation, we write $\widehat{g}_n$ instead of $\widehat{g}_{n,\mathcal{G}}^{\mathrm{ERM}}$. In view of (6.4), we have

$$R_{\mathsf{W}_1,P^*}(\widehat{g}_n) \leq \inf_{g\in\mathcal{G}} \mathsf{W}_1(g\sharp\mathcal{U}_d, P^*) + 2\mathsf{W}_1(\widehat{P}_n, P^*).$$

Using the variational formulation of the Wasserstein-1 distance we write

$$\begin{aligned}
\mathsf{W}_1(\widehat{P}_n, P^*) &= \sup_{f\in \mathrm{Lip}_1([0,1]^D)} \Big| \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i) - \mathbf{E}_{\boldsymbol{X}\sim P^*} f(\boldsymbol{X}) \Big| \\
&= \sup_{f\in \mathrm{Lip}_1([0,1]^D)} \Big| \frac{1}{n}\sum_{i=1}^n f\circ g^*(\boldsymbol{U}_i) - \mathbf{E}_{\boldsymbol{U}\sim\mathcal{U}_d} f\circ g^*(\boldsymbol{U}) \Big| \\
&= \sup_{h\in\mathcal{H}_L} \Big(\frac{1}{n}\sum_{i=1}^n h(\boldsymbol{U}_i) - \mathbf{E}_{\boldsymbol{U}\sim\mathcal{U}_d} h(\boldsymbol{U}) \Big)
\end{aligned}$$

where we define the class $\mathcal{H}_L = \{h : [0,1]^d \to \mathbb{R} : h = f\circ g^*, f \in \mathrm{Lip}_1([0,1]^D)\}$. Finally, taking the expectation and noting that $\mathcal{H}_L$ is a subset of the L-Lipschitz functions on $[0,1]^d$ with values in $\mathbb{R}$, we get

$$\begin{aligned}
\mathbf{E}[\mathsf{W}_1(\widehat{P}_n, P^*)] &\leq \mathbf{E}\Big[\sup_{h\in \mathrm{Lip}_L([0,1]^d)} \Big(\frac{1}{n}\sum_{i=1}^n h(\boldsymbol{U}_i) - \mathbf{E}_{\boldsymbol{U}\sim\mathcal{U}_d} h(\boldsymbol{U}) \Big) \Big] \\
&\leq L\, \mathbf{E}[\mathsf{W}_1(\widehat{P}_{U,n}, \mathcal{U}_d)] \\
&\leq \frac{cL\sqrt{d}}{n^{1/d}\wedge n^{1/2}} (1 + \mathbb{1}_{d=2}\log n),
\end{aligned}$$

with c a universal constant. The last inequality follows from Niles-Weed and Rigollet 2019, Proposition 1.

6.7.2 Proof of Theorem 6.5.1

In view of (6.4), we need to establish an upper bound on the expected stochastic error

$$\text{NoisyStochErr}_n = \mathbf{E}[\mathsf{d}_{\mathcal{F}}(P_n, P^*)] = \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i) - \mathbf{E}[f(g^*(\boldsymbol{U}))]\right|\right],$$

where $\boldsymbol{U} \sim \mathcal{U}_d$ and $\mathcal{F} = \mathcal{W}^\alpha([0,1]^D, 1)$. The first step in the proof is a lemma showing the influence of the noise and the corruption on the error StochErr_n.

Lemma 6.7.1. *If P_X satisfies Assumption A with $\varepsilon \in [0,1]$ and all the functions in $\mathcal{F}$ are bounded by a constant $L_{\mathcal{F}}$ and Lipschitz with constant $L_{\mathcal{F}}$, then*

$$\text{NoisyStochErr}_n \le L_{\mathcal{F}}\sigma + 2M_{\mathcal{F}}\varepsilon + \text{NoiseFreeStochErr}_n,$$

where

$$\text{NoiseFreeStochErr}_n = \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f\circ g^*)(\boldsymbol{U}_i) - \mathbf{E}[(f\circ g^*)(\boldsymbol{U})]\right|\right],$$

with $\boldsymbol{U}, \boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ iid random vectors drawn from $\mathcal{U}_d$.

Proof. The triangle inequality yields

$$\text{StochErr}_n \le \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \{f(\boldsymbol{X}_i) - (f\circ g^*)(\boldsymbol{U}_i)\}\right|\right] + \text{NoiseFreeStochErr}_n.$$

Let us define $\boldsymbol{Y}_i = g^*(\boldsymbol{U}_i) + \boldsymbol{\xi}_i$ for $i = 1, \ldots, n$. The third item of Assumption A implies that $\boldsymbol{Y}_i = \boldsymbol{X}_i$ for $i \in \mathcal{I}$. For $i \notin \mathcal{I}$, we have $|f(\boldsymbol{X}_i) - f(\boldsymbol{Y}_i)| \le 2M_{\mathcal{F}}$. Therefore, the first term in the right hand side of the last display can be further bounded as follows:

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \{f(\boldsymbol{X}_i) - (f\circ g^*)(\boldsymbol{U}_i)\}\right| &\le \left|\frac{1}{n}\sum_{i=1}^n \{f(\boldsymbol{Y}_i) - (f\circ g^*)(\boldsymbol{U}_i)\}\right| + \frac{2M_{\mathcal{F}}(n - n_{\mathcal{I}})}{n} \\ &\le \frac{L_{\mathcal{F}}}{n}\sum_{i=1}^n \|\boldsymbol{Y}_i - g^*(\boldsymbol{U}_i)\| + 2M_{\mathcal{F}}\,\varepsilon \\ &= \frac{L_{\mathcal{F}}}{n}\sum_{i=1}^n \|\boldsymbol{\xi}_i\| + 2M_{\mathcal{F}}\,\varepsilon.\end{aligned}$$

To get the claimed result, it suffices to take the expectation of both sides of the last display. □

The next step consists in upper bounding the stochastic error in the noise free case. If we use the notation $\mathcal{F}\circ g^* = \{f\circ g^* : f \in \mathcal{F}\}$, the noise free stochastic error can be written as

$$\text{NoiseFreeStochErr}_n = \mathbf{E}[\mathsf{d}_{\mathcal{F}\circ g^*}(\widehat{P}_{U,n}, \mathcal{U}_d)]. \tag{6.7}$$

We see that the problem is reduced to that of evaluating the distance between the uniform distribution and the empirical distribution of n independent random points uniformly distributed on the unit hypercube. In order to upper bound this distance, we first show that the class $\mathcal{F} \circ g^*$, under the assumptions of Theorem 6.5.1, is included in a smoothness class of order α. The precise statement is the following.

Lemma 6.7.2. *Let $g : [0,1]^d \to [0,1]^D$ and $h : [0,1]^D \to [-1,1]$ two mappings such that $g \in \mathcal{W}^\alpha([0,1]^d, L)$ and $h \in \mathcal{W}^\alpha([0,1]^D, 1)$ for some $\alpha \in \mathbb{N}^*$ and some $L \geq 1$. Then, there exists a constant $C = C(D,d,\alpha)$ such that*

$$|\mathtt{D}^{\boldsymbol{k}}(h \circ g)(\boldsymbol{x})| \leq CL^\alpha, \quad \forall \boldsymbol{x} \in [0,1]^d,$$

for every multi-index $\boldsymbol{k} = (k_1, \ldots, k_d) \in \mathbb{N}^d$ such that $|\boldsymbol{k}| \leq \alpha$.

This lemma, in conjunction with (6.7) and the assumption $g^* \in \mathcal{W}^\alpha([0,1]^d, L)$, implies that

$$\begin{aligned} \text{NoiseFreeStochErr}_n &\leq \mathbf{E}[\mathtt{d}_{\mathcal{W}^\alpha([0,1]^d, CL^\alpha)}(\widehat{P}_{U,n}, \mathcal{U}_d)] \\ &= CL^\alpha \mathbf{E}[\mathtt{d}_{\mathcal{W}^\alpha([0,1]^d,1)}(\widehat{P}_{U,n}, \mathcal{U}_d)]. \end{aligned}$$

The last step is to use Schreuder (2020, Theorem 4), which provides the inequality

$$\mathbf{E}[\mathtt{d}_{\mathcal{W}^\alpha([0,1]^d, CL^\alpha)}(\widehat{P}_{U,n}, \mathcal{U}_d)] \leq \tilde{C} L^\alpha n^{-(\alpha \wedge d/2)/d}(1 + \mathbb{1}_{\alpha = d/2} \log n).$$

This completes the proof of the theorem.

6.7.3 Image of a smoothness class by a smooth function

Proof of Lemma 6.7.2 The proof relies on Fraenkel 1978, Formula B providing an explicit formula for derivatives of composite functions: for any multi-index $\boldsymbol{k}$ such that $1 \leq |\boldsymbol{k}| \leq \alpha$ and for any $\boldsymbol{x} \in [0,1]^d$,

$$\mathtt{D}^{\boldsymbol{k}}(h \circ g)(\boldsymbol{x}) = \boldsymbol{k}! \sum_{\boldsymbol{a}: 1 \leq |\boldsymbol{a}| \leq |\boldsymbol{k}|} \frac{(\mathtt{D}^{\boldsymbol{a}} h)(g(\boldsymbol{x}))}{\boldsymbol{a}!} Q_{\boldsymbol{k},\boldsymbol{a}}(g; \boldsymbol{x}), \tag{6.8}$$

where $Q_{\boldsymbol{k},\boldsymbol{a}}(g; \cdot)$ is a homogeneous polynomial of degree $|\boldsymbol{a}|$ in derivatives of $g_1, \ldots, g_D$. Since the partial derivatives of h of any order up to α are bounded by one, we infer from the last display that

$$|\mathtt{D}^{\boldsymbol{k}}(h \circ g)(\boldsymbol{x})| = \boldsymbol{k}! \sum_{\boldsymbol{a}: 1 \leq |\boldsymbol{a}| \leq |\boldsymbol{k}|} \frac{1}{\boldsymbol{a}!} |Q_{\boldsymbol{k},\boldsymbol{a}}(g; \boldsymbol{x})|. \tag{6.9}$$

We can give an explicit expression of $Q_{\boldsymbol{k},\boldsymbol{a}}$ using the following notation. Let r be the cardinality of the set $\{\boldsymbol{\beta} \in \mathbb{N}^d \,|\, 0 < \boldsymbol{\beta} \leq \boldsymbol{\gamma}\}$ and $\boldsymbol{\beta}(1), \ldots, \boldsymbol{\beta}(r)$ be its elements somehow enumerated. Define, for $\boldsymbol{\gamma} \in \mathbb{N}^d$ and for $a \in \mathbb{N}$, the set of multi-indices

$$R(\boldsymbol{\gamma}, a) = \Big\{ \boldsymbol{\rho} \in \mathbb{N}^r \,|\, \sum_{j=1}^r \rho_j \boldsymbol{\beta}(j) = \boldsymbol{\gamma}, |\boldsymbol{\rho}| = a \Big\},$$

and, for any $v : \mathbb{R}^d \to \mathbb{R}$, the polynomials

$$P_{\gamma}(a, v; \boldsymbol{x}) = \sum_{\boldsymbol{\rho} \in R(\gamma, a)} \frac{a!}{\boldsymbol{\rho}!} \prod_{j=1}^{r} \frac{(\mathtt{D}^{\boldsymbol{\beta}(j)} v(\boldsymbol{x}))^{\rho_j}}{\boldsymbol{\beta}(j)!}. \tag{6.10}$$

The functions $Q_{\boldsymbol{k},\boldsymbol{a}}$ in (6.9) are given by

$$Q_{\boldsymbol{k},\boldsymbol{a}}(g; \boldsymbol{x}) = \sum_{\boldsymbol{\gamma}(1)+\ldots+\boldsymbol{\gamma}(D)=\boldsymbol{k}} \prod_{m=1}^{D} P_{\boldsymbol{\gamma}(m)}(a_m, g_m; \boldsymbol{x}).$$

Since, according to the conditions of the lemma, all the partial derivatives of g appearing in (6.10) for $v = g_m$ are bounded by $L \geq 1$, we have

$$|P_{\boldsymbol{\gamma}(m)}(a_m, g_m; \boldsymbol{x})| \leq \sum_{\boldsymbol{\rho} \in R(\boldsymbol{\gamma}(m), a_m)} L^{|\boldsymbol{\rho}|} \frac{a_m!}{\boldsymbol{\rho}!} \prod_{j=1}^{r} \frac{1}{\boldsymbol{\beta}(j)!}.$$

Since $|\boldsymbol{\rho}| \leq a_m$ and $|\boldsymbol{a}| \leq |\boldsymbol{k}| \leq \alpha$, this leads to

$$|Q_{\boldsymbol{k},\boldsymbol{a}}(g; \boldsymbol{x})| \leq L^{\alpha} \boldsymbol{a}! \sum_{\boldsymbol{\gamma}(1)+\ldots+\boldsymbol{\gamma}(D)=\boldsymbol{k}} \prod_{m=1}^{D} \left(\sum_{\boldsymbol{\rho} \in R(\boldsymbol{\gamma}(m), a_m)} \frac{1}{\boldsymbol{\rho}!} \prod_{j=1}^{r} \frac{1}{\boldsymbol{\beta}(j)!} \right).$$

Combining this inequality with (6.9), we arrive at

$$|\mathtt{D}^{\boldsymbol{k}}(h \circ g)(\boldsymbol{x})| \leq L^{\alpha} \boldsymbol{k}! \sum_{1 \leq |\boldsymbol{a}| \leq |\boldsymbol{k}|} \sum_{\boldsymbol{\gamma}(1)+\cdots+\boldsymbol{\gamma}(D)=\boldsymbol{a}} \prod_{m=1}^{D} \left(\sum_{\boldsymbol{\rho} \in R(\boldsymbol{\gamma}(m), a_m)} \frac{1}{\boldsymbol{\rho}!} \prod_{j=1}^{r} \frac{1}{\boldsymbol{\beta}(r)!} \right).$$

Denoting by $C(D, d, \alpha)$ the maximum of the right hand side over all multi-indices $\boldsymbol{k}$ such that $|\boldsymbol{k}| \leq \alpha$, we get the claim of the lemma. □

6.7.4 Proof of the lower bounds in Theorem 6.5.2

Since the bound we wish to prove does not depend on the dimension, we assume without loss of generality that $D = d$. First, we start by considering the case $\sigma + \varepsilon \geq 2/n^{1/2}$.

Let us define $g^*(\boldsymbol{x}) = (2\boldsymbol{x} + 1)/4$. This function is clearly 1-Lipschitz. Let ξ_1 be a random variable drawn from the uniform in $[0, 1]$ distribution. We define $P_0^{(n)}$ to be the distribution of i.i.d. vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ such that $\boldsymbol{X}_i \overset{\text{dist}}{\sim} g^*(\boldsymbol{U}) + \sigma\xi_1$ for $i = 1, \ldots, n\varepsilon$ and $\boldsymbol{X}_i = (1, \ldots, 1)^\top$ for $i > n\varepsilon$. Then, it is clear that $P_0^{(n)} \in \mathcal{P}_{n,D}(d, \sigma, (1-\varepsilon)n)$ and

$$\mathbf{E}_{P_0^{(n)}}[\mathtt{d}_{\mathcal{F}}(\widehat{P}_n, P^*)] = \mathbf{E}_{P_0^{(n)}}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) - \mathbf{E}[(f \circ g^*)(\boldsymbol{U})] \right| \right] \tag{6.11}$$

$$\geq \mathbf{E}_{P_0^{(n)}}\left[\left| \frac{1}{n} \sum_{i=1}^{n} X_{i,1} - \mathbf{E}[g^*(\boldsymbol{U})_1] \right| \right] \tag{6.12}$$

$$= \mathbf{E}_{P_0^{(n)}}\left[\left| \frac{1}{n} \sum_{i=1}^{n} (X_{i,1} - \mathbf{E}[X_{i,1}]) + \varepsilon + 0.5\sigma - \varepsilon \mathbf{E}[g^*(\boldsymbol{U})_1] \right| \right] \tag{6.13}$$

$$= \mathbf{E}_{P_0^{(n)}}\left[\left| \frac{1}{n} \sum_{i=1}^{n} (X_{i,1} - \mathbf{E}[X_{i,1}]) + 0.5(\sigma + \varepsilon) \right| \right]. \tag{6.14}$$

The first inequality above follows by replacing the sup over $\mathcal{F}$ by the corresponding expression evaluated at the representer $f_0(\boldsymbol{x}) = x_1$. The third line above follows from $\mathbf{E}[X_{i,1}] = \mathbf{E}[g^*(\boldsymbol{U})_1] + 0.5\sigma$ if $i \leq n\varepsilon$ whereas $\mathbf{E}[X_{i,1}] = 1$ if $i > n\varepsilon$. The last line is a consequence of $\mathbf{E}[g(\boldsymbol{U})_1] = 0.5$. Combining the above lower bound with the triangle inequality, we arrive at

$$\begin{aligned}
\mathbf{E}_{P_0^{(n)}}[\mathsf{d}_{\mathcal{F}}(\widehat{P}_n, P^*)] &\geq 0.5(\sigma+\varepsilon) - \mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^n (X_{i,1} - \mathbf{E}[X_{i,1}])\right|\right] \\
&\geq 0.5(\sigma+\varepsilon) - \left(\mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^n (X_{i,1} - \mathbf{E}[X_{i,1}])\right|^2\right]\right)^{1/2} \\
&\geq 0.5(\sigma+\varepsilon) - 0.5/\sqrt{n} \\
&\geq (\sigma+\varepsilon+1/\sqrt{n})/6.
\end{aligned}$$

To get the second line above, we used that the first-order moment is bounded by the second-order moment. In the third line, we used that the variance of the sum of independent random variables is the sum of variances and that the variance of a random variables taking its values in $[0,1]$ is always $\leq 1/4$. Finally, the last line is derived from the assumption $\sigma + \varepsilon \geq 2/\sqrt{n}$.

We now turn to the case $\sigma + \varepsilon \leq 2/\sqrt{n}$. In this case, we use the same distribution $P_0^{(n)}$ as in the previous case but we choose $\sigma = \varepsilon = 0$. From (6.14) we derive that

$$\mathbf{E}_{P_0^{(n)}}[\mathsf{d}_{\mathcal{F}}(\widehat{P}_n, P^*)] \geq \mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^n (X_{i,1} - \mathbf{E}[X_{i,1}])\right|\right] \tag{6.15}$$

$$\geq 0.5\,\mathbf{E}_{U_i \overset{\text{iid}}{\sim} \mathcal{U}_1}\left[\left|\frac{1}{n}\sum_{i=1}^n (U_i - 0.5)\right|\right] \geq 0.105/\sqrt{n} \tag{6.16}$$

In view of the assumption $\sigma + \varepsilon \leq 2/\sqrt{n}$, this leads to

$$\begin{aligned}
\mathbf{E}_{P_0^{(n)}}[\mathsf{d}_{\mathcal{F}}(\widehat{P}_n, P^*)] &\geq \mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^n (X_{i,1} - \mathbf{E}[X_{i,1}])\right|\right] \\
&\geq 0.5\,\mathbf{E}_{U_i \overset{\text{iid}}{\sim} \mathcal{U}_1}\left[\left|\frac{1}{n}\sum_{i=1}^n (U_i - 0.5)\right|\right] \geq 0.035(\sigma+\varepsilon+1/\sqrt{n}),
\end{aligned}$$

which completes the proof of the first inequality of the theorem. For the second inequality, it suffices to combine the first inequality with the lower bound established in the seminal paper Dudley (1969).

6.7.5 Proof of the lower bound in Theorem 6.5.3

We split the proof of Theorem 6.5.3 into two propositions: The first one shows the tightness of the dependence on the contamination rate whereas the second one establishes the tightness of the dependence on the noise-level.

Proposition 6.7.3 (Tightness wrt to the contamination rate)**.** *Under the assumptions of Theorem 6.5.3,*

$$\inf_{\widehat{g}_n} \sup_{g^*} \sup_{P^{(n)} \in \mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)} \mathbf{E}[R_{\mathsf{d}_{\mathcal{F}},P^*}(\widehat{g}_n)] \geq \varepsilon/3.$$

Proof. It can be easily checked that the supremum of the expected risk over $\mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)$ is always not smaller than the supremum of the same quantity over $\mathcal{P}_{n,1}(d,0,\varepsilon,g^*)$. To ease notation, we write $\mathcal{P}_n(d,\varepsilon,g^*)=\mathcal{P}_{n,1}(d,0,\varepsilon,g^*)$ and also set $\mu=\mathcal{U}_d$.

Step 1: Reduction to Huber contamination model. Note that the set of admissible data distributions $\mathcal{P}_{n,D}(d,\varepsilon,g^*)$ comprises the data distributions from Huber's deterministic contamination model Bateni and Dalalyan 2020, Section 2.2, namely data distributions such that a (deterministic) proportion $(1-\varepsilon)$ of the data is distributed according to a reference distribution P^* while the remaining proportion ε is independently drawn from another distribution Q. Therefore, denoting by $\mathcal{P}_n^{\mathrm{HDC}}(d,\varepsilon,g^*)$ such distributions, it holds, for any estimator $\widehat{g}_n$ and generator g^*,

$$\begin{aligned}\sup_{P^{(n)}\in\mathcal{P}_n(d,\varepsilon,g^*)}\mathbf{E}[R_{\mathsf{d}_{\mathcal{F}},P^*}(\widehat{g}_n)] &= \sup_{P^{(n)}\in\mathcal{P}_n(d,\varepsilon,g^*)}\mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^*\sharp\mu,\widehat{g}_n\sharp\mu)]\\ &\geq \sup_{P^{(n)}\in\mathcal{P}_n^{\mathrm{HDC}}(d,\varepsilon,g^*)}\mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^*\sharp\mu,\widehat{g}_n\sharp\mu)].\end{aligned}$$

Furthermore, let us denote by $\mathcal{P}_D^{\mathrm{HC}}(d,\varepsilon,g^*)$ the set of data distributions such that there is a distribution Q defined on the same space as a reference distribution $P^*=g^*\sharp\mu$ such that the observations $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ are independent and drawn from the mixture distribution $(1-\varepsilon)P^*+\varepsilon Q$. In view of $\sup_{g,g'}d_{\mathcal{F}}(g\sharp\mu,g'\sharp\mu)\leq L$ and Bateni and Dalalyan 2020, Proposition 1, for any estimator $\widehat{g}_n$ and generator g^*, we have

$$\sup_{P^{(n)}\in\mathcal{P}_n^{\mathrm{HDC}}(d,\varepsilon,g^*)}\mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^*\sharp\mu,\widehat{g}_n\sharp\mu)]\geq\sup_{P^{(n)}\in\mathcal{P}_n^{\mathrm{HC}}(d,\varepsilon/2,g^*)}\mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^*\sharp\mu,\widehat{g}_n\sharp\mu)]-e^{-n\varepsilon/6}L.$$

The second step consists in lower bounding the risk in the Huber contamination model using an argument based on two simple hypotheses.

Step 2: Construction of hypotheses. Let us define the generators $g_1^*,g_2^*:[0,1]^d\to[0,1]$ as

$$g_1^*(\boldsymbol{u})=(1-\varepsilon)u_1\quad\text{and}\quad g_2^*(\boldsymbol{u})=(1-\varepsilon)u_1+\varepsilon,\quad\text{for }\boldsymbol{u}=(u_1,\dots,u_d)\in[0,1]^d.$$

For contamination distributions $Q_1:=\mathcal{U}([1-\varepsilon,1])$ and $Q_2:=\mathcal{U}([0,\varepsilon])$, define the data generating distributions

$$P_1^{(n)}=[(1-\varepsilon)g_1^*\sharp\mu+\varepsilon Q_1]^{\otimes n}\text{ and }P_2^{(n)}=[(1-\varepsilon)g_2^*\sharp\mu+\varepsilon Q_2]^{\otimes n}.$$

One can easily check that $P_1^{(n)}=P_2^{(n)}=\mathcal{U}([0,1])^{\otimes n}$ and $P_j^{(n)}\in\mathcal{P}_n^{\mathrm{HC}}(d,\varepsilon,g_j^*)$ for $j=1,2$. Using the fact that the maximum is larger than the arithmetic mean, in conjunction with the triangular inequality, we obtain

$$\begin{aligned}\sup_{g^*}\sup_{P^{(n)}\in\mathcal{P}_n^{\mathrm{HC}}(d,\varepsilon,g^*)}\mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^*\sharp\mu,\widehat{g}_n\sharp\mu)] &\geq \frac{1}{2}\left[\mathbf{E}_{P_1^{(n)}}\mathsf{d}_{\mathcal{F}}(g_1^*\sharp\mu,\widehat{g}_n\sharp\mu)+\mathbf{E}_{P_2^{(n)}}\mathsf{d}_{\mathcal{F}}(g_2^*\sharp\mu,\widehat{g}_n\sharp\mu)\right]\\ &=\frac{1}{2}\mathbf{E}_{P_0^{(n)}}\left[\mathsf{d}_{\mathcal{F}}(g_1^*\sharp\mu,\widehat{g}_n\sharp\mu)+\mathsf{d}_{\mathcal{F}}(g_2^*\sharp\mu,\widehat{g}_n\sharp\mu)\right]\\ &\geq\frac{1}{2}\mathsf{d}_{\mathcal{F}}(g_1^*\sharp\mu,g_2^*\sharp\mu)\geq\varepsilon/2.\end{aligned}$$

The last inequality comes from choosing the representer $f(\boldsymbol{u})=u_1$ from $\mathcal{F}$.

Conclusion. Combining the previous two steps, we get

$$\sup_{P^{(n)} \in \mathcal{P}_n^{\mathrm{HDC}}(d,\varepsilon,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^* \sharp \mu, \widehat{g}_n \sharp \mu)] \geq (1/4)\varepsilon - e^{-n\varepsilon/6} L.$$

Choosing $n \geq (6/\varepsilon)\log(20L/\varepsilon)$, we get the claim of the proposition. □

Proposition 6.7.4 (Tightness wrt to the noise level)**.** *Under the assumptions of Theorem 6.5.3, we have*

$$\inf_{\widehat{g}_n \in \mathcal{G}} \sup_{g^* \in \mathcal{G}^*} \sup_{P^{(n)} \in \mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^* \sharp \mu, \widehat{g}_n \sharp \mu)] \geq \sigma/2.$$

Proof. Once again, without loss of generality we assume that $D = 1$, $\varepsilon = 0$ and drop the dependence of different quantities on these two parameters. Recall that $\mu = \mathcal{U}_d$. Let us define the generators $g_j^* : [0,1]^d \to [0,1]^D$, $j = 1, 2$, by

$$g_1^*(\boldsymbol{u}) \equiv 0 \qquad \text{and} \qquad g_2^*(\boldsymbol{u}) \equiv \sigma, \qquad \text{for } \boldsymbol{u} = (u_1, \ldots, u_d) \in [0,1]^d.$$

These functions allow us to define the data generating distributions

$$P_1^{(n)} = [g_1^* \sharp \mu * \delta_\sigma]^{\otimes n} \qquad \text{and} \qquad P_2^{(n)} = [g_2^* \sharp \mu * \delta_0]^{\otimes n}.$$

One can easily check that $P_1^{(n)} = P_2^{(n)} = \delta_\sigma^{\otimes n}$, which belongs to $\mathcal{P}_n(d,\sigma,g_1^*) \cap \mathcal{P}_n(d,\sigma,g_2^*)$. Furthermore, $g_j^* \in \mathcal{G}^*$ for $j = 1, 2$ since the latter contains all the affine functions. Using the same arguments as in the proof of the previous proposition, we arrive at

$$\begin{aligned}\sup_{g^* \in \mathcal{G}^*} \sup_{P^{(n)} \in \mathcal{P}_n(d,\sigma,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^* \sharp \mu, \widehat{g}_n \sharp \mu)] &\geq \frac{1}{2} \mathbf{E}_{P_1^{(n)}} [\mathsf{d}_{\mathcal{F}}(g_1^* \sharp \mu, \widehat{g}_n \sharp \mu) + \mathsf{d}_{\mathcal{F}}(g_2^* \sharp \mu, \widehat{g}_n \sharp \mu)] \\ &\geq \frac{1}{2} \mathsf{d}_{\mathcal{F}}(g_1^* \sharp \mu, g_2^* \sharp \mu) \geq \sigma/2.\end{aligned}$$

This completes the proof of the proposition. □

To get the claim of Theorem 6.5.3, it suffices to combine the claims of the last two propositions with the fact that $(0.2\varepsilon \vee 0.5\sigma) \geq 0.1(\varepsilon + \sigma)$.

CHAPTER 7

A minimax framework for quantifying risk-fairness trade-off in regression

We propose a theoretical framework for the problem of learning a real-valued function which meets fairness requirements. This framework is built upon the notion of α-relative (fairness) improvement of the regression function which we introduce using the theory of optimal transport. Setting $\alpha = 0$ corresponds to the regression problem under the Demographic Parity constraint, while $\alpha = 1$ corresponds to the classical regression problem without any constraints. For $\alpha \in (0, 1)$ the proposed framework allows to continuously interpolate between these two extreme cases and to study partially fair predictors. Within this framework we precisely quantify the cost in risk induced by the introduction of the fairness constraint. We put forward a statistical minimax setup and derive a general problem-dependent lower bound on the risk of any estimator satisfying α-relative improvement constraint. We illustrate our framework on a model of linear regression with Gaussian design and systematic group-dependent bias, deriving matching (up to absolute constants) upper and lower bounds on the minimax risk under the introduced constraint. Finally, we perform a simulation study of the latter setup.

Based on Evgenii Chzhen and Nicolas Schreuder (2020a). "A minimax framework for quantifying risk-fairness trade-off in regression". In: *arXiv preprint arXiv:2007.14265*.

Contents

7.1 Introduction

Data driven algorithms are deployed in almost all areas of modern daily life and it becomes increasingly more important to adequately address the fundamental issue of historical biases present in the data (Barocas, Hardt, and Narayanan 2019). The goal of algorithmic fairness is to bridge the gap between the statistical theory of decision making and the understanding of justice, equality, and diversity. The literature on fairness is broad and its volume increases day by day, we refer the reader to (Mehrabi et al. 2019; Barocas, Hardt, and Narayanan 2019) for a general introduction on the subject and to (Oneto and Chiappa 2020; Barrio, Gordaliza, and Loubes 2020) for reviews of the most recent theoretical advances.

Basically, the mathematical definitions of fairness can be divided into two groups (Dwork et al. 2012): *individual fairness* and *group fairness*. The former notion reflects the principle that similar individuals must be treated similarly, which translates into Lipschitz type constraints on possible prediction rules. The latter defines fairness on population level via (conditional) statistical independence of a prediction from a sensitive attribute (*e.g.,* gender, ethnicity). A popular formalization of such notion is through the *Demographic Parity* constraint, initially introduced in the context of binary classification (Calders, Kamiran, and Pechenizkiy 2009). Despite of some limitations (Hardt, Price, and Srebro 2016), the concept of Demographic Parity is natural and suitable for a range of applied problems (Köeppen, Yoshida, and Ohnishi 2014; Zink and Rose 2019).

In this work we study the regression problem of learning a real-valued prediction function, which complies with an approximate notion of Demographic Parity while minimizing expected squared loss.

Unlike its classification counterpart, the problem of fair regression has received far less attention in the literature. However, as argued by Agarwal, Dudík, and Wu 2019, classifiers only provide binary decisions, while in practice final decisions are taken by humans based on predictions from the machine. In this case a continuous prediction is more informative than a binary one and justifies the need for studying fairness in the regression framework.

Notation For any univariate probability measure μ we denote by F_μ (*resp.* F_μ^{-1}) the cumulative distribution function (*resp.* the quantile function) of μ. For two random variables U and V we denote by $\mathrm{Law}(U \mid V{=}v)$ the conditional distribution of the random variable $U \mid V{=}v$ and we write $U \overset{d}{=} V$ to denote their equality in distribution. For any integer $K \geq 1$, we denote by Δ^{K-1} the probability simplex in $\mathbb{R}^K$ and we write $[K] = \{1, \dots, K\}$. For any $a, b \in \mathbb{R}$ we denote by $a \vee b$ (*resp.* $a \wedge b$) the maximum (*resp.* the minimum) between a, b. We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ with finite second-order moment.

7.2 Problem statement and contributions

We study the regression problem when a sensitive attribute is available. The statistician observes triplets $(\boldsymbol{X}_1, S_1, Y_1), \dots, (\boldsymbol{X}_n, S_n, Y_n) \in \mathbb{R}^p \times [K] \times \mathbb{R}$, which are connected by the following regression-type relation

$$Y_i = f^*(\boldsymbol{X}_i, S_i) + \xi_i \ , \qquad i \in [n] \ , \tag{7.1}$$

where $\xi_i \in \mathbb{R}$ is a centered random variable and $f^* : \mathbb{R}^p \times [K] \to \mathbb{R}$ is the regression function. Here for each $i \in [n]$, $\boldsymbol{X}_i$ is a feature vector taking values in $\mathbb{R}^p$, S_i is a sensitive attribute taking values in $[K]$, and Y_i is a real-valued dependent variable. A prediction is any measurable function of the form $f : \mathbb{R}^p \times [K] \to \mathbb{R}$. We define the risk of a prediction function f via the $\mathbb{L}_2$ distance to the regression function f^* as

$$\mathcal{R}(f) := \|f - f^*\|_2^2 := \sum_{s=1}^{K} w_s \mathbb{E}\left[(f(\boldsymbol{X}, S) - f^*(\boldsymbol{X}, S))^2 \mid S = s\right] \ , \qquad \textbf{(Risk measure)}$$

where $\mathbb{E}[\cdot \mid S{=}s]$ is the expectation *w.r.t.* the distribution of the features $\boldsymbol{X}$ in the group $S = s$ and $\boldsymbol{w} = (w_1, \dots, w_K)^\top \in \Delta^{K-1}$ is a probability vector, which weights the group-wise risks.

For any $s \in [K]$ define ν_s^* as $\mathrm{Law}(f^*(\boldsymbol{X}, S) \mid S{=}s)$ – the distribution of the optimal prediction inside the group $S = s$. Throughout this work we make the following assumption on those measures, which is, for instance, satisfied in linear regression with Gaussian design.

Assumption 7.2.1. *The measures $\{\nu_s^*\}_{s \in [K]}$ are non-atomic and have finite second moments.*

7.2.1 Regression with fairness constraints

Any predictor f induces a group-wise distribution of the predicted outcomes $\mathrm{Law}(f(\boldsymbol{X}, S) \mid S{=}s)$ for $s \in [K]$. The high-level idea of *group fairness* notions is to bound or diminish an eventual discrepancy between these distributions.

We define the *unfairness* of a predictor f as the sum of the weighted distances between $\{\mathrm{Law}(f(\boldsymbol{X}, S) \mid S{=}s)\}_{s\in[K]}$ and their common barycenter *w.r.t.* the Wasserstein-2 distance[1]:

$$\mathcal{U}(f) := \min_{\nu\in\mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{K} w_s W_2^2\big(\,\mathrm{Law}(f(\boldsymbol{X}, S) \mid S{=}s), \nu\big) \ . \qquad \textbf{(Unfairness measure)}$$

In particular, since the Wasserstein-2 distance is a metric on the space probability distributions with finite second-order moment $\mathcal{P}_2(\mathbb{R}^d)$, a predictor f is such that $\mathcal{U}(f) = 0$ if and only if it satisfies the Demographic Parity (DP) constraint defined as

$$(f(\boldsymbol{X}, S) \mid S = s) \overset{d}{=} (f(\boldsymbol{X}, S) \mid S = s'), \quad \forall s, s' \in [K] \ . \qquad \textbf{(DP)}$$

Exact DP is not necessarily desirable in practice and it is common in the literature to consider *relaxations* of this constraint. In this work we introduce the α-*Relative Improvement* (α-RI) constraint – a novel DP relaxation based on our unfairness measure. We say that a predictor f satisfies the α-RI constraint for some $\alpha \in [0, 1]$ if its unfairness is at most an α fraction of the unfairness of the regression function f^*, that is, $\mathcal{U}(f) \le \alpha\mathcal{U}(f^*)$. Importantly, the fairness requirement is stated relatively to the unfairness of the regression function f^*, which allows to make a more informed choice of α.

Formally, for a fixed $\alpha \in [0, 1]$, the goal of a statistician in our framework is to build an estimator $\widehat{f}$ using data, which enjoys two guarantees (with high probability)

$$\textbf{α-RI guarantee:} \quad \mathcal{U}(\widehat{f}) \le \alpha\mathcal{U}(f^*) \qquad \text{and} \qquad \textbf{Risk guarantee:} \quad \mathcal{R}(\widehat{f}) \le r_{n,\alpha,f^*} \ .$$

The former ensures that $\widehat{f}$ satisfies the α-RI constraint. In the latter guarantee we seek the sequence r_{n,α,f^*} being as small as possible in order to quantify *two effects*: the introduction of the α-RI *fairness constraint* and the *statistical estimation*. We note that r_{n,α,f^*} depends on the sample size n, the fairness parameter α, as well as the regression function f^* to be estimated, we clarify the reason for this dependency later in the text.

7.2.2 Contributions

The first natural question that we address is: assuming that the underlying distribution of $X \mid S$ and the regression function f^* are known, which prediction rule f^*_α minimizes the expected squared loss under the α-RI constraint $\mathcal{U}(f^*_\alpha) \le \alpha\mathcal{U}(f^*)$? To answer this question we shift the discussion to the population level and define a collection $\{f^*_\alpha\}_{\alpha\in[0,1]}$ of *oracle* α-*RI* indexed by the parameter α as

$$f^*_\alpha \in \arg\min\{\mathcal{R}(f) \,:\, \mathcal{U}(f) \le \alpha\mathcal{U}(f^*)\} \ , \qquad \forall\alpha \in [0, 1] \ . \qquad \textbf{(Oracle α-RI)}$$

For $\alpha = 0$ the predictor f^*_0 corresponds to the optimal fair predictor in the sense of DP while for $\alpha = 1$ the corresponding predictor f^*_1 coincides with the regression function f^*. Those two extreme cases have been previously studied but, up to our knowledge, nothing is known about those "partially fair" predictors. Our study of the family $\{f^*_\alpha\}_{\alpha\in[0,1]}$ serves as a basis for our statistical framework and analysis. It also reveals the intrinsic interplay of the fairness constraint with the risk measure.

The contributions of this work can be roughly split into three interconnected groups:

[1] See Appendix 7.8.1 for a reminder on Wasserstein distances.

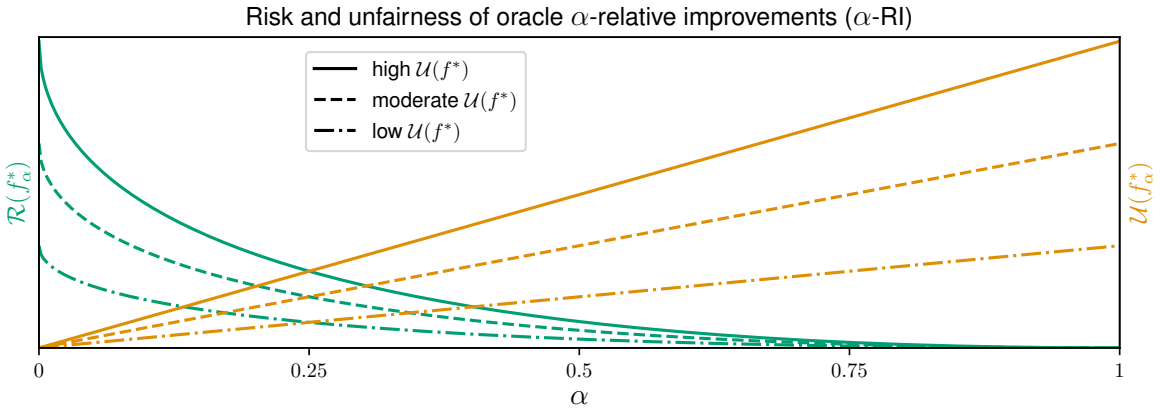


Figure 7.1: Risk $\mathcal{R}$ and unfairness $\mathcal{U}$ of α-RI oracles $\{f^*_\alpha\}_{\alpha\in[0,1]}$. Green curves (decreasing, convex) correspond to the risk, while orange curves (increasing, linear) correspond to the unfairness. Each pair of curves (solid, dashed, dashed dotted) corresponds to three regimes: high, moderate, and low unfairness of the regression function f^* respectively.

1. We provide a theoretical study of the family of oracle α-RI $\{f^*_\alpha\}_{\alpha\in[0,1]}$ on the population level;

2. We introduce a minimax statistical framework and derive a general problem-dependent minimax lower bound for the problem of regression under the α-RI constraint;

3. We derive minimax optimal rate of convergence for the statistical model of linear regression with systematic group-dependent bias and Gaussian design under the α-RI constraint.

Properties of oracle α-RI $\{f^*_\alpha\}_{\alpha\in[0,1]}$ It has been shown that, under the squared loss, the optimal fair predictor f^*_0 can be obtained as the solution of a Wasserstein-2 barycenter problem (Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c). In Section 7.4 we study the whole family $\{f^*_\alpha\}_{\alpha\in[0,1]}$ for arbitrary choice of $\alpha \in [0,1]$. To provide complete characterization of $\{f^*_\alpha\}_{\alpha\in[0,1]}$ we derive Lemma 7.4.3, which could be of independent interest. This result can be summarized as follows: given a fixed collection of points $a_1, \dots, a_K$ in an abstract metric space $(\mathcal{X}, d)$, if one walks along the (constant speed) geodesics starting from a_s and leading to their (weighted) barycenter until it reaches a proportion α of the full path, then these intermediate points $b_1, \dots, b_K$ minimize the weighted distance to the initial points while being α-closer to their own barycenter. This abstract result enables us to characterize explicitly oracle α-RI $\{f^*_\alpha\}_{\alpha\in[0,1]}$. In particular, we show that the family of oracle α-RI $\{f^*_\alpha\}_{\alpha\in[0,1]}$ admits a simple structure: for any $\alpha \in [0,1]$ the prediction f^*_α is the point-wise convex combination of the regression function $f^* \equiv f^*_1$ and the optimal fair predictor f^*_0, that is,

$$f^*_\alpha(\boldsymbol{x}, s) = \sqrt{\alpha} f^*_1(\boldsymbol{x}, s) + (1-\sqrt{\alpha}) f^*_0(\boldsymbol{x}, s), \qquad \forall\, (\boldsymbol{x}, s) \in \mathbb{R}^p \times [K] \ .$$

The final contribution of Section 7.4 is the quantification of the risk-fairness trade-off on the population level. In particular, Lemma 7.4.5 establishes that for every $\alpha \in [0,1]$ it holds that

$$\mathcal{R}(f^*_\alpha) = (1-\sqrt{\alpha})^2 \mathcal{R}(f^*_0) \quad \text{and} \quad \mathcal{U}(f^*_\alpha) = \alpha \mathcal{U}(f^*) \ .$$

Observe that f^*_0, which is the optimal fair predictor in terms of DP, has the highest risk and the lowest unfairness, while the situation is reversed for $f^*_1 \equiv f^*$ – the risk is the lowest and the unfairness is the highest. Since the function $\alpha \to (1-\sqrt{\alpha})^2$ grows rapidly in the vicinity of zero, even a mild relaxation of the exact fairness constraint ($\alpha = 0$) yields a noticeable improvement in terms of the risk while having a low unfairness inflation. For instance, the risk of $f^*_{1/2}$ is only around 8.5% of the risk of f^*_0, while its fairness is two times better than that of f^*. This observation is illustrated in Figure 7.1.

Minimax framework In order to quantify the *statistical* price of fairness, in Section 7.5 we propose a minimax framework and in Section 7.5.1 we derive a general problem-dependent lower bound on the minimax risk of estimators satisfying the α-RI constraint. Statistical study of the model in Eq. (7.1) typically requires additional assumptions to provide meaningful statistical guarantees. Classically, one chooses a set $\mathcal{F}$ of possible candidates for the regression function f^* (*e.g.,* linear functions) and, possibly, introduces additional conditions on nuisance parameters of the model via some set Θ (*e.g.,* variance of the noise). The goal of our lower bound is to understand fundamental limits of the problem of prediction under α-RI constraint in arbitrary statistical model for Eq. (7.1). To this end, we show in Theorem 7.5.3 that *any estimator* $\widehat{f}$ *satisfying the* α*-RI* constraint with high probability must incur

$$\mathcal{R}(\widehat{f}) \geq \delta_n(\mathcal{F}, \Theta) \vee (1-\sqrt{\alpha})^2 \mathcal{U}(f^*) \ ,$$

where $\delta_n(\mathcal{F}, \Theta)$ is the rate one would obtain *without* restricting the set of possible estimators.

Application to linear model The goal of Section 7.6 is to demonstrate that the general problem-dependent lower bound does indeed yield minimax optimal rates. To this end, we apply our machinery to the problem of linear regression with systematic bias formalized by the following linear model

$$Y_i = \langle \boldsymbol{X}_i, \boldsymbol{\beta}^* \rangle + b^*_{S_i} + \xi_i, \quad i = 1, \dots, n \ ,$$

where the ξ_i's are i.i.d. zero mean Gaussian with variance σ^2 and the p-dimensional covariates $\{\boldsymbol{X}_i\}_{i=1}^n$ are i.i.d. Gaussian random vectors. We propose an estimator $\widehat{f}$ which, with probability at least $1-\delta$, satisfies $\mathcal{U}(\widehat{f}) \leq \alpha \mathcal{U}(f^*)$ and achieves the following minimax optimal rate

$$\mathcal{R}(\widehat{f}) \asymp \left\{ \sigma^2 \left(\frac{p+K}{n} + \frac{\log(1/\delta)}{n} \right) \right\} \bigvee \left\{ (1-\sqrt{\alpha})^2 \mathcal{U}(f^*) \right\} \ .$$

Finally, we conduct a simulation study of the proposed estimator $\widehat{f}$ and compare its performance with more straightforward approaches in terms of unfairness and risk.

7.3 Prior and related works

Until very recently, contributions on fair regression were almost exclusively focused on the practical incorporation of proxy fairness constraints in classical learning methods, such as random forest, ridge regression, kernel based methods to name a few (Calders et al. 2013; Komiyama and Shimao 2017; Berk et al. 2017; Pérez-Suay et al. 2017; Raff, Sylvester, and Mills 2018; Fitzsimons et al. 2018). Several works empirically study the impact of (relaxed) fairness constraints on the risk (Bertsimas, Farias, and Trichakis 2012; Zliobaite 2015; Haas 2019; Wick, Panda, and Tristan 2019; Zafar et al. 2017). Yet, the problem of precisely quantifying the effect of such constraints on the risk has not been tackled.

More recently, statistical and learning guarantees for fair regression were derived (Agarwal, Dudík, and Wu 2019; Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c; Chiappa et al. 2020; Fitzsimons et al. 2019; Plečko and Meinshausen 2019; Chzhen et al. 2020a). The closest works to our contribution are that of Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c; Chiappa et al. 2020, who draw a connection between the problem of exactly fair regression of demographic parity and the multi-marginal optimal transport formulation (Gangbo and Święch 1998; Agueh and Carlier 2011).

As already mentioned in the previous section, considering predictors which satisfy the DP constraint incurs an unavoidable price in terms of the risk. Depending on the application at hand, this price might or might not be reasonable. However, since the notion of DP is completely fairness driven, it does not allow to quantify the price of considering "fairer" predictions than the regression function f^*. For this reason, several contributions relax this constraint, forcing a milder fairness requirement. A natural idea is to define a functional $\mathcal{U}$ which quantifies the violation of the DP constraint and to declare a prediction approximately fair if this functional does not exceed a user pre-specified threshold. In recent years a large variety of such relaxations has been proposed: correlation based (Baharlouei et al. 2019; Mary, Calauzènes, and El Karoui 2019; Komiyama et al. 2018); Kolmogorov-Smirnov distance (Agarwal, Dudík, and Wu 2019); Mutual information (Steinberg et al. 2020; Steinberg, Reid, and O'Callaghan 2020); Total Variation distance (Oneto, Donini, and Pontil 2019b; Oneto et al. 2019); Equality of means and higher moment matching (Raff, Sylvester, and Mills 2018; Fitzsimons et al. 2019; Calders et al. 2013; Berk et al. 2017; Olfat et al. 2020; Donini et al. 2018); Maximum Mean Discrepancy (Quadrianto and Sharmanska 2017; Madras et al. 2018); Wasserstein distance (Chiappa et al. 2020; Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c; Gordaliza et al. 2019).

7.3.1 Other notions of unfairness

The most common relaxations of the Demographic Parity constraint are based on the Total Variation (TV) and the Kolmogorov-Smirnov (KS) distances (Agarwal, Dudík, and Wu 2019; Oneto, Donini, and Pontil 2019a; Agarwal et al. 2018; Chzhen et al. 2020a). There are various ways to use the TV or KS in order to build a functional $\mathcal{U}$, which quantifies the violation of the DP constraint. To compare those measures of discrepancy with the one that we introduce

in our work, we define $\mathcal{U}_{\mathrm{TV}}$ and $\mathcal{U}_{\mathrm{KS}}$ as follows

$$\textbf{TV unfairness:} \qquad \mathcal{U}_{\mathrm{TV}}(f) := \sum_{s\in[K]} \mathrm{TV}\left(\mathrm{Law}(f(\boldsymbol{X},S)\mid S=s), \mathrm{Law}(f(\boldsymbol{X},S))\right) \ ,$$

$$\textbf{KS unfairness:} \qquad \mathcal{U}_{\mathrm{KS}}(f) := \sum_{s\in[K]} \mathrm{KS}\left(\mathrm{Law}(f(\boldsymbol{X},S)\mid S=s), \mathrm{Law}(f(\boldsymbol{X},S))\right) \ .$$

Using these notions, one wishes to study those predictors f which satisfy relaxed fairness constraint $\mathcal{U}_{\square}(f) \le \varepsilon$, where $\square$ is KS or TV and $\varepsilon \ge 0$ is a user specified parameter. Note that since both KS and TV are metrics, setting $\varepsilon = 0$ is equivalent to the DP constraint. Meanwhile, for $\varepsilon > 0$ these formulations allow some slack. It is known that the TV distance is rather strong and extremely sensitive to small changes in distributions which is the major drawback of the TV unfairness. This limitation can be addressed by the KS unfairness due to an obvious relation $\mathcal{U}_{\mathrm{KS}}(f) \le \mathcal{U}_{\mathrm{TV}}(f)$.

In our work we argue that the introduced notion of unfairness $\mathcal{U}$ is better suited for the problem of regression with squared loss under fairness constraint. Indeed, we prove in Lemma 7.4.5 that $\mathcal{U}$ can be naturally connected to the squared risk and allows to give a precise quantification of the risk-fairness trade-off. This result is the major advantage of $\mathcal{U}$ over both $\mathcal{U}_{\mathrm{KS}}$ and $\mathcal{U}_{\mathrm{TV}}$. Nevertheless, it is still interesting to understand whether a more popular KS unfairness can be related to $\mathcal{U}$ that we introduce. In Appendix we prove the following connection.

Proposition 7.3.1. *Fix some predictor $f : \mathbb{R}^p \times [K] \to \mathbb{R}$. Assume that the distribution $\mathrm{Law}(f(\boldsymbol{X},S) \mid S{=}s) \in \mathcal{P}_2(\mathbb{R})$ and that it admits a density bounded by $C_{f,s} > 0$ for all $s \in [K]$. Then* [2]

$$\mathcal{U}_{\mathrm{KS}}(f) \le \|{}^1\!/\!_{\boldsymbol{w}}\|_\infty \sqrt{8\bar{C}_f}\cdot \mathcal{U}^{1/4}(f) \ ,$$

where $\bar{C}_f = \sum_{s=1}^K w_s C_{f,s}$ and ${}^1\!/\!_{\boldsymbol{w}} = ({}^1\!/\!_{w_1}, \dots, {}^1\!/\!_{w_K})^\top$.

The latter result indicates that if one can control the unfairness $\mathcal{U}$ introduced in this work, one also has some control over the KS unfairness. Note that the leading constant of the previous bound depends on the predictor f. More precisely, this constant corresponds to the upper bound on the density of $f(\boldsymbol{X}, S)$.

Another advantage of the introduced unfairness measure, and, in particular, the notion of α-relative improvement is the fact that the parameter α has a clear practical interpretation, while the interpretation of ε is not intuitive. Of course, using $\mathcal{U}_{\mathrm{KS}}$ or $\mathcal{U}_{\mathrm{TV}}$ one can also define unfairness of a predictor f relatively to the regression function f^*. However, due to completely different geometries induced by $\mathcal{R}$ in the space of functions and by $\mathcal{U}_{\mathrm{KS\,/\,TV}}$ in the space of distributions, precise theoretical study of such formulations is notoriously complicated if possible.

7.3.2 Optimal transport and fair regression

The use of optimal transport tools in the study of fairness is relatively recent. Initially, contributions in this direction were mainly dealing with the problem of binary classification (Gordaliza

[2] One can erase the term $\|{}^1\!/\!_{\boldsymbol{w}}\|_\infty$ from the bound defining $\mathcal{U}_{\mathrm{KS}}(f)$ as $\sum_{s\in[K]} w_s \,\mathrm{KS}\left(\mathrm{Law}(f(\boldsymbol{X},S)\mid S=s), \mathrm{Law}(f(\boldsymbol{X},S))\right)$.

et al. 2019; Jiang et al. 2019). Later on, the tools of the optimal transport theory migrated to the setup of fair regression (Chiappa et al. 2020; Chzhen et al. 2020c; Le Gouic, Loubes, and Rigollet 2020). The main theoretical motivation to consider $\mathcal{U}$ instead of the KS and TV unfairnesses lies in the following recent result due to (Chzhen et al. 2020c; Le Gouic, Loubes, and Rigollet 2020).

Theorem 7.3.2 (Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c)**.** *Let Assumption 7.2.1 be satisfied, then*

$$\min\left\{\mathcal{R}(f) \,:\, (f(\boldsymbol{X},S) \mid S=s) \overset{d}{=} (f(\boldsymbol{X},S) \mid S=s') \;\forall s,s' \in [K]\right\} = \mathcal{U}(f^*) \;. \tag{7.2}$$

Moreover, the distribution of the minimizer of the problem on the l.h.s. *is given by*

$$\operatorname*{arg\,min}_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{K} w_s W_2^2 \left(\mathrm{Law}(f^*(\boldsymbol{X},S) \mid S{=}s), \nu\right) \;.$$

An important consequence of Theorem 7.3.2 is that it puts the risk $\mathcal{R}$ and the unfairness $\mathcal{U}$ – two conflicting quantities – on the same scale. In particular, it allows to measure both fairness and risk using the same unit measurements, hence, study the trade-off between the two. In order to build our framework, we remark that since W_2 is a metric then the problem on the *l.h.s.* of Eq. (7.2) can be equivalently written as $\min\{\mathcal{R}(f) \,:\, \mathcal{U}(f) \le 0 \times \mathcal{U}(f^*)\}$. Moreover, one can observe that the regression function $f^* \in \min\{\mathcal{R}(f) \,:\, \mathcal{U}(f) \le 1 \times \mathcal{U}(f^*)\}$. Thus, a natural relaxation of the above formulation is the introduced notion of *α-relative improvement*, which interpolates between the exactly fair predictor f_0^* and the regression function $f_1^* \equiv f^*$. In this retrospect, the result of Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c provides characterization of f_0^* but says nothing about the whole family of oracle α-RI $\{f_\alpha^*\}_{\alpha\in[0,1]}$.

7.4 Oracle α-relative improvement

This section is devoted to the study of the α-relative improvement f_α^* on population level, that is, in this section we study

$$f_\alpha^* \in \operatorname{arg\,min}\{\mathcal{R}(f) \,:\, \mathcal{U}(f) \le \alpha\,\mathcal{U}(f^*)\} \quad , \qquad \forall \alpha \in [0,1] \;. \tag{7.3}$$

The next result establishes a closed form solution to the minimization Problem (7.3) under Assumption 7.2.1 for any value of $\alpha \in [0,1]$.

Proposition 7.4.1. *Let Assumption 7.2.1 be satisfied, then for all $\alpha \in [0,1]$ and all $(\boldsymbol{x},s) \in \mathbb{R}^p \times [K]$ (up to a set of null measure) it holds that*

$$\begin{aligned} f_\alpha^*(\boldsymbol{x},s) &= \sqrt{\alpha} f^*(\boldsymbol{x},s) + (1{-}\sqrt{\alpha}) \sum_{s'=1}^{K} w_{s'} F^{-1}_{\nu^*_{s'}} \circ F_{\nu^*_s} \circ f^*(\boldsymbol{x},s) \\ &= \sqrt{\alpha} f_1^*(\boldsymbol{x},s) + (1{-}\sqrt{\alpha}) f_0^*(\boldsymbol{x},s) \;. \end{aligned}$$

Recall that $f^* = f_1^*$, hence the α-relative improvement f_α^* is the point-wise convex combination of exactly fair prediction f_0^* and the regression function f_1^*. Besides, setting $\alpha = 0$ we recover the result of Chzhen et al. 2020c; Le Gouic, Loubes, and Rigollet 2020 as a particular case of our framework. The set of oracle α-RI $\{f_\alpha^*\}_{\alpha\in[0,1]}$ satisfies the following properties.

1. **Risk and fairness monotonicity:** if $\alpha \leq \alpha'$, then $\mathcal{R}(f^*_\alpha) \geq \mathcal{R}(f^*_{\alpha'})$ and $\mathcal{U}(f^*_\alpha) \leq \mathcal{U}(f^*_{\alpha'})$.
2. **Point-wise convexity:** for all $\alpha, \alpha' \in [0,1]$ and all $\tau \in [0,1]$ it holds that $\tau f^*_\alpha + (1-\tau) f^*_{\alpha'} \in \{f^*_\alpha\}_{\alpha \in [0,1]}$. Moreover $\tau f^*_\alpha + (1-\tau) f^*_{\alpha'} = f^*_{\bar{\alpha}}$ with $\bar{\alpha} = (\tau\sqrt{\alpha} + (1-\tau)\sqrt{\alpha'})^2$.
3. **Order preservation:** for all $s \in [K], \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$, if $f^*(\boldsymbol{x}, s) \geq f^*(\boldsymbol{x}', s)$, then for all $\alpha \in [0,1]$ it holds that $f^*_\alpha(\boldsymbol{x}, s) \geq f^*_\alpha(\boldsymbol{x}', s)$.

The first property is intuitive and does not require the result of Proposition 7.4.1. The second property can be directly derived using the expression of f^*_α and it describes additional algebraic structure of the family $\{f^*_\alpha\}_{\alpha \in [0,1]}$. The third group-wise order preserving property of f^*_α is particularly attractive. Its proof is straightforward after the observation that $F_{\nu^*_s}$ and $\sum_{s'=1}^K w_{s'} F^{-1}_{\nu^*_{s'}}$ are non-decreasing functions and the fact that the composition of two non-decreasing functions is non-decreasing. For the special case of $\alpha = 0$, this observation has already been made in (Chzhen et al. 2020c) and a practical algorithm that follows the group-wise order preservation property was proposed by Plečko and Meinshausen 2019. In words, this property says: given any two individuals $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^p$ from the same sensitive group $s \in [K]$, if the optimal prediction $f^*(\boldsymbol{x}, s)$ for $\boldsymbol{x}$ is larger than that for $\boldsymbol{x}'$, then across all levels α of fairness parameter the oracle α-RI f^*_α is not changing this order.

7.4.1 An abstract geometric lemma

The proof of Proposition 7.4.1 relies on a abstract geometric result, Lemma 7.4.3, which might be interesting on its own. First, let us introduce the following definition, which asks for existence of finitely supported barycenters in a metric space $(\mathcal{X}, d)$.

Definition 7.4.2. *We say that a metric space $(\mathcal{X}, d)$ satisfies the barycenter property if for any weights $\boldsymbol{w} \in \Delta^{K-1}$ and tuple $\boldsymbol{a} = (a_1, \ldots, a_K) \in \mathcal{X}^K$ there exists a barycenter*

$$C_{\boldsymbol{a}_{\boldsymbol{w}}} \in \arg\min_{C \in \mathcal{X}} \sum_{s=1}^K w_s d^2(a_s, C) \ .$$

Moreover, for any tuple $\boldsymbol{a} = (a_1, \ldots, a_K) \in \mathcal{X}^K$ we denote[3] by $C_{\boldsymbol{a}_{\boldsymbol{w}}}$ a barycenter of $\boldsymbol{a}$ weighted by $\boldsymbol{w} \in \Delta^{K-1}$.

Lemma 7.4.3 (Abstract geometric lemma)**.** *Let $(\mathcal{X}, d)$ be a metric space satisfying the barycenter property. Let $\boldsymbol{a} = (a_1, \ldots, a_K) \in \mathcal{X}^K$, $\boldsymbol{w} = (w_1, \ldots, w_K)^\top \in \Delta^{K-1}$ and let $C_{\boldsymbol{a}}$ be a barycenter of $\boldsymbol{a}$ with respect to weights $\boldsymbol{w}$. For a fixed $\alpha \in [0,1]$ assume that there exists $\boldsymbol{b} = (b_1, \ldots, b_K) \in \mathcal{X}^K$ which satisfies*

$$d(a_s, C_{\boldsymbol{a}}) = d(a_s, b_s) + d(b_s, C_{\boldsymbol{a}}) \ , \qquad s = 1, \ldots, K \ , \tag{P_1}$$

$$d(b_s, a_s) = (1 - \sqrt{\alpha}) d(a_s, C_{\boldsymbol{a}}) \ , \qquad s = 1, \ldots, K \ . \tag{P_2}$$

Then, $\boldsymbol{b}$ is a solution of

$$\inf_{\boldsymbol{b} \in \mathcal{X}^K} \left\{ \sum_{s=1}^K w_s d^2(b_s, a_s) \ : \ \sum_{s=1}^K w_s d^2(b_s, C_{\boldsymbol{b}}) \leq \alpha \sum_{s=1}^K w_s d^2(a_s, C_{\boldsymbol{a}}) \right\} \ . \tag{7.4}$$

[3] When there is no ambiguity in the weights $\boldsymbol{w}$ we simply write $C_{\boldsymbol{a}}$.

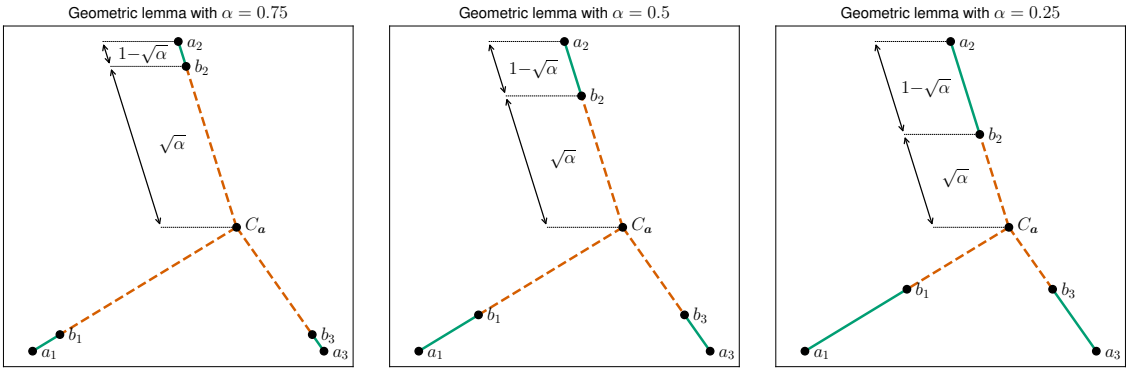


Figure 7.2: Illustration of Lemma 7.4.3 for $(\mathcal{X}, d) = (\mathbb{R}^2, \|\cdot\|_2)$ and $\alpha \in \{0.25, 0.5, 0.75\}$. The initial points a_1, a_2, a_3 are the vertices of an isosceles triangle. The weights are set as follows: $w_1 = 0.1$, $w_2 = 0.4$ and $w_3 = 0.5$.

Remark 7.4.4. *Property (P_1) essentially requires that each b_i lies on the geodesic between a_i and $C_{\boldsymbol{a}}$ while Property (P_2) specifies the location of b_i on this geodesic: b_i should be $(1-\sqrt{\alpha})$ times closer to a_i, than $C_{\boldsymbol{a}}$ to a_i. An illustration provided on Figure 7.2 describes these properties in Euclidean geometry. For general case, the straight lines should be replaced by geodesics.*

The setting of Lemma 7.4.3 is quite general and only requires existence of barycenters (also known as the Fréchet means) for any finite weighted combination of points in accordance with Definition 7.4.2. For our purposes, Lemma 7.4.3 will be applied to the metric space $(\mathcal{X}, d) = (\mathcal{P}_2(\mathbb{R}), W_2)$. We refer to (Agueh and Carlier 2011; Le Gouic and Loubes 2017) who investigate and prove the existence of Wasserstein barycenters of random probabilities defined on geodesic spaces.

Proof of Lemma 7.4.3. Fix some $\boldsymbol{a} = (a_1, \dots, a_K) \in \mathcal{X}^K$, $\boldsymbol{w} = (w_1, \dots, w_K)^\top \in \Delta^{K-1}$ and let $C_{\boldsymbol{a}}$ be a barycenter of $\boldsymbol{a}$ with respect to weights $\boldsymbol{w}$. Fix $\alpha \in [0, 1]$ and any $\boldsymbol{b} = (b_1, \dots, b_K) \in \mathcal{X}^K$ which satisfies properties (P_1)–(P_2). Let $\boldsymbol{b}_k = (b_1^k, \dots, b_K^k) \in \mathcal{X}^K$ be a minimizing sequence of the problem (7.4) and for any $\boldsymbol{b}' = (b_1', \dots, b_K') \in \mathcal{X}^K$ denote by $G(\boldsymbol{b}') = \sum_{s=1}^K w_s d^2(b_s', a_s)$ the objective function of the problem (7.4). Then, by the definition of a minimizing sequence, the following two properties hold

$$\lim_{k\to\infty} G(\boldsymbol{b}_k) = \inf_{\boldsymbol{b}\in\mathcal{X}^K} \left\{ G(\boldsymbol{b}) \,:\, \sum_{s=1}^K w_s d^2(b_s, C_{\boldsymbol{b}}) \le \alpha \sum_{s=1}^K w_s d^2(a_s, C_{\boldsymbol{a}}) \right\} \,, \tag{7.5}$$

$$\sum_{s=1}^K w_s d^2(b_s^k, C_{\boldsymbol{b}_k}) \le \alpha \sum_{s=1}^K w_s d^2(a_s, C_{\boldsymbol{a}}) \,, \quad \forall k \in \mathbb{N} \,. \tag{7.6}$$

Furthermore, using properties (P_1)–(P_2) we deduce that

$$\sum_{s=1}^K w_s d^2(b_s, C_{\boldsymbol{b}}) \overset{(a)}{=} \sum_{s=1}^K w_s d^2(b_s, C_{\boldsymbol{a}}) \overset{(P_1)}{=} \sum_{s=1}^K w_s \left(d(a_s, C_{\boldsymbol{a}}) - d(a_s, b_s)\right)^2 \overset{(P_2)}{=} \alpha \sum_{s=1}^K w_s d^2(a_s, C_{\boldsymbol{a}}) \,,$$

where (a) follows from Lemma 7.9.3 in appendix. Therefore, $\boldsymbol{b} = (b_1, \dots, b_s) \in \mathcal{X}^K$ is feasible for the problem (7.4).

By Lemma 7.9.2 it holds for all $k \in \mathbb{N}$ that

$$\left\{\sum_{s=1}^{K} w_s d^2(a_s, C_{\boldsymbol{b}_k})\right\}^{1/2} \leq \left\{\sum_{s=1}^{K} w_s d^2(a_s, b_s^k)\right\}^{1/2} + \left\{\sum_{s=1}^{K} w_s d^2(b_s^k, C_{\boldsymbol{b}_k})\right\}^{1/2}$$
$$= G^{1/2}(\boldsymbol{b}_k) + \left\{\sum_{s=1}^{K} w_s d^2(b_s^k, C_{\boldsymbol{b}_k})\right\}^{1/2} .$$

We continue using the definition of $C_{\boldsymbol{a}}$ and Eq. (7.6) to obtain for all $k \in \mathbb{N}$

$$\left\{\sum_{s=1}^{K} w_s d^2(a_s, C_{\boldsymbol{a}})\right\}^{1/2} \leq \left\{\sum_{s=1}^{K} w_s d^2(a_s, C_{\boldsymbol{b}_k})\right\}^{1/2} \leq G^{1/2}(\boldsymbol{b}_k) + \sqrt{\alpha}\left\{\sum_{s=1}^{K} w_s d^2(a_s, C_{\boldsymbol{a}})\right\}^{1/2} ,$$

which after rearranging implies that

$$(1-\sqrt{\alpha})\left\{\sum_{s=1}^{K} w_s d^2(a_s, C_{\boldsymbol{a}})\right\}^{1/2} \leq G^{1/2}(\boldsymbol{b}_k), \qquad \forall k \in \mathbb{N} .$$

Finally, using property (P_2) we derive that

$$G(\boldsymbol{b}) \leq G(\boldsymbol{b}_k), \qquad \forall k \in \mathbb{N} .$$

Recall that we have already shown that $\boldsymbol{b}$ is feasible for the problem (7.4), hence taking the limit *w.r.t.* to k concludes the proof of Lemma 7.4.3. □

The complete proof of Proposition 7.4.1 is omitted in the main body. We only provide a short intuition.

Sketch of the proof. The idea of the proof is to apply Lemma 7.4.3 with $(\mathcal{X}, d) = (\mathcal{P}_2(\mathbb{R}), W_2)$ and with measures $a_s := \nu_s^*$, which belong to $\mathcal{P}_2(\mathbb{R})$ due to Assumption 7.2.1. Then, we need to construct measures $\boldsymbol{b} = (b_1, \ldots, b_K)^\top \in \mathcal{P}_2^K(\mathbb{R})$, which satisfy the properties (P_1)–(P_2). To this end, let γ_s be the (constant-speed) geodesic between a_s and $C_{\boldsymbol{a}}$ *i.e.*, $\gamma_s(0) = a_s$, $\gamma_s(1) = C_{\boldsymbol{a}}$. We define $b_s := \gamma_s(1-\sqrt{\alpha})$ for $s \in [K]$, similarly to the intuition provided by Figure 7.2. One can verify that that $\boldsymbol{b} = (b_s)_{s\in[K]}$ satisfies (P_1) and (P_2). Then, by Lemma 7.4.3 we know that $\boldsymbol{b}$ solves the minimization problem in Eq. (7.4). For the final part of the proof we propagate the optimality of $\boldsymbol{b}$ in the space of distributions to the optimality of f_α^* in the space of predictions using the assumption that $\boldsymbol{a}$ admits a density and an explicit construction of the geodesic γ_s. □

7.4.2 Risk-fairness trade-off on the population level

The next key result of our framework establishes the risk-fairness trade-off provided by the parameter $\alpha \in [0, 1]$ on the population level. In particular, it establishes a simple user-friendly relation between the risk and unfairness of α-relative improvement. Note that such a result is not available neither for $\mathcal{U}_{\text{TV}}$ nor for $\mathcal{U}_{\text{KS}}$, due to fundamentally different geometries of the squared risk and the aforementioned distances.

Lemma 7.4.5. *Let Assumption 7.2.1 be satisfied, then for any $\alpha \in [0,1]$ it holds that*

$$\mathcal{R}(f^*_\alpha) = (1-\sqrt{\alpha})^2 \mathcal{R}(f^*_0) = (1-\sqrt{\alpha})^2 \mathcal{U}(f^*) \ . \tag{7.7}$$

Proof. Proposition 7.4.1 gives the following explicit expression for the best α-improvement of f^*:

$$f^*_\alpha(\boldsymbol{x}, s) = \sqrt{\alpha} f^*(\boldsymbol{x}, s) + (1-\sqrt{\alpha}) f^*_0(\boldsymbol{x}, s) \ .$$

Plugging it in the risk gives

$$\mathcal{R}(f^*_\alpha) = \|f^*_\alpha - f^*\|_2^2 = (1-\sqrt{\alpha})^2 \|f^*_0 - f^*\|_2^2 = (1-\sqrt{\alpha})^2 \mathcal{R}(f^*_0) \ .$$

This proves the first equality. Given the definition of f^*_0, the second equality is exactly the result stated in Theorem 7.3.2. □

Recall that thanks to Theorem 7.3.2 we have $\mathcal{R}(f^*_0) = \mathcal{U}(f^*)$. Hence, the α-relative improvement f^*_α enjoys the following two properties

$$\mathcal{R}(f^*_\alpha) = (1-\sqrt{\alpha})^2 \mathcal{R}(f^*_0) \quad \text{and} \quad \mathcal{U}(f^*_\alpha) = \alpha \mathcal{U}(f^*) \ .$$

For instance, if $\alpha = 1/2$, that is, we want to half the unfairness of f^*, it incurs the risk which is equal to $\approx 8.5\%$ of the risk of exactly fair predictor f^*_0. We illustrate this general behaviour in Figure 7.1 (Section 7.2), where the risk and the unfairness of f^*_α are shown for different levels of $\mathcal{U}(f^*)$. A striking observation we can make from this plot is that, letting α vary between 0 and 1, the risk of f^*_α growth rapidly in the vicinity of zero, while it behaves almost linearly in a large neighbourhood of one. That is, one can find a prediction f whose unfairness $\mathcal{U}(f)$ is smaller than that of f^* by a constant multiplicative factor, without a large increase in risk.

Remark 7.4.6. *Let us remark that the results of this section apply in the case of* arbitrary *weights $\boldsymbol{w}$. This can be potentially useful for applications where the group-wise risks must be re-weighted. For instance, one can consider uniform weights $\boldsymbol{w} = (1/K, \ldots, 1/K)^\top$ or weights which are proportional to $1/\mathbb{P}(S = s)$.*

7.4.3 Pareto efficiency: a systematic way to select α

Even though the parameter $\alpha \in [0,1]$ has a clear interpretation in our framework, one still might have to figure out which α to pick in practice. The ultimate theoretical goal is to find a prediction f which simultaneously minimizes the risk $\mathcal{R}$ and the unfairness $\mathcal{U}$. Yet, unless f^* satisfies $\mathcal{U}(f^*) = 0$, this goal is unreachable and some trade-offs must be examined. A standard approach to study such *multi-criteria optimization problems* is via the notion of *Pareto dominance* and *Pareto efficiency* (Osborne and Rubinstein 1994). In words, the idea of Pareto analysis is to restrict the attention of a practitioner to some set of "good" predictors, termed *Pareto frontier* of the multi-criteria optimization problem, instead of considering all possible predictions. In this section, we show that the set of oracle α-RI $\{f^*_\alpha\}_{\alpha \in [0,1]}$ is the Pareto frontier of the multi-criteria minimization problem with target functions $f \mapsto \mathcal{R}(f)$ and $f \mapsto \mathcal{U}(f)$.

Let us first introduce the terminology of the Pareto analysis specified for our setup. We say that a prediction f *Pareto dominates* a prediction f' if one of the following holds

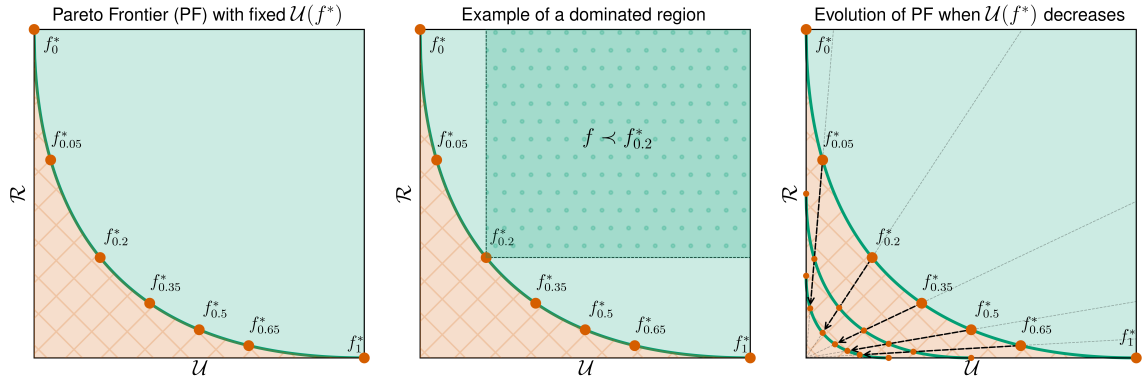


Figure 7.3: Illustration of Pareto frontiers and Pareto dominance. *Left*: Orange (hatched) part is not realisable by any prediction f; Each point of green (not hatched) part is realizable by some prediction f; The curve that separates the two is the Pareto frontier. *Center*: The darker green (dotted) rectangle in the upper right corner is the set of predictors dominated by $f^*_{0.2}$. *Right*: Evolution of the Pareto frontier when $\mathcal{U}(f^*)$ decreases.

- $\mathcal{R}(f) \leq \mathcal{R}(f')$ and $\mathcal{U}(f) < \mathcal{U}(f')$;
- $\mathcal{R}(f) < \mathcal{R}(f')$ and $\mathcal{U}(f) \leq \mathcal{U}(f')$.

To denote the fact that f' is dominated by f we write $f' \prec f$. Moreover, we say that f' and f are *comparable* if either $f' \prec f$ or $f \prec f'$. Intuitively, whenever $f' \prec f$, the prediction f is strictly preferable, since it is at least as good as f' for both criteria and it is strictly better for at least one of them.

Note that not every two predictions are actually comparable, that is, the relation $\prec$ only defines a partial-order. It is a known fact that partially ordered sets can be partitioned into well-ordered chains, that is, every pair within the chain is comparable and the restriction of $\prec$ on this chain defines an order relation. In this set-theoretic terminology, a prediction f is *Pareto efficient* if it is *maximal* within some chain in the sense of the partial order $\prec$. In other words, a prediction f is *Pareto efficient* if it is not dominated by any other prediction. The set of all Pareto efficient predictions is called the *Pareto frontier* and is denoted by PF.

Note that it would be more accurate to say that f' $\mathbb{P}$-Pareto dominates f and f' is $\mathbb{P}$-Pareto efficient, since the above definitions are acting on the level of population and they do depend on the underlying distribution. We omit this notation for simplicity.

In general, an analytic description of the Pareto frontier PF is not necessarily feasible. However, in our case, thanks to the analysis of the previous section, we can precisely describe the Pareto frontier of this problem.

Proposition 7.4.7. *Let Assumption 7.2.1 be satisfied. Then, the Pareto frontier for the multi-criteria minimization problem with objective functions $\mathcal{R}(f)$ and $\mathcal{U}(f)$ is given by $\{f^*_\alpha\}_{\alpha\in[0,1]}$.*

Proof. On the one hand, by definition of f^*_α it holds that $\{f^*_\alpha\}_{\alpha\in[0,1]} \subset \mathrm{PF}$. On the other hand, let $f \in \mathrm{PF}$ with $\mathcal{U}(f) \neq 0$ and let $\alpha_f := \mathcal{U}(f)/\mathcal{U}(f^*)$. Then by the definition of α_f it holds that $\mathcal{U}(f) = \alpha_f \mathcal{U}(f^*)$. Furthermore, by definition of $f^*_{\alpha_f}$ it holds that $\mathcal{R}(f^*_{\alpha_f}) \leq \mathcal{R}(f)$ and

by Lemma 7.4.5 it holds that $\mathcal{U}(f^*_{\alpha_f}) = \alpha_f \mathcal{U}(f^*)$. Finally, since f is Pareto efficient it holds that $\mathcal{R}(f) \leq \mathcal{R}(f^*_{\alpha_f})$. If $f \in \mathrm{PF}$ is such that $\mathcal{U}(f) = 0$, then it is as good as f^*_0 in the sense of Pareto. The proof is concluded[4]. □

Note that any predictor f defines a point $(\mathcal{U}(f), \mathcal{R}(f))$ in the coordinate system $(\mathcal{U}, \mathcal{R})$. The left plot of Figure 7.3 illustrates the Pareto frontier and those values of $(\mathcal{U}, \mathcal{R})$ that are attainable by some prediction f. We remark that the convexity of the Pareto frontiers curve is due to the specific trade-off provided by the parameter α. For a general multi-criteria optimization problem this convexity is not ensured. The right plot of Figure 7.3 demonstrates the evolution of the Pareto frontier when $\mathcal{U}(f^*)$ decreases. A general conclusion of this plot is that if $\mathcal{U}(f^*)$ is low, then one can set $\alpha = 1$.

Finally, Proposition 7.4.7 provides simple practical guidelines for the study of the trade-off given by α. Note that since thanks to Lemma 7.4.5 it holds that $\mathcal{R}(f^*_0) = \mathcal{U}(f^*)$ and $\mathcal{U}(f^*_1) = \mathcal{U}(f^*)$, then the practitioner needs to estimate only one quantity $\mathcal{U}(f^*)$ and trace the curve of Pareto frontier in order to establish the desired trade-off for the problem at hand.

7.5 Minimax setup

While the previous section was dealing with the general framework on the population level, the goal of this section is to put forward a minimax setup for the statistical problem of regression with the introduced fairness constraints.

Let $(\boldsymbol{X}_1, S_1, Y_1), \ldots, (\boldsymbol{X}_n, S_n, Y_n)$ be i.i.d. sample with joint distribution $\mathbf{P}_{(f^*, \boldsymbol{\theta})}$, where the pair $(f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta$ for some class $\mathcal{F}$ and Θ. In this notation f^* is the regression function and $\boldsymbol{\theta}$ is a nuisance parameter. For example $\mathcal{F}$ can be the set of all affine or Lipschitz continuous functions and Θ defines additional assumptions on the model in Eq. (7.1) (see Section 7.6 for a concrete example). For a given fairness parameter $\alpha \in [0, 1]$ and a given confidence parameter $t > 0$, the goal of the statistician is to construct an estimator[5] $\widehat{f}$, which simultaneously satisfies the following two properties

1. Uniform fairness guarantee:

$$\forall (f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta \quad \mathbf{P}_{(f^*, \boldsymbol{\theta})} \left(\mathcal{U}(\widehat{f}) \leq \alpha \mathcal{U}(f^*) \right) \geq 1 - t \ , \tag{7.8}$$

2. Uniform risk guarantee:

$$\forall (f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta \quad \mathbf{P}_{(f^*, \boldsymbol{\theta})} \left(\mathcal{R}(\widehat{f}) \leq r_{n,\alpha,f^*}(\mathcal{F}, \Theta, t) \right) \geq 1 - t \ . \tag{7.9}$$

Eq. (7.8) states that the constructed estimator satisfies the fairness requirement with high probability uniformly over the class $\mathcal{F} \times \Theta$. Meanwhile, in Eq. (7.9) we seek for the smallest rate $r_{n,\alpha,f^*}(\mathcal{F}, \Theta, t)$ to quantify the statistical price of being α-relatively fair. Note that $r_{n,\alpha,f^*}(\mathcal{F}, \Theta, t)$ depends explicitly on f^*. This is explained by the fact that the fairness of

[4] To be more precise, one needs to introduce the equivalence relation $\sim$ defined as $f \sim f'$ iff $\mathcal{R}(f) = \mathcal{R}(f')$ and $\mathcal{U}(f) = \mathcal{U}(f')$ and to perform the exact same proof on the quotient space. For the sake of presentation we omit this benign technicality.

[5] As usual, an estimator $\widehat{f}$ is a measurable mapping of data to the space of predictions.

$\widehat{f}$ is measured relatively to f^*, hence the price of this constraint also depends on the initial unfairness level of the regression function f^*.

The actual construction of the estimator $\widehat{f}$ is problem dependent and the proving that it satisfies Eqs. (7.8)–(7.9) requires a careful case-by-case study. In Section 7.6 we provide an example of such analysis for a simple statistical model of linear regression with systematic group-dependent bias.

7.5.1 Generic lower bound

While the upper bounds of Eqs. (7.8)–(7.9) require a problem dependent analysis, a general problem dependent lower bound can be derived. In this section we develop such lower bound. Let us first introduce some useful definitions.

Assumption 7.5.1 (Unconstrained rate)**.** *For a fixed confidence level $t \in (0,1)$ and a class $(\mathcal{F}, \Theta)$, there exists a positive sequence $\delta_n(\mathcal{F}, \Theta, t)$ such that*

$$\inf_{\widehat{f}} \sup_{(f^*,\boldsymbol{\theta})\in\mathcal{F}\times\Theta} \mathbf{P}_{(f^*,\boldsymbol{\theta})}\left(\mathcal{R}(\widehat{f}) \geq \delta_n(\mathcal{F}, \Theta, t)\right) \geq t \ ,$$

where the infimum is taken over all estimators.

Assumption 7.5.1 can be used with any sequence $\delta_n(\mathcal{F}, \Theta, t)$, however, we implicitly assume that $\delta_n(\mathcal{F}, \Theta, t)$ corresponds to the minimax optimal rate of estimation of f^* by any estimator (without constraints) in expected squared loss.

Definition 7.5.2 (Valid estimators)**.** *For some $\alpha \in [0,1]$ and confidence level $t' \in (0,1)$ we say that an estimator $\widehat{f}$ is (α, t')-valid* w.r.t. *the class $(\mathcal{F}, \Theta)$ if*

$$\inf_{(f^*,\boldsymbol{\theta})\in\mathcal{F}\times\Theta} \mathbf{P}_{(f^*,\boldsymbol{\theta})}\left(\mathcal{U}(\widehat{f}) \leq \alpha\mathcal{U}(f^*)\right) \geq 1 - t' \ .$$

The set of all (α, t')-valid estimators w.r.t. *the class $(\mathcal{F}, \Theta)$ is denoted by $\widehat{\mathcal{F}}_{(\alpha,t')}$.*

Definition 7.5.2 characterizes estimators which satisfy the α-RI constraint at least with constant probability uniformly over the class $(\mathcal{F}, \Theta)$.

Equipped with Assumption 7.5.1 and Definition 7.5.2 we are in position to state the main result of this section, which establishes the statistical risk-fairness trade-off. As we will see in Section 7.6, supported by appropriate upper bounds, Theorem 7.5.3 yields optimal rates of convergence up to a multiplicative factor.

Theorem 7.5.3. *Let Assumption 7.2.1 be satisfied. Let $\delta_n(\mathcal{F}, \Theta, t)$ be a sequence that satisfies Assumption 7.5.1. Then*

$$\inf_{\widehat{f}\in\widehat{\mathcal{F}}_{(\alpha,t')}} \sup_{(f^*,\boldsymbol{\theta})\in\mathcal{F}\times\Theta} \mathbf{P}_{(f^*,\boldsymbol{\theta})}\left(\mathcal{R}^{1/2}(\widehat{f}) \geq \delta_n^{1/2}(\mathcal{F}, \Theta, t) \vee (1-\sqrt{\alpha})\mathcal{U}^{1/2}(f^*)\right) \geq t \wedge (1 - t') \ .$$

Drawing an analogy with Lemma 7.4.5, the two terms of the derived bound have natural interpretations: the first term $\delta_n(\mathcal{F}, \Theta, t)$ is the price of statistical estimation; the second term $(1-\sqrt{\alpha})^2\mathcal{U}(f^*)$ is the price of fairness. Consequently, the rate $r_{n,\alpha,f^*}(\mathcal{F}, \Theta, t)$ in Eq. (7.9) is

lower bounded (up to a multiplicative constant factor) by $\delta_n^{1/2}(\mathcal{F}, \Theta, t) \vee (1-\sqrt{\alpha})\mathcal{U}^{1/2}(f^*)$. The confidence parameter on the *r.h.s.* of the bound is $t \wedge (1-t')$. The reasonable choice of t' is in the vicinity of zero, which corresponds to estimators satisfying the fairness constraint with high probability. Finally, observe that this bound in not conventional in the sense of classical statistics, where the bound would converge to zero with the growth of sample size. This behavior is not surprising, since the infimum is taken *w.r.t.* to (α, t')-valid estimators and not *w.r.t.* all possible estimators. One can draw an analogy of the obtained bound with recent results in robust statistics (Chen, Gao, and Ren 2016; Chen, Gao, and Ren 2018), where the minimax rate converges to a function of the proportion of outliers, which might be different from zero.

7.6 Application to linear model with systematic bias

Additional notation We denote by $\|\cdot\|_2$ and by $\|\cdot\|_n = (1/\sqrt{n})\|\cdot\|_2$ the Euclidean and the normalized Euclidean norm. The standard scalar product is denoted by $\langle\cdot,\cdot\rangle$. We denote by $\mathbf{1}_p$ the vector of all ones of size p. For square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}, n \geq 1$, we write $\mathbf{A} \succ 0$ if $\mathbf{A}$ is symmetric positive-definite.

The goal of this part is to provide an example of a complete statistical analysis for a regression problem under the α-RI constraint. In particular, we show how to apply the plug-and-play results of Section 7.5.1 in order to derive minimax rate optimal bounds under the α-RI constraint. To this end we apply the developed theory to the following model of linear regression with systematic group-dependent bias

$$Y = \langle \boldsymbol{X}, \boldsymbol{\beta}^* \rangle + b_S^* + \xi \ , \tag{7.10}$$

where $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is a feature vector independent from the sensitive attribute S with $\boldsymbol{\Sigma} \succ 0$; $\xi \sim \mathcal{N}(0, \sigma^2)$ is an additive independent noise; and the vector $\boldsymbol{b}^* = (b_1^*, \ldots, b_K^*)$ is the vector of systematic bias. We assume that the noise level σ is known to the statistician. Note that in this case the regression function f^* is given by the expression $f^*(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta}^* \rangle + b_s^*$ and Assumption 7.2.1 is satisfied. We assume that the observations are

$$\boldsymbol{Y}_s = \mathbf{X}_s \boldsymbol{\beta}^* + b_s^* \mathbf{1}_{n_s} + \boldsymbol{\xi}_s, \quad s = 1, \ldots, K \ , \tag{7.11}$$

with $\boldsymbol{Y}_s, \boldsymbol{\xi}_s \in \mathbb{R}^{n_s}$, $\mathbf{X}_s \in \mathbb{R}^{n_s \times p}$, and $\mathbf{1}_{n_s}$ is the vector of all ones of size n_s. The rows of $\mathbf{X}_s$ are i.i.d. realization of $\boldsymbol{X}$, the components of $\boldsymbol{\xi}_s$ are i.i.d. from $\mathcal{N}(0, \sigma^2)$. Additionally, we set $n = n_1 + \ldots + n_K$ and $w_s = {}^{n_s}/_{n}$. The risk of a prediction rule $f : \mathbb{R}^p \times [K] \to \mathbb{R}$ is defined as

$$\mathcal{R}(f) = \sum_{s=1}^{K} w_s \mathbb{E}\left(\langle \boldsymbol{X}, \boldsymbol{\beta}^* \rangle + b_s^* - f(\boldsymbol{X}, s)\right)^2 \ .$$

Remark 7.6.1. *We set $w_s = {}^{n_s}/_{n}$ instead of $w_s = \mathbb{P}(S = s)$ to simplify the presentation and proofs of the main results. Thanks to Remark 7.4.6, all of the statements of Sections 7.4-7.5 are applied for this choice. Finally, note that if $\mathbb{P}(S = s) = w'_s$ and $S_1, \ldots, S_n$ is an* i.i.d. *sample, then $n_s = \sum_{i=1}^n \mathbb{I}\{S_i = s\}$ and $\mathbf{E}[{}^{n_s}/_{n}] = w'_s$, that is our choice of weights essentially corresponds to the scenario of* i.i.d. *sampling of sensitive attribute.*

Using the terminology of Section 7.5.1 the joint distribution of data sample $\mathbf{P}_{(f^*,\boldsymbol{\theta})}$ is uniquely defined by $(\boldsymbol{\beta}^*, \boldsymbol{b}^*)$ and $(\boldsymbol{\Sigma}, \sigma)$. That is, $(\boldsymbol{\beta}^*, \boldsymbol{b}^*)$ defines the regression function f^* and $(\boldsymbol{\Sigma}, \sigma)$ is the nuisance parameter $\boldsymbol{\theta}$. To simplify the notation we write $\mathbf{P}_{(\beta^*,\boldsymbol{b}^*)}$ instead of $\mathbf{P}_{(\beta^*,\boldsymbol{b}^*,\boldsymbol{\Sigma},\sigma)}$.

The following result is the application of Proposition 7.4.1 to the model in Eq. (7.10).

Proposition 7.6.2. *For all $\alpha \in [0,1]$, the α-relative improvement of f^* is given for all $(\boldsymbol{x}, s) \in \mathbb{R}^p \times [K]$ by*

$$f^*_\alpha(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta}^* \rangle + \sqrt{\alpha} b^*_s + (1-\sqrt{\alpha}) \sum_{s=1}^K w_s b^*_s \ .$$

In order to build an estimator $\widehat{f}$, which improves the fairness of f^*, while providing minimal risk among such predictions, we first estimate parameters of model in Eq. (7.10) using least-squares estimators

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}) \in \underset{(\boldsymbol{\beta},\boldsymbol{b})\in\mathbb{R}^p\times\mathbb{R}^K}{\arg\min} \sum_{s=1}^K w_s \left\| \boldsymbol{Y}_s - \mathbf{X}_s\boldsymbol{\beta} - b_s \mathbf{1}_{n_s} \right\|^2_{n_s} \ . \tag{7.12}$$

Based on the above quantities we then define a family of linear estimators $\widehat{f}_\tau$ parametrized by $\tau \in [0,1]$ as

$$\widehat{f}_\tau(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \widehat{\boldsymbol{\beta}} \rangle + \sqrt{\tau}\widehat{b}_s + (1-\sqrt{\tau}) \sum_{s=1}^K w_s \widehat{b}_s, \qquad (\boldsymbol{x}, s) \in \mathbb{R}^p \times [K] \ . \tag{7.13}$$

We would like to find a value of $\tau = \tau_n(\alpha)$ such that Eqs. (7.8)–(7.9) are satisfied. Note that the choice of $\tau = \alpha$ would not yield the desired fairness guarantee stated in Eq. (7.8). As it will be shown later, τ should be smaller than α, in order to account for finite sample effects and derive high confidence fairness guarantee. The next result shows that under the model in Eq. (7.10), the unfairness of $\widehat{f}_\tau$ can be computed in a data-driven manner, which is crucial for the consequent choice of τ.

Lemma 7.6.3. *For any $\tau \in [0,1]$, the unfairness of $\widehat{f}_\tau$ is given by*

$$\mathcal{U}(\widehat{f}_\tau) = \tau \sum_{s=1}^K w_s \left(\widehat{b}_s - \sum_{s'=1}^K w_{s'} \widehat{b}_{s'} \right)^2 \ ,$$

almost surely.

Apart from being computable in practice, Lemma 7.6.3 provides an intuitive result that $\mathcal{U}(\widehat{f}_\tau)$ is the variance of the bias term $\widehat{\boldsymbol{b}}$.

7.6.1 Upper bound

Linear regression is one of the most well-studied problems of statistics (Nemirovski 2000; Tsybakov 2003; Györfi et al. 2006; Mourtada 2019; Catoni 2004; Hsu, Kakade, and Zhang 2012; Audibert and Catoni 2011). In the context of fairness, linear regression is considered in (Calders et al. 2013; Berk et al. 2017; Donini et al. 2018), where the fairness constraint is

formulated via the approximate equality of group-wise means. In this section we establish a statistical guarantee on the risk and fairness of $\widehat{f}_\tau$ for an appropriate data-driven choice of τ. Our theoretical analysis in this part is inspired by that of Hsu, Kakade, and Zhang 2012, who derived high probability bounds on least squares estimator for linear regression with random design.

The following rate plays a crucial rule in the analysis of this section

$$\delta_n(p,K,t) = 8\left(\frac{p}{n}+\frac{K}{n}\right) + 16\left(\sqrt{\frac{p}{n}}+\sqrt{\frac{K}{n}}\right)\sqrt{\frac{t}{n}}+\frac{32t}{n} \ .$$

Not taking into account the confidence parameter $t > 0$, $\delta_n(p,K,t) \asymp (p+K)/n$ up to a constant multiplicative factor, which as it is shown in Theorem 7.6.5 is the minimax optimal rate for the model in Eq. (7.10) without the fairness constraint.

Theorem 7.6.4 (Fairness and risk upper bound)**.** *Define*

$$\widehat{\tau} = \begin{cases} \alpha\left(1+\frac{\sigma\delta_n^{1/2}(p,K,t)}{\mathcal{U}^{1/2}(\widehat{f}_1)-\sigma\delta_n^{1/2}(p,K,t)}\right)^{-2} & \text{if} \quad \mathcal{U}^{1/2}(\widehat{f}_1) > \sigma\delta_n^{1/2}(p,K,t) \\ 0, & \text{otherwise} \end{cases} \ .$$

Consider $p, K \in \mathbb{N}, t \geq 0$ *and define* $\gamma(p,K,t) = (4\sqrt{K}+5\sqrt{t}+6\sqrt{p})/(\sqrt{p}+\sqrt{t})$. *Assume that* $\sqrt{n} \geq 2(\sqrt{p}+\sqrt{t})/(\gamma(p,K,t)-\sqrt{\gamma^2(p,K,t)-3})$. *Then, for any* $\alpha \in [0,1]$, *with probability at least* $1-4\exp(-t/2)$ *it holds that*

$$\mathcal{U}(\widehat{f}_{\widehat{\tau}}) \leq \alpha\mathcal{U}(f^*) \quad \text{and} \quad \mathcal{R}^{1/2}(\widehat{f}_{\widehat{\tau}}) \leq 2\sigma(1+\sqrt{\alpha})\delta_n^{1/2}(p,K,t) + (1-\sqrt{\alpha})\mathcal{U}^{1/2}(f^*) \ .$$

Theorem 7.6.4 simultaneously provides two results: first, it shows that the estimator $\widehat{f}_{\widehat{\tau}}$ is $(\alpha, 4e^{-t/2})$-valid, that is, it satisfies the fairness constraint with high probability; second it provides the rate of convergence which consists of two parts. The first part of the rate, $\sigma\delta_n^{1/2}(p,K,t)$, is the price of statistical estimation of $(\boldsymbol{\beta}^*, \boldsymbol{b}^*)$, while the second part, $(1-\sqrt{\alpha})\mathcal{U}^{1/2}(f^*)$, is the price one has to pay when introducing the α-RI fairness constraint. In order to achieve the fairness validity, we need to loosen the value of α to reflect the base level of unfairness, that is, $\widehat{\tau}$ is adjusted by $\mathcal{U}(\widehat{f}_1)$. Let us point out that the bound of Theorem 7.6.4 slightly differs from the conditions required by Eqs. (7.8)–(7.9). In particular, it provides a joint guarantee on risk and fairness.

Let us remark that the previous result requires n to be sufficiently large, similarly to the conditions in (Hsu, Kakade, and Zhang 2012; Audibert and Catoni 2011). One can obtain a more explicit, but more restrictive bound on n by finding sufficient conditions under which the assumption on n is satisfied. For instance, rough computations show that it is sufficient to assume that $\sqrt{n} \geq 16\sqrt{K}$ and $\sqrt{n} \geq 12.5(\sqrt{p}+\sqrt{t})$.

At last, we emphasize that the choice of $\widehat{\tau}$ requires the knowledge of the noise level σ, that is, this choice is not adaptive. However, our proof can effortlessly be extended to the case when only an upper bound $\bar{\sigma}$ on the noise level σ is known. In this case σ should be replaced by $\bar{\sigma}$ in the definition of $\widehat{\tau}$ and in the resulting rate. The question of adaptation to σ without any prior knowledge should be treated separately and is out of the scope of this work.

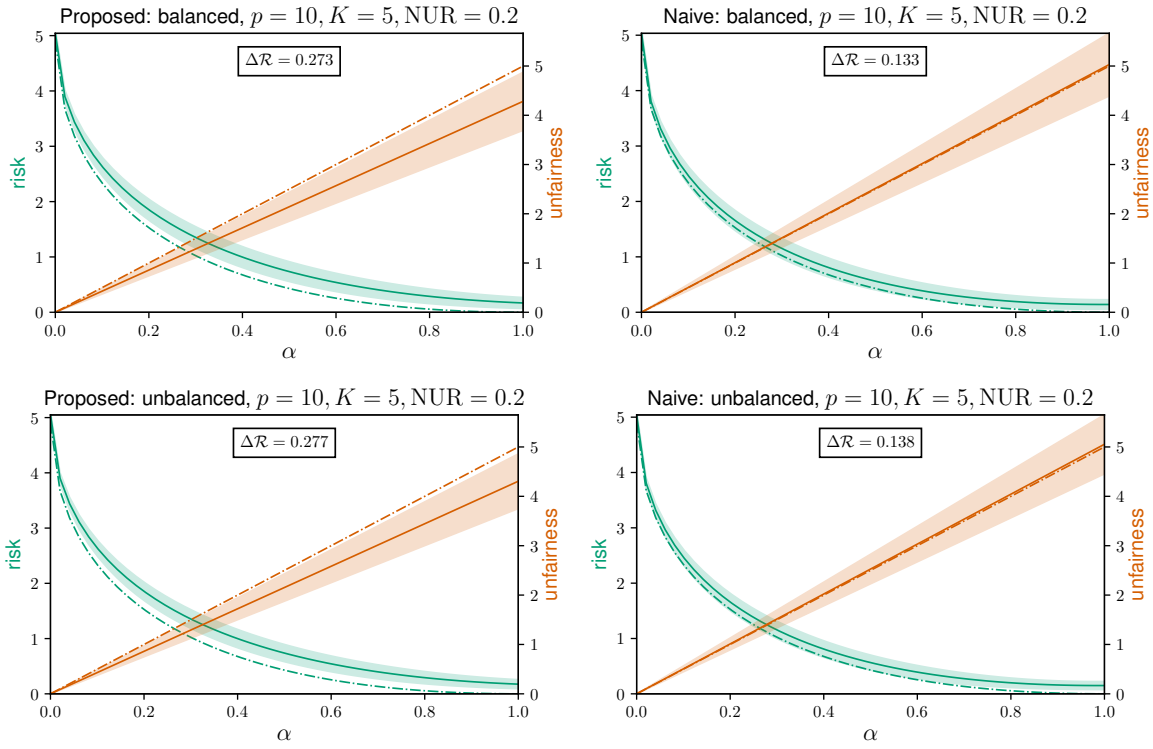


Figure 7.4: Dashed green and brown lines correspond to the risk and unfairness of f^*_α respectively. Solid green and brown lines correspond to the average risk and unfairness of $\widehat{f}_{\tau(\alpha)}$ and the shaded region shows three standard deviations over 50 repetitions. On the left $\tau(\alpha) = \widehat{\tau}$ and on the risk $\tau(\alpha) = \alpha$.

7.6.2 Lower bound

The goal of this section is to provide a lower bound, demonstrating that the result of Theorem 7.6.4 is minimax optimal up to a multiplicative constant factor. Recall that thanks to the general lower bound derived in Theorem 7.5.3 it is sufficient to prove a lower bound on the risk without constraining the set of possible estimators. Even though the problem of linear regression is well studied, to the best of our knowledge there is no known lower bound for the model in Eq. (7.10) which *i)* holds for the random design *ii)* is stated in probability *iii)* considers explicitly the confidence parameter t. Next theorem establishes such lower bound.

Theorem 7.6.5. *For all $n, p, K \in \mathbb{N}$, $t \geq 0$, $\sigma > 0$ it holds that*

$$\inf_{\widehat{f}} \sup_{(\boldsymbol{\beta}^*, \boldsymbol{b}^*) \in \mathbb{R}^p \times \mathbb{R}^K, \boldsymbol{\Sigma} \succ 0} \mathbf{P}_{(\boldsymbol{\beta}^*, \boldsymbol{b}^*)} \left(\mathcal{R}(\widehat{f}) \geq \frac{\sigma^2}{3 \cdot 2^9 n} (\sqrt{p+K} + \sqrt{32t})^2 \right) \geq \frac{1}{12} e^{-t} \ ,$$

where the infimum is taken w.r.t. *all estimators.*

The proof of Theorem 7.6.5 relies on standard information theoretic results. In particular, in order to prove optimal exponential concentration we follow similar strategy as that of Bellec 2017; Kerkyacharian et al. 2014 who derived optimal exponential concentrations in the context

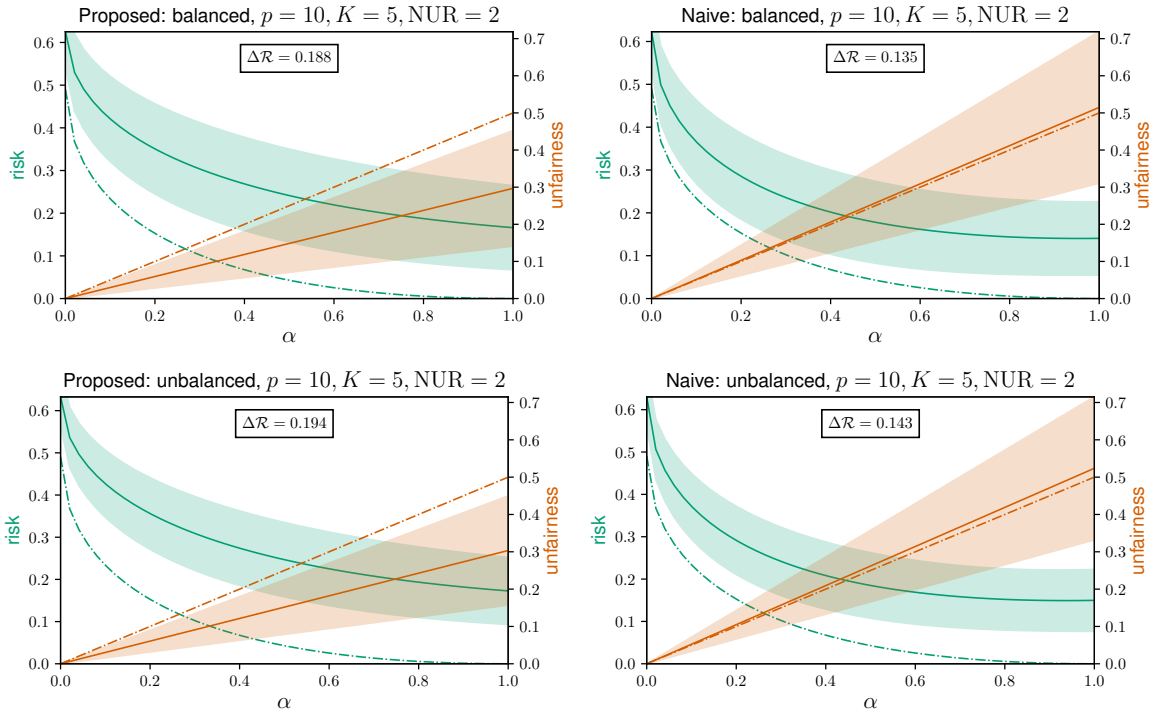


Figure 7.5: Dashed green and brown lines correspond to the risk and unfairness of f^*_α respectively. Solid green and brown lines correspond to the average risk and unfairness of $\widehat{f}_{\tau(\alpha)}$ and the shaded region shows three standard deviations over 50 repetitions. On the left $\tau(\alpha) = \widehat{\tau}$ while on the right $\tau(\alpha) = \alpha$.

of density aggregation and binary classification. Theorem 7.6.5 combined with generic lower bound derived in Theorem 7.5.3 yields the following corollary.

Corollary 7.6.6. *Let* $\bar{\delta}_n(p, K, t) = (\sqrt{(p+K)/n} + \sqrt{32t/n})^2/(3 \cdot 2^9)$. *For all* $n, p, K \in \mathbb{N}$, $t \geq 0$, $\sigma > 0$, $\alpha \in [0, 1]$ *it holds for all* $t \geq 0$ *and all* $t' \leq 1 - e^{-t}/12$ *that*

$$\inf_{\widehat{f} \in \widehat{\mathcal{F}}_{\alpha,t'}} \sup_{(\boldsymbol{\beta}^*, \boldsymbol{b}^*) \in \mathbb{R}^p \times \mathbb{R}^K, \boldsymbol{\Sigma} \succ 0} \mathbf{P}_{(\boldsymbol{\beta}^*, \boldsymbol{b}^*)} \left(\mathcal{R}^{1/2}(\widehat{f}) \geq \sigma \bar{\delta}_n^{1/2}(p, K, t) \vee (1 - \sqrt{\alpha}) \mathcal{U}^{1/2}(f^*) \right) \geq \frac{1}{12} e^{-t} \ .$$

Comparing the upper bound of Theorem 7.6.4 and the lower bound of Corollary 7.6.6 we conclude that the two obtained rates are the same up to a multiplicative constant factor. Hence confirming the tightness of the results derived in Section 7.5.1.

7.6.3 Simulation study

In this section we perform simulation study to empirically validate our theoretical analysis[6]. Before continuing let us discuss the notion of signal-to-unfairness ratio. Setting $\boldsymbol{\beta}^* = 0$ in

[6]For our empirical validation and illustrations we have relied on the following `python` packages: `scikit-learn` (Pedregosa et al. 2011a), `numpy` (Van Der Walt, Colbert, and Varoquaux 2011), `matplotlib` (Hunter 2007), `seaborn`.

the model (7.10), if the amplitudes of b_s^* is much smaller than the noise level σ^2, then the observations $\boldsymbol{Y}_s$ are mainly composed of noise. While for the prediction problem it is not a problem, since our rates will scale with the noise level, it becomes important for the estimation of unfairness $\mathcal{U}(f^*)$. Motivated by this discussion, we define the noise-to-unfairness ratio as

$$\mathrm{NUR}^2 := \frac{\sigma^2}{\mathcal{U}(f^*)} \ .$$

The signal-to-unfairness ratio tells as how the level of unfairness compares to the noise level. The regime $\mathrm{NUR} \gg 1$ means that the unfairness of the distributions is below the noise level, and it is statistically difficult to estimate it. In contrast, $\mathrm{NUR} \ll 1$ implies that the unfairness dominates the noise. Instead of varying $\mathcal{U}(f^*)$ and σ we fix σ and perform our study for different values of NUR.

We follow the following protocol. For some fixed $K, n_1, \ldots, n_K, p, \sigma, \mathrm{NUR}$ we simulate the model in Eq. (7.11) with $\boldsymbol{\Sigma} = \mathbf{I}_p$. In all the experiments we set $\boldsymbol{\beta}^* = (1, \ldots, 1)^\top \in \mathbb{R}^p$. For $\boldsymbol{b}^*$ we first define $\boldsymbol{v} = (1, -1, 1, -1, \ldots)^\top \in \mathbb{R}^K$ and set $\boldsymbol{b}^* = \boldsymbol{v}\sqrt{\sigma^2/\mathrm{NUR} \cdot \mathsf{Var}_S(\boldsymbol{v})}$, where $\mathsf{Var}_S(\boldsymbol{v})$ is the variance of $\boldsymbol{v}$ with weights $w_1, \ldots, w_K$. So that the unfairness of this model is exactly equal to σ^2/NUR^2. On each simulation round of the model, we compute the estimator in Eq. (7.13) with two choices of parameter τ:

1. **Proposed**: $\tau(\alpha) = \widehat{\tau}$ from Theorem 7.6.4;
2. **Naive**: $\tau(\alpha) = \alpha$.

Remark 7.6.7. *While performing experiments we have noticed that setting $\widehat{\tau}$ with $\delta_n(p, K, t)$ defined in Theorem 7.6.4 results in too pessimistic estimates in terms of unfairness, for this reason in all of our experiments we set $\delta_n(p, K, t) = (p/n) + (K/n)$, which is of the same order as that of Theorem 7.6.4.*

Then, for each $\widehat{f}_{\tau(\alpha)}$ we evaluate $\mathcal{R}(\widehat{f}_{\tau(\alpha)})$ and $\mathcal{U}(\widehat{f}_{\tau(\alpha)})$. This procedure is repeated 50 times, which results in 50 values of $\mathcal{R}(\widehat{f}_{\tau(\alpha)})$ and $\mathcal{U}(\widehat{f}_{\tau(\alpha)})$ for each $\alpha \in (0, 1)$. For these 50 values we compute mean and standard deviation. We considered $p = 10$, $K = 5$, $\sigma = 1$, and $\mathrm{NUR} \in \{0.2, 0.5, 2\}$. Furthermore, for the choice of $n_1, \ldots, n_K$ we study the following two regimes

1. **Balanced**: $n_1 = \ldots = n_5 = 100$.
2. **Unbalanced**: $n_1 = 5, n_2 = 45, n_3 = 100, n_4 = 100, n_5 = 250$.

The reason we consider two regimes is to confirm the theoretical findings of Theorem 7.6.4, which indicate that the rate is governed by $n_1 + \ldots + n_K$ instead of the their individual values. Finally, for a given fairness parameter function $\alpha \mapsto \tau(\alpha)$ we report cumulative risk increase over all $\alpha \in [0, 1]$ defined as

$$\Delta\mathcal{R}(\tau) := \int_0^1 \left(\mathcal{R}(\widehat{f}_{\tau(\alpha)}) - \mathcal{R}(f_\alpha^*)\right) d\alpha \ .$$

This quantity describes the cumulative risk loss of the rule $\tau(\alpha)$ across all the levels of fairness α compared to the best α-relative improvement f_α^*.

On Figures 7.4–7.5 we draw the evolution of the risk and of the unfairness when α traverses the interval $[0, 1]$. We also report $\Delta\mathcal{R}(\tau)$ defined above. Inspecting the plots we can see that

BALANCED

	Oracle		Proposed		Naive	
α	$\mathcal{R}(f^*_\alpha)$	$\mathcal{U}(f^*_\alpha)$	$\mathcal{R}(\widehat{f}_{\widehat{\tau}})$	$\mathcal{U}(\widehat{f}_{\widehat{\tau}})$	$\mathcal{R}(\widehat{f}_\alpha)$	$\mathcal{U}(\widehat{f}_\alpha)$
0	2.0	0.0	2.13 ± 0.03	0.0 ± 0.0	2.13 ± 0.03	0.0 ± 0.0
0.2	0.61	0.4	0.87 ± 0.05	0.31 ± 0.02	0.74 ± 0.04	0.40 ± 0.03
0.4	0.27	0.8	0.52 ± 0.05	0.62 ± 0.05	0.40 ± 0.04	0.81 ± 0.05
0.6	0.10	1.2	0.34 ± 0.04	0.93 ± 0.07	0.24 ± 0.04	1.21 ± 0.08
0.8	0.02	1.6	0.23 ± 0.04	1.25 ± 0.09	0.16 ± 0.03	1.61 ± 0.11
1	0.0	2.0	0.17 ± 0.03	1.56 ± 0.12	0.14 ± 0.03	2.02 ± 0.14

Table 7.1: Summary for $p = 10, K = 5, \text{NUR} = 0.5$. We report the mean and the standard deviation.

that the main disadvantage of the naive choice of $\tau = \alpha$ is its poor fairness guarantee, that is, in almost half of the outcomes, the unfairness of $\widehat{f}_\alpha$ exceeded the prescribed value. In contrast, the proposed choice of $\tau(\alpha) = \widehat{\tau}$ consistently improves the unfairness of the regression function f^*, empirically validating our findings in Theorem 7.6.4. However, good fairness results come at the cost of consistently higher risk. One can also see that the effect of unbalanced distributions is negligible for the considered model (it only affects the variance of the result). This is explained by the definition of the risk, which weights the groups proportionally to their frequencies. Finally, observing the behavior of naive approach for $\text{NUR} = 0.2$ and $\text{NUR} = 2$ we note that in the latter case the unfairness of $\widehat{f}_\alpha$ starts to deviate from the true value (with consistently positive bias). Meanwhile, since the proposed choice $\tau(\alpha) = \widehat{\tau}$ is more conservative, the bias remains negative, that is, the unfairness of f^* is still improved.

Table 7.1 presents the numeric results for $p = 10$, $K = 5$, $\text{NUR} = 0.5$. We remark the striking drop in the risk for $\alpha = 0.2$, indicating that a slight relaxation of the Demographic Parity constraint results in a significant improvement in terms of the risk. Of course, the justification of such a relaxation must be considered based on the application at hand.

7.7 Conclusion

In this work we proposed a theoretical framework for rigorous analysis of regression problems under fairness requirements. Our framework allows to interpolate between the regression of demographic parity and the unconstrained regression using univariate parameter between zero and one. Within this framework we precisely quantified the risk-fairness trade-off and derived general plug-n-play lower bound. To demonstrate the generality of our results we provided minimax analysis of the linear model with systematic group-dependent bias. Finally, we have performed empirical validation. For future work it would be interesting to extend our analysis to other statistical model, providing estimators with high confidence fairness improvement.

7.8 Reminder

7.8.1 The Wasserstein-2 distance

Additional notation For any $s \in [K]$, we denote by $\mu_{\boldsymbol{X}|s}$ the conditional distribution of the feature vector $\boldsymbol{X}$ knowing the attribute s. For a probability measure μ on $\mathbb{R}^p$ and a measurable function $g : \mathbb{R}^p \to \mathbb{R}$, we denote by $g\#\mu$ the push-forward (image) measure. That is for all measurable set $\mathcal{C} \subset \mathbb{R}$ it holds that $(g\#\mu)(\mathcal{C}) := \mu\{\boldsymbol{x} \in \mathbb{R}^p \,:\, g(\boldsymbol{x}) \in \mathcal{C}\}$.

We recall basic results on the Wasserstein-2 distance on the real line. We recall that the *Wasserstein-2* distance between probability distributions μ and ν in $\mathcal{P}_2(\mathbb{R}^d)$, the space of measures on $\mathbb{R}^d$ with finite second moment, is defined as

$$W_2^2(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 d\gamma(\boldsymbol{x},\boldsymbol{y}) \right\} \quad , \tag{7.14}$$

where $\Gamma(\mu,\nu)$ denotes the collection of measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. See Santambrogio (2015) and Villani (2003) for more details about Wasserstein distances and optimal transport.

The following lemma gives a closed form expression for the Wasserstein-2 distance between two univariate Gaussian distributions.

Lemma 7.8.1 (Fréchet 1957)**.** *For any $m_0, m_1 \in \mathbb{R}, \sigma_0, \sigma_1 \geq 0$ it holds that*

$$W_2^2\left(\mathcal{N}(m_0,\sigma_0^2), \mathcal{N}(m_1,\sigma_1^2)\right) = (m_0-m_1)^2 + (\sigma_0-\sigma_1)^2 \quad .$$

The next lemma gives a closed form expression for the barycenter of K univariate Gaussian distributions. It shows in particular that such barycenter is also a univariate Gaussian distribution.

Lemma 7.8.2 (Agueh and Carlier 2011)**.** *Let $\boldsymbol{w} \in \mathbb{R}^K$ be a probability vector, then the solution of*

$$\min_{\nu\in\mathcal{P}_2(\mathbb{R})} \sum_{s=1}^K w_s W_2^2\left(\mathcal{N}(m_s,\sigma_s^2),\nu\right) \quad ,$$

is given by $\mathcal{N}(\bar{m},\bar{\sigma}^2)$ with

$$\bar{m} = \sum_{s=1}^K w_s m_s \quad \text{and} \quad \bar{\sigma} = \sum_{s=1}^K w_s \sigma_s \quad .$$

Finally we state a lemma giving an explicit form for the transport map to the barycenter of probability distributions supported on the real line and the corresponding constant speed geodesics. See Agueh and Carlier 2011, Section 6.1.

Lemma 7.8.3. *Let $a_1, \ldots, a_K$ be non-atomic probability measures on the real line that have finite second moments, and let $w_1, \ldots, w_K$ be positive reals that sum to 1. Denote by $\bar{a}$ a*

barycenter of those measures (w.r.t. to the Wasserstein-2 distance). For any $s \in [K]$, the transport map from a_s to the barycenter $\bar{a}$ is given by

$$T_{a_s \to \bar{a}} = \left(\sum_{s'=1}^{K} w_{s'} F_{s'}^{-1} \circ F_s \right) \ ,$$

where F_s is the cumulative distribution function of a_s and F_s^{-1} denotes the generalized inverse of F_s defined as

$$F_s^{-1}(t) = \inf\{x : F_s(x) \geq t\} \ .$$

In particular, the constant speed geodesic $\gamma_s(\cdot)$ from a_s to $\bar{a}$ is given by

$$\gamma_s(t) = ((1-t)\,\mathrm{Id} + tT_{a_s \to \bar{a}})\#a_s, \quad t \in [0,1] \ .$$

7.8.2 Tail inequalities

The next result can be found in Laurent and Massart 2000, Lemma 1.

Lemma 7.8.4. *Let $\zeta_1, \ldots, \zeta_p$ be* i.i.d. *standard Gaussian random variables and let $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$ be component-wise non-negative. Then*

$$\mathbf{P}\left(\sum_{j=1}^{p} a_j(\zeta_j^2 - 1) \geq 2\,\|\boldsymbol{a}\|_2 \sqrt{t} + 2\,\|\boldsymbol{a}\|_\infty t \right) \leq \exp(-t), \quad \forall t \geq 0 \ .$$

In particular, setting $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_p)^\top$ and applying the previous result with $a_1 = \ldots = a_p = 1$ we get

$$\mathbf{P}\left(\|\boldsymbol{\zeta}\|_2^2 \geq p + 2\sqrt{pt} + 2t \right) \leq \exp(-t), \quad \forall t \geq 0$$

We need one result from random matrix theory to control the smallest and largest singular values of a Gaussian matrix, see Vershynin 2010, Corollary 5.35.

Lemma 7.8.5. *Let $\mathbf{A}$ be an $N \times m$ matrix whose entries are independent standard normal random variables. Then,*

$$\mathbf{P}\left(\sigma_{\min}(\mathbf{A}) \leq \sqrt{N} - \sqrt{m} - t \right) \vee \mathbf{P}\left(\sigma_{\max}(\mathbf{A}) \geq \sqrt{N} + \sqrt{m} + t \right) \leq \exp(-t^2/2), \quad \forall t \geq 0 \ .$$

7.9 Proofs for Section 7.4

7.9.1 Auxiliary results

The next result is taken from (Le Gouic, Loubes, and Rigollet 2020, Theorem 3).

Lemma 7.9.1. *Let $f : \mathbb{R}^p \times [K] \to \mathbb{R}$ be any measurable function. Let Assumption 7.2.1 be satisfied, then*

$$\mathcal{R}(f) \geq \sum_{s=1}^{K} w_s W_2^2\Big(f(\cdot, s)\#\mu_{\boldsymbol{X}|s}, f^*(\cdot, s)\#\mu_{\boldsymbol{X}|s}\Big) \ .$$

Lemma 7.9.2 (Minkowski's inequality)**.** *Let $(\mathcal{X}, d)$ be a metric space. Fix an integer $K \geq 2$, a weight vector $w \in \Delta^{K-1}$ and define the mapping $d_w : \mathcal{X}^K \times \mathcal{X}^K \to \mathbb{R}$ as*

$$d_w(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, b_s)}, \qquad \text{for any } \boldsymbol{a}, \boldsymbol{b} \in \mathcal{X}^K \ .$$

Then, d_w is a pseudo-metric on the product space $\mathcal{X}^K$.

Proof. The mapping d_w is clearly symmetric and non-negative. We only have to check the triangle inequality. Fix arbitrary $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathcal{X}^K$. Then, by triangular inequalities on the distance d and Jensen's inequality,

$$\begin{aligned}\sum_{s=1}^{K} w_s d^2(a_s, b_s) &\leq \sum_{s=1}^{K} w_s d(a_s, b_s) d(a_s, c_s) + \sum_{s=1}^{K} w_s d(a_s, b_s) d(c_s, b_s) \\ &\leq \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, b_s)} \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, c_s)} + \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, b_s)} \sqrt{\sum_{s=1}^{K} w_s d^2(c_s, b_s)} \ .\end{aligned}$$

That is,

$$\begin{aligned}d_w(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, b_s)} &\leq \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, c_s)} + \sqrt{\sum_{s=1}^{K} w_s d^2(c_s, b_s)} \\ &= d_w(\boldsymbol{a}, \boldsymbol{c}) + d_w(\boldsymbol{c}, \boldsymbol{b}) \ .\end{aligned}$$

□

Lemma 7.9.3. *Let $\boldsymbol{a} = (a_1, \ldots, a_K) \in \mathcal{X}^K$, $\boldsymbol{w} = (w_1, \ldots, w_K)^\top \in \Delta^{K-1}$. Assume that $\boldsymbol{b} = (b_1, \cdots, b_K) \in \mathcal{X}^K$ satisfies (P_1)–(P_2), then*

$$\sqrt{\sum_{s=1}^{K} w_s d^2(b_s, C_{\boldsymbol{b}})} = \sqrt{\sum_{s=1}^{K} w_s d^2(b_s, C_{\boldsymbol{a}})} \ .$$

Proof. Let $C_{\boldsymbol{b}}$ be a barycenter of $(b_s)_{s\in[K]}$ with weights $(w_s)_{s\in[K]}$, then by Lemma 7.9.2 it holds that

$$\sqrt{\sum_{s=1}^{K} w_s d^2(a_s, C_{\boldsymbol{b}})} \leq \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, b_s)} + \sqrt{\sum_{s=1}^{K} w_s d^2(b_s, C_{\boldsymbol{b}})} \ . \tag{7.15}$$

The following chain of inequalities holds thanks to Eq. (7.15) and properties (P_1)–(P_2)

$$\begin{aligned}\sqrt{\sum_{s=1}^{K} w_s d^2(b_s, C_{\boldsymbol{b}})} &\geq \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, C_{\boldsymbol{b}})} - \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, b_s)} \\ &\geq \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, C_{\boldsymbol{a}})} - \sqrt{\sum_{s=1}^{K} w_s d^2(a_s, b_s)} \\ &= \frac{1}{\sqrt{\alpha}} \sqrt{\sum_{s=1}^{K} w_s d^2(b_s, C_{\boldsymbol{a}})} - \frac{1-\sqrt{\alpha}}{\sqrt{\alpha}} \sqrt{\sum_{s=1}^{K} w_s d^2(b_s, C_{\boldsymbol{a}})} \\ &= \sqrt{\sum_{s=1}^{K} w_s d^2(b_s, C_{\boldsymbol{a}})} \ .\end{aligned}$$

The converse inequality follows from the definition of $C_{\boldsymbol{b}}$, which concludes the proof. □

7.9.2 Proof of Proposition 7.4.1

Let $\alpha \in [0,1]$. For any $s \in [K]$, define

$$a_s = f^*(\cdot, s)\#\mu_{\boldsymbol{X}|s} = \nu_s^* \ , \tag{7.16}$$

Let γ_s be the (constant-speed) geodesic between a_s and $C_{\boldsymbol{a}}$ *i.e.*, $\gamma_s(0) = a_s$, $\gamma_s(1) = C_{\boldsymbol{a}}$ and $W_2(\gamma_s(t_1), \gamma_s(t_2)) = |t_2 - t_1| W_2(a_s, C_{\boldsymbol{a}})$ for any $t_1, t_2 \in [0,1]$. Note that the uniqueness of the geodesic come from the particular structure of the Wasserstein-2 space on the real line, see *e.g.,* Kloeckner 2010, Section 2.2. We define $b_s := \gamma_s(1-\sqrt{\alpha})$ for $s \in [K]$. Let us show that $\boldsymbol{b} = (b_s)_{s\in[K]}$ satisfies the properties (P_1)–(P_2) of the Geometric Lemma 7.4.3 when considering $\boldsymbol{a} = (a_s)_{s\in[K]}$ with the weights $(w_s)_{s\in[K]}$ and $d \equiv W_2$. By construction of $b_s = \gamma_s(1-\sqrt{\alpha})$, we have

$$W_2(b_s, C_{\boldsymbol{a}}) = \sqrt{\alpha} W_2(a_s, C_{\boldsymbol{a}}) \ , \tag{7.17}$$

$$W_2(b_s, a_s) = (1-\sqrt{\alpha}) W_2(a_s, C_{\boldsymbol{a}}) \ . \tag{7.18}$$

This shows that $\boldsymbol{b} = (b_s)_{s\in[K]}$ satisfies (P_1) and (P_2). Therefore, using Lemma 7.4.3 we get

$$\sum_{s=1}^{K} w_s W_2^2(b_s, a_s) = \inf_{\boldsymbol{b}\in\mathcal{P}_2^K(\mathbb{R})} \left\{ \sum_{s=1}^{K} w_s W_2^2(b_s, a_s) \ : \ \sum_{s=1}^{K} w_s W_2^2(b_s, C_{\boldsymbol{b}}) \leq \alpha \sum_{s=1}^{K} w_s d^2(a_s, C_{\boldsymbol{a}}) \right\} \ . \tag{7.19}$$

Finally, thanks to the Assumption 7.2.1 which says that that $a_s = \nu_s^*$ is atomless the constant speed geodesic γ_s between a_s and $C_{\boldsymbol{a}}$ can be written as

$$\begin{aligned}\gamma_s(t) &= \left((1-t)\,\mathrm{Id} + t \left(\sum_{s'=1}^{K} w_{s'} F^{-1}_{a_{s'}} \circ F_{a_s} \right) \right) \#a_s \\ &= \left\{ \left((1-t)\,\mathrm{Id} + t \left(\sum_{s'=1}^{K} w_{s'} F^{-1}_{a_{s'}} \circ F_{a_s} \right) \right) \circ f^*(\cdot, s) \right\} \#\mu_{\boldsymbol{X}|s}, \quad t \in [0,1] \ .\end{aligned}$$

See Appendix 7.8.1 for details about the first equality. Substituting $t = 1-\sqrt{\alpha}$ to γ_s, the expression for b_s is

$$b_s = \left\{ \left(\sqrt{\alpha}\,\mathrm{Id} + (1-\sqrt{\alpha}) \left(\sum_{s'=1}^{K} w_{s'} F^{-1}_{a_{s'}} \circ F_{a_s} \right) \right) \circ f^*(\cdot, s) \right\} \# \mu_{\boldsymbol{X}|s} \ . \tag{7.20}$$

We define f^*_α for all $(\boldsymbol{x}, s) \in \mathbb{R}^p \times [K]$ as

$$f^*_\alpha(\boldsymbol{x}, s) = \sqrt{\alpha} f^*(\boldsymbol{x}, s) + (1-\sqrt{\alpha}) \sum_{s'=1}^{K} w_{s'} F^{-1}_{a_{s'}} \left(F_{a_s}(f^*(\boldsymbol{x}, s)) \right) \ , \tag{7.21}$$

then after Eq. (7.20) it holds that $b_s = f^*_\alpha(\cdot, s) \# \mu_{\boldsymbol{X}|s}$ and

$$W_2^2(b_s, a_s) = \mathbb{E}\left[(f^*(\boldsymbol{X}, S) - f^*_\alpha(\boldsymbol{X}, S))^2 \mid S = s \right] \ . \tag{7.22}$$

with $\mathcal{U}(f^*_\alpha) = \alpha \mathcal{U}(f^*)$. Moreover, Lemma 7.9.1 implies that for any f such that $\mathcal{U}(f) \leq \alpha \mathcal{U}(f^*)$ we have

$$\begin{aligned} \mathbb{E}(f^*(X, S) - f(X, S))^2 \geq \sum_{s=1}^{K} w_s W_2^2(b_s, a_s) &= \sum_{s=1}^{K} w_s \mathbb{E}\left[(f^*(\boldsymbol{X}, S) - f^*_\alpha(\boldsymbol{X}, S))^2 \mid S = s \right] \\ &= \mathcal{R}(f^*_\alpha) \ . \end{aligned}$$

Thus, f^*_α is the optimal fair prediction with α relative improvement. The proof is concluded.

7.10 Proof of Theorem 7.5.3

To ease the notation we write δ_n instead of $\delta_n(\mathcal{F}, \Theta, t)$. We also define

$$\Psi(\widehat{f}, (f^*, \boldsymbol{\theta})) := \mathbf{P}_{(f^*, \boldsymbol{\theta})} \left(\mathcal{R}^{1/2}(\widehat{f}) \geq \delta_n^{1/2} \vee (1-\sqrt{\alpha}) \mathcal{U}^{1/2}(f^*) \right) \ .$$

We split the proof according to two complementary cases.

Case 1: there exists $(f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta$ such that $\delta_n \leq (1-\sqrt{\alpha})^2 \mathcal{U}(f^*)$. In this case, for such couple $(f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta$ and for any estimator $\widehat{f} \in \widehat{\mathcal{F}}_{(\alpha,t')}$ we have

$$\begin{aligned} \Psi(\widehat{f}, (f^*, \boldsymbol{\theta})) &\geq \mathbf{P}_{(f^*, \boldsymbol{\theta})} \left(\mathcal{R}^{1/2}(\widehat{f}) \geq \delta_n^{1/2} \vee (1-\sqrt{\alpha}) \mathcal{U}^{1/2}(f^*), \mathcal{U}(\widehat{f}) \leq \alpha \mathcal{U}(f^*) \right) \\ &\overset{\text{def. of } f^*_\alpha}{\geq} \mathbf{P}_{(f^*, \boldsymbol{\theta})} \left(\mathcal{R}^{1/2}(f^*_\alpha) \geq \delta_n^{1/2} \vee (1-\sqrt{\alpha}) \mathcal{U}^{1/2}(f^*), \mathcal{U}(\widehat{f}) \leq \alpha \mathcal{U}(f^*) \right) \\ &\overset{\text{Lemma 7.4.5}}{=} \mathbf{P}_{(f^*, \boldsymbol{\theta})} \left(\mathcal{U}(\widehat{f}) \leq \alpha \mathcal{U}(f^*) \right) \mathbb{I}\left\{ \delta_n \leq (1-\sqrt{\alpha})^2 \mathcal{U}(f^*) \right\} \ . \end{aligned}$$

Note that by definition of $\widehat{\mathcal{F}}_{(\alpha,t')}$ it holds that

$$\forall \widehat{f} \in \widehat{\mathcal{F}}_{(\alpha,t')}, \forall (f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta, \quad \mathbf{P}_{(f^*, \boldsymbol{\theta})} \left(\mathcal{U}(\widehat{f}) \leq \alpha \mathcal{U}(f^*) \right) \geq 1 - t' \ .$$

Since in the considered case there exists a couple $(f^*, \boldsymbol{\theta}) \in \widehat{\mathcal{F}}_{(\alpha,t')} \times \Theta$ such that $\delta_n \leq (1-\sqrt{\alpha})^2 \mathcal{U}(f^*)$, by definition of $\widehat{\mathcal{F}}_{(\alpha,t')}$ we have

$$\inf_{\widehat{f} \in \widehat{\mathcal{F}}_{(\alpha,t')}} \sup_{(f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta} \Psi(\widehat{f}, (f^*, \boldsymbol{\theta})) \geq 1 - t' \ . \tag{7.23}$$

Case 2: for any couple $(f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta$ it holds that $\delta_n > (1-\sqrt{\alpha})^2 \mathcal{U}(f^*)$. In this case, for any couple $(f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta$ and for any estimator $\widehat{f} \in \widehat{\mathcal{F}}_{(\alpha,t')}$,

$$\Psi(\widehat{f}, (f^*, \boldsymbol{\theta})) = \mathbf{P}_{(f^*,\boldsymbol{\theta})} \left(\mathcal{R}(\widehat{f}) \geq \delta_n \right) \ .$$

By definition of δ_n it holds in this case that

$$\begin{aligned} \inf_{\widehat{f} \in \widehat{\mathcal{F}}_{(\alpha,t')}} \sup_{(f^*,\boldsymbol{\theta}) \in \mathcal{F} \times \Theta} \Psi(\widehat{f}, (f^*, \boldsymbol{\theta})) &\geq \inf_{\widehat{f}} \sup_{(f^*,\boldsymbol{\theta}) \in \mathcal{F} \times \Theta} \Psi(\widehat{f}, (f^*, \boldsymbol{\theta})) \\ &= \inf_{\widehat{f}} \sup_{(f^*,\boldsymbol{\theta}) \in \mathcal{F} \times \Theta} \mathbf{P}_{(f^*,\boldsymbol{\theta})} \left(\mathcal{R}(\widehat{f}) \geq \delta_n \right) \geq t \ . \end{aligned} \tag{7.24}$$

Putting two cases together, and in particular using Eqs. (7.23) and (7.24) we obtain

$$\inf_{\widehat{f} \in \widehat{\mathcal{F}}_{(\alpha,t')}} \sup_{(f^*,\boldsymbol{\theta}) \in \mathcal{F} \times \Theta} \Psi(\widehat{f}, (f^*, \boldsymbol{\theta})) \geq \begin{cases} 1 - t' & \text{if} \quad \exists (f^*, \boldsymbol{\theta}) \in \mathcal{F} \times \Theta \text{ s.t. } \delta_n \leq (1-\sqrt{\alpha})^2 \mathcal{U}(f^*) \\ t & \text{otherwise} \end{cases} .$$

We conclude the proof observing that the *r.h.s.* of the last inequality is lower bounded by $t \wedge (1 - t')$.

7.11 Proofs for Section 7.6

Additional notation We denote by $\mathbb{S}^{p-1}$ the unit sphere in $\mathbb{R}^p$. For any matrix $\mathbf{A}$ we denote by $\|\mathbf{A}\|_{\text{op}}$, the operator norm of $\mathbf{A}$. We denote by $\chi^2(p)$ the standard chi-square distribution with p degrees of freedom and by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We denote by $\mathbf{I}_p$ the identity matrix of size $p \times p$.

7.11.1 Proof of Lemma 7.6.3

Throughout the proof we implicitly condition on the observations. Let $\tau \in [0, 1]$. For each $s \in [K]$ we set $\widehat{m}_s = \sqrt{\tau}\widehat{b}_s + (1 - \sqrt{\tau}) \sum_{s=1}^K w_s \widehat{b}_s$. Note that for all $s \in [K]$, $(\widehat{f}_\tau(\boldsymbol{X}, S) | S = s) \sim \mathcal{N}(\widehat{m}_s, \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}} \rangle)$. Therefore, by the definition of the unfairness and Lemma 7.8.2

$$\begin{aligned} \mathcal{U}(\widehat{f}_\tau) &= \min_{\nu} \sum_{s=1}^K w_s W_2^2 \left(\mathcal{N}(\widehat{m}_s, \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}} \rangle), \nu \right) \\ &= \sum_{s=1}^K w_s W_2^2 \left(\mathcal{N}(\widehat{m}_s, \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}} \rangle), \mathcal{N}(\bar{m}, \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}} \rangle) \right) \ , \end{aligned}$$

where $\bar{m} = \sum_{s=1}^K w_s \widehat{m}_s$. We conclude the proof by noticing that thanks to Lemma 7.8.1 it holds that

$$\begin{aligned} W_2^2 \left(\mathcal{N}(\widehat{m}_s, \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}} \rangle), \mathcal{N}(\bar{m}, \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}} \rangle) \right) &= (\widehat{m}_s - \bar{m})^2 \\ &= \left\{ \sqrt{\tau}\widehat{b}_s + (1 - \sqrt{\tau}) \sum_{s=1}^K w_s \widehat{b}_s - \sum_{s=1}^K w_s \widehat{b}_s \right\}^2 \ . \end{aligned}$$

The proof is concluded.

7.11.2 Auxiliary results for Theorem 7.6.4

Lemma 7.11.1 (Fixed design analysis)**.** *Define the following matrix of size* $(p+K)\times(p+K)$

$$\widehat{\mathbf{\Psi}} = \left[\begin{array}{c|c} \frac{1}{2}\sum_{s=1}^K w_s \mathbf{X}_s^\top \mathbf{X}_s/n_s & \mathbf{O} \\ \hline \mathbf{O}^\top & \frac{1}{2}\mathbf{W} \end{array}\right] \ ,$$

where $\mathbf{O} = [w_1\bar{\boldsymbol{X}}_1, \ldots, w_K\bar{\boldsymbol{X}}_K] \in \mathbb{R}^{p\times K}$ *and* $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_K)$*. For all* $t \geq 0$ *it holds that*

$$\mathbf{P}\left(\|\widehat{\mathbf{\Psi}}^{1/2}\widehat{\mathbf{\Delta}}\|_2^2 \geq \sigma^2\left\{\left(\frac{p}{n}+\frac{K}{n}\right) + 2\left(\sqrt{\frac{p}{n}}+\sqrt{\frac{K}{n}}\right)\sqrt{\frac{t}{n}} + 4\frac{t}{n}\right\} \,\middle|\, \mathbf{X}_{1:K}\right) \leq 2\exp(-t) \ ,$$

where $\widehat{\mathbf{\Delta}} = (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*, \widehat{\boldsymbol{b}}-\boldsymbol{b}^*) \in \mathbb{R}^p \times \mathbb{R}^K$ *and* $\mathbf{X}_{1:K} = (\mathbf{X}_1, \ldots, \mathbf{X}_K)$*.*

Proof. By optimality of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}})$ and the linear model assumption in Eq. (7.11) it holds that

$$\sum_{s=1}^K w_s \left\|\boldsymbol{Y}_s - \mathbf{X}_s\widehat{\boldsymbol{\beta}} - \widehat{b}_s\mathbf{1}_{n_s}\right\|_{n_s}^2 \leq \sum_{s=1}^K w_s \left\|\boldsymbol{\xi}_s\right\|_{n_s}^2 \ .$$

After simplification, the above yields

$$\begin{aligned}\sum_{s=1}^K w_s \left\|\mathbf{X}_s(\boldsymbol{\beta}^*-\widehat{\boldsymbol{\beta}}) + (b_s^*-\widehat{b}_s)\mathbf{1}_{n_s}\right\|_{n_s}^2 &\leq 2\sum_{s=1}^K w_s \left\langle \mathbf{X}_s(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) + (\widehat{b}_s - b_s^*)\mathbf{1}_{n_s}, \boldsymbol{\xi}_s/n_s\right\rangle \\ &= 2\left\langle \widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*, \sum_{s=1}^K \mathbf{X}_s^\top\boldsymbol{\xi}_s/n\right\rangle + 2\sum_{s=1}^K w_s(\widehat{b}_s - b_s^*)\bar{\boldsymbol{\xi}}_s \ ,\end{aligned}$$

where $\bar{\boldsymbol{\xi}}_s = (1/n_s)\sum_{i=1}^{n_s}(\boldsymbol{\xi}_s)_i$. Using Young's inequality, we can write

$$\begin{aligned}2\left\langle \widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*, \sum_{s=1}^K \mathbf{X}_s^\top\boldsymbol{\xi}_s/n\right\rangle &\leq \frac{1}{2}\sum_{s=1}^K w_s\|\mathbf{X}_s(\boldsymbol{\beta}^*-\widehat{\boldsymbol{\beta}})\|_{n_s}^2 + 2\left(\frac{\left\langle \widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*, \sum_{s=1}^K \mathbf{X}_s^\top\boldsymbol{\xi}_s/n\right\rangle}{\sqrt{\sum_{s=1}^K w_s\|\mathbf{X}_s(\boldsymbol{\beta}^*-\widehat{\boldsymbol{\beta}})\|_{n_s}^2}}\right)^2 \\ &\leq \frac{1}{2}\sum_{s=1}^K w_s\|\mathbf{X}_s(\boldsymbol{\beta}^*-\widehat{\boldsymbol{\beta}})\|_{n_s}^2 + 2\sup_{\mathbf{\Delta}\in\mathbb{R}^p}\left(\frac{\left\langle \mathbf{\Delta}, \sum_{s=1}^K \mathbf{X}_s^\top\boldsymbol{\xi}_s/n\right\rangle}{\sqrt{\sum_{s=1}^K w_s\|\mathbf{X}_s\mathbf{\Delta}\|_{n_s}^2}}\right)^2 \ .\end{aligned}$$

We also observe that again thanks to Young's inequality

$$2\sum_{s=1}^K w_s(\widehat{b}_s - b_s^*)\bar{\boldsymbol{\xi}}_s \leq \frac{1}{2}\sum_{s=1}^K w_s(\widehat{b}_s-b_s^*)^2 + 2\sum_{s=1}^K w_s\bar{\boldsymbol{\xi}}_s^2 \ .$$

Putting everything together, we have shown that

$$\|\widehat{\mathbf{\Psi}}^{1/2}\widehat{\mathbf{\Delta}}\|_2^2 \leq 2\sup_{\mathbf{\Delta}\in\mathbb{R}^p}\left(\frac{\left\langle \mathbf{\Delta}, \sum_{s=1}^K \mathbf{X}_s^\top\boldsymbol{\xi}_s/n\right\rangle}{\sqrt{\sum_{s=1}^K w_s\|\mathbf{X}_s\mathbf{\Delta}\|_{n_s}^2}}\right)^2 + 2\sum_{s=1}^K w_s\bar{\boldsymbol{\xi}}_s^2 \ . \tag{7.25}$$

Notice that since $\boldsymbol{\xi}_s \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_s})$, then conditionally on $\mathbf{X}_1, \dots, \mathbf{X}_K$,

$$\sum_{s=1}^K \mathbf{X}_s^\top \boldsymbol{\xi}_s / n \overset{d}{=} \frac{\sigma}{n} \left(\sum_{s=1}^K \mathbf{X}_s^\top \mathbf{X}_s \right)^{1/2} \boldsymbol{\zeta} \ ,$$

where $\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Besides, since $w_s = {}^{n_s}/_n$, it holds for all $\boldsymbol{\Delta} \in \mathbb{R}^p$ that

$$\sum_{s=1}^K w_s \|\mathbf{X}_s \boldsymbol{\Delta}\|_{n_s}^2 = \boldsymbol{\Delta}^\top \left(\frac{1}{n} \sum_{s=1}^K \mathbf{X}_s^\top \mathbf{X}_s \right) \boldsymbol{\Delta} = \left\| \left(\frac{1}{n} \sum_{s=1}^K \mathbf{X}_s^\top \mathbf{X}_s \right)^{1/2} \boldsymbol{\Delta} \right\|_2^2 \ .$$

The above implies that conditionally on $\mathbf{X}_1, \dots, \mathbf{X}_K$,

$$\sqrt{U} := \sup_{\boldsymbol{\Delta} \in \mathbb{R}^p} \frac{\left\langle \boldsymbol{\Delta}, \sum_{s=1}^K \mathbf{X}_s^\top \boldsymbol{\xi}_s / n \right\rangle}{\sqrt{\sum_{s=1}^K w_s \|\mathbf{X}_s \boldsymbol{\Delta}\|_{n_s}^2}} \overset{d}{=} \frac{\sigma}{\sqrt{n}} \sup_{\boldsymbol{\Delta} \in \mathbb{R}^p} \frac{\left\langle \left(\sum_{s=1}^K \mathbf{X}_s^\top \mathbf{X}_s \right)^{1/2} \boldsymbol{\Delta}, \boldsymbol{\zeta} \right\rangle}{\left\| \left(\sum_{s=1}^K \mathbf{X}_s^\top \mathbf{X}_s \right)^{1/2} \boldsymbol{\Delta} \right\|_2} \ . \tag{7.26}$$

Note that for any random variable $\boldsymbol{\zeta}$ taking values in $\mathbb{R}^p$,

$$\sup_{\boldsymbol{\Delta} \in \mathbb{R}^p} \frac{\left\langle \left(\sum_{s=1}^K \mathbf{X}_s^\top \mathbf{X}_s \right)^{1/2} \boldsymbol{\Delta}, \boldsymbol{\zeta} \right\rangle}{\left\| \left(\sum_{s=1}^K \mathbf{X}_s^\top \mathbf{X}_s \right)^{1/2} \boldsymbol{\Delta} \right\|_2} \leq \|\boldsymbol{\zeta}\|_2 \ \text{ almost surely.} \tag{7.27}$$

Furthermore, recalling that $\bar{\boldsymbol{\xi}}_s \sim \mathcal{N}(\mathbf{0}, {}^1/_{n_s})$ we get

$$V := \sum_{s=1}^K w_s \bar{\boldsymbol{\xi}}_s^2 \sim \frac{\sigma^2}{n} \chi^2(K) \ . \tag{7.28}$$

For any $u, v \in \mathbb{R}$ it holds that

$$\begin{aligned} \mathbf{P}\left(\|\widehat{\boldsymbol{\Psi}}^{1/2} \widehat{\boldsymbol{\Delta}}\|_2^2 \geq 2(u+v) \,|\, \mathbf{X}_{1:K} \right) &\overset{(7.25)}{\leq} \mathbf{P}\left(2(U+V) \geq 2(u+v) \,|\, \mathbf{X}_{1:K} \right) \\ &\overset{(a)}{\leq} \mathbf{P}\left(\frac{\sigma^2}{n} \chi^2(p) \geq u \,|\, \mathbf{X}_{1:K} \right) + \mathbf{P}\left(\frac{\sigma^2}{n} \chi^2(K) \geq v \,|\, \mathbf{X}_{1:K} \right) \ , \end{aligned}$$

where inequality (a) uses Eqs. (7.26) and (7.28) and the fact that $\mathbf{P}(U+V \geq u+v) \leq \mathbf{P}(U \geq u) + \mathbf{P}(V \geq v)$ for all random variables U, V and all $u, v \in \mathbb{R}$. Finally, setting $u = u_n(\sigma, p, t), v = v_n(\sigma, p, t)$ with

$$u_n(\sigma, p, t) = \frac{\sigma^2 p}{n} + 2\sigma^2 \sqrt{\frac{p}{n}} \sqrt{\frac{t}{n}} + 2\frac{\sigma^2 t}{n}, \quad v_n(\sigma, K, t) = \frac{\sigma^2 K}{n} + 2\sigma^2 \sqrt{\frac{K}{n}} \sqrt{\frac{t}{n}} + 2\frac{\sigma^2 t}{n} \ ,$$

we obtain the stated result after application of Lemma 7.8.4 in appendix

$$\mathbf{P}\left(\|\widehat{\boldsymbol{\Psi}}^{1/2} \widehat{\boldsymbol{\Delta}}\|_2^2 \geq 2(u_n(\sigma, p, t) + v_n(\sigma, p, t)) \,|\, \mathbf{X}_{1:K} \right) \leq 2\exp(-t) \ .$$

□

Theorem 7.11.2 (From fixed to random design)**.** *Define,*

$$\delta_n(p,K,t) = 8\left(\frac{p}{n}+\frac{K}{n}\right) + 16\left(\sqrt{\frac{p}{n}}+\sqrt{\frac{K}{n}}\right)\sqrt{\frac{t}{n}}+\frac{32t}{n} \ .$$

Consider $p, K \in \mathbb{N}, t \geq 0$ *and define* $\gamma(p,K,t) = {}^{(4\sqrt{K}+5\sqrt{t}+6\sqrt{p})}/_{(\sqrt{p}+\sqrt{t})}$*. Assume that* $\sqrt{n} \geq 2{}^{(\sqrt{p}+\sqrt{t})}/_{(\gamma(p,K,t)-\sqrt{\gamma^2(p,K,t)-3})}$*, then with probability at least* $1-4\exp(-t/2)$

$$\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_2^2 + \sum_{s=1}^{K} w_s(b_s^* - \widehat{b}_s)^2 \leq \sigma^2\delta_n(p,K,t) \ .$$

Proof. Define the $(p+K)\times(p+K)$ matrix

$$\boldsymbol{\Psi} = \frac{1}{2}\left[\begin{array}{c|c} \boldsymbol{\Sigma} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{W} \end{array}\right] \ , \tag{7.29}$$

then under notation of Lemma 7.11.1 we can write

$$\begin{aligned}
\|\widehat{\boldsymbol{\Psi}}^{1/2}\widehat{\boldsymbol{\Delta}}\|_2^2 &= \widehat{\boldsymbol{\Delta}}^\top\boldsymbol{\Psi}^{1/2}\boldsymbol{\Psi}^{-1/2}\widehat{\boldsymbol{\Psi}}\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Psi}^{1/2}\widehat{\boldsymbol{\Delta}} \\
&= \widehat{\boldsymbol{\Delta}}^\top\boldsymbol{\Psi}^{1/2}\boldsymbol{\Psi}^{-1/2}\left(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}\right)\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Psi}^{1/2}\widehat{\boldsymbol{\Delta}} + \widehat{\boldsymbol{\Delta}}^\top\boldsymbol{\Psi}\widehat{\boldsymbol{\Delta}} \\
&\geq \left(1+\lambda_{\min}\left(\boldsymbol{\Psi}^{-1/2}\left(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}\right)\boldsymbol{\Psi}^{-1/2}\right)\right)\|\boldsymbol{\Psi}^{1/2}\widehat{\boldsymbol{\Delta}}\|_2^2 \ .
\end{aligned} \tag{7.30}$$

If we set $\widehat{\boldsymbol{\Sigma}} = \sum_{s=1}^K w_s\mathbf{X}_s^\top\mathbf{X}_s/n_s$, then

$$\boldsymbol{\Psi}^{-1/2}\left(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}\right)\boldsymbol{\Psi}^{-1/2} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1/2} & 2\boldsymbol{\Sigma}^{-1/2}\mathbf{O}\mathbf{W}^{-1/2} \\ \hline 2\mathbf{W}^{-1/2}\mathbf{O}^\top\boldsymbol{\Sigma}^{-1/2} & \mathbf{0} \end{array}\right] \ .$$

Furthermore, by Courant-Fisher theorem it holds that

$$\lambda_{\min}\left(\boldsymbol{\Psi}^{-1/2}\left(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}\right)\boldsymbol{\Psi}^{-1/2}\right) \geq \lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1/2}\right) - 4\|\boldsymbol{\Sigma}^{-1/2}\mathbf{O}\mathbf{W}^{-1/2}\|_{\text{op}} \ . \tag{7.31}$$

Using the definition of $\mathbf{O}$ we can write

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{O}\mathbf{W}^{-1/2} = [w_1^{1/2}\boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{X}}_1, \ldots, w_K^{1/2}\boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{X}}_K] \ .$$

Note that the random variable on right hand side of Eq. (7.31) is independent from $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$. Recall that since $w_s = {}^{n_s}/_n$ and $\bar{\boldsymbol{X}}_s \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}/n)$, then for all $s = 1, \ldots, K$ it holds that

$$w_s^{1/2}\boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{X}}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p/n) \ ,$$

and these vectors are independent. Hence, the matrix $\boldsymbol{\Sigma}^{-1/2}\mathbf{O}\mathbf{W}^{-1/2} \in \mathbb{R}^{p\times K}$ has i.i.d. Gaussian entries with variance ${}^1/_n$. Therefore, by Lemma 7.8.5 we get

$$\mathbf{P}\left(\|\boldsymbol{\Sigma}^{-1/2}\mathbf{O}\mathbf{W}^{-1/2}\|_{\text{op}} \geq \sqrt{\frac{p}{n}}+\sqrt{\frac{K}{n}}+\sqrt{\frac{t}{n}}\right) \leq \exp(-t/2) \ . \tag{7.32}$$

Furthermore, we observe that

$$\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} \stackrel{d}{=} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top \ ,$$

where $\boldsymbol{\zeta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. It implies that

$$\boldsymbol{\Sigma}^{-1/2}\left(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1/2} \stackrel{d}{=} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top - \mathbf{I}_p = \frac{1}{n}\left(\mathbf{Z}^\top\mathbf{Z} - n\mathbf{I}_p\right) \ ,$$

where $\mathbf{Z}$ is a matrix of size $n \times p$ with i^{th}-row being equal to $\boldsymbol{\zeta}_i^\top$. Note that the spectral theorem and the relation between eigenvalues of $\mathbf{Z}^\top\mathbf{Z}$ and the singular values of $\mathbf{Z}$ imply that

$$n\lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\left(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1/2}\right) \stackrel{d}{=} \lambda_{\min}\left(\mathbf{Z}^\top\mathbf{Z} - n\mathbf{I}_p\right) = \sigma^2_{\min}(\mathbf{Z}) - n \ .$$

where $\sigma_{\min}(\mathbf{Z})$ is the maximal singular value of $\mathbf{Z}$. Applying Lemma 7.8.5 from appendix we get for all $t \geq \sqrt{n} - \sqrt{p}$ that $\mathbf{P}\left(\frac{1}{n}\sigma^2_{\min}(Z) \leq \frac{1}{n}(\sqrt{n} - \sqrt{p} - t)^2\right)$ equals to

$$\mathbf{P}\left(\frac{1}{n}(\sigma^2_{\min}(Z) - n) \leq \frac{p}{n} + 2\sqrt{\frac{p}{n}}\frac{t}{\sqrt{n}} + \frac{t^2}{n} - 2\sqrt{\frac{p}{n}} - 2\frac{t}{\sqrt{n}}\right) \leq \exp(-t^2/2) \ .$$

Changing variables $t^2 \mapsto t$ we get

$$\mathbf{P}\left(\lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\left(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1/2}\right) \leq \frac{p}{n} + 2\sqrt{\frac{p}{n}}\sqrt{\frac{t}{n}} + \frac{t}{n} - 2\sqrt{\frac{p}{n}} - \sqrt{\frac{t}{n}}\right) \leq \exp(-t/2) \ . \tag{7.33}$$

Combining Eqs. (7.31),(7.32), and (7.33) we deduce that

$$\mathbf{P}\left(\lambda_{\min}\left(\boldsymbol{\Psi}^{-1/2}\left(\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\right)\boldsymbol{\Psi}^{-1/2}\right) \leq \psi_n(p, K.t)\right) \leq 2\exp(-t/2) \ ,$$

where $\psi_n(p, K, t) = \frac{p}{n} - 6\sqrt{\frac{p}{n}} + 2\sqrt{\frac{p}{n}}\sqrt{\frac{t}{n}} - 4\sqrt{\frac{K}{n}} + \frac{t}{n} - 5\sqrt{\frac{t}{n}}$. Applying Lemma 7.11.3 we deduce that under the assumption on n that $\psi_n(p, K, t) \geq -0.75$. Thus,

$$\mathbf{P}\left(\lambda_{\min}\left(\boldsymbol{\Psi}^{-1/2}\left(\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\right)\boldsymbol{\Psi}^{-1/2}\right) \leq -0.75\right) \leq 2\exp(-t/2) \ .$$

Combining the above fact with Eq. (7.30) and Lemma 7.11.1 we conclude that with probability at least $1 - 2\exp(-t) - 2\exp(-t/2)$

$$\|\boldsymbol{\Psi}^{1/2}\widehat{\boldsymbol{\Delta}}\|_2^2 \leq \sigma^2\left\{4\left(\frac{p}{n} + \frac{K}{n}\right) + 8\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{K}{n}}\right)\sqrt{\frac{t}{n}} + 16\frac{t}{n}\right\} = \sigma^2\frac{\delta_n(p, K, t)}{2} \ .$$

The statement of the lemma follows from the fact that

$$\|\boldsymbol{\Psi}^{1/2}\widehat{\boldsymbol{\Delta}}\|_2^2 = \frac{1}{2}\left(\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_2^2 + \sum_{s=1}^{K} w_s(b_s^* - \widehat{b}_s)^2\right) \ .$$

□

Lemma 7.11.3. *Consider $p, K \in \mathbb{N}, t \geq 0$ and define*

$$\gamma(p, K, t) = \frac{4\sqrt{K} + 5\sqrt{t} + 6\sqrt{p}}{\sqrt{p} + \sqrt{t}} \ .$$

For all $n, K, p \in \mathbb{N}, t \geq 0$, the following two conditions are equivalent

- $n \geq \left(\frac{2(\sqrt{p}+\sqrt{t})}{\gamma(p,K,t) - \sqrt{\gamma^2(p,K,t)-3}} \right)^2$;
- $\frac{p}{n} - 6\sqrt{\frac{p}{n}} + 2\sqrt{\frac{p}{n}}\sqrt{\frac{t}{n}} - 4\sqrt{\frac{K}{n}} + \frac{t}{n} - 5\sqrt{\frac{t}{n}} \geq -0.75.$

Proof. To simplify the notation and to save space we write γ instead of $\gamma(p, K, t)$. Let $x = n^{-1/2}$, we want to solve

$$x^2(\sqrt{p} + \sqrt{t})^2 - x(6\sqrt{p} + 4\sqrt{K} + 5\sqrt{t}) \geq -0.75$$

Set $y = x(\sqrt{p} + \sqrt{t})$, then thanks to the definition of γ, the previous inequality amounts to

$$y^2 - \gamma y + 0.75 \geq 0 \ .$$

The roots of the polynomial above are

$$x_-, x_+ = \frac{\gamma \pm \sqrt{\gamma^2 - 3}}{2} \ ,$$

which are both positive. The polynomial is non-negative outside the interval $(x_-, x_+) \subset \mathbb{R}_+$. Hence, a sufficient condition is to have

$$y \leq \frac{\gamma - \sqrt{\gamma^2 - 3}}{2} \ .$$

Substituting $x = n^{-1/2}$ and the expression for γ we conclude.

□

Lemma 7.11.4 (General unfairness control)**.** *Under notation of Lemma 7.6.3 it holds that, for any $\alpha \in [0, 1]$,*

$$\mathcal{U}(\widehat{f}_\alpha) \leq \alpha \mathcal{U}(f^*) \left\{ 1 + \mathrm{NUR} \sqrt{\frac{\sum_{s=1}^K w_s (\widehat{b}_s - b_s^*)^2}{\sigma^2}} \right\}^2 , \quad \textit{almost surely.} \tag{7.34}$$

Moreover,

$$\left| \mathcal{U}^{1/2}(\widehat{f}_1) - \mathcal{U}^{1/2}(f^*) \right| \leq \left\{ \sum_{s=1}^K w_s (\widehat{b}_s - b_s^*)^2 \right\}^{1/2} , \quad \textit{almost surely.} \tag{7.35}$$

Proof. Let U and V be discrete random variables such that $\mathbf{P}(U = \widehat{b}_s, V = b^*_{s'}) = w_s \delta_{s,s'}$, for any $s, s' \in [K]$. Note that, in particular, $\mathbf{P}(U = \widehat{b}_s) = w_s$ and $\mathbf{P}(V = b^*_s) = w_s$. Then, according to Lemma 7.6.3 and the definition of $\widehat{f}_\alpha$ it holds that

$$\mathcal{U}(\widehat{f}_\alpha) = \alpha \mathsf{Var}(U) \quad \text{and} \quad \mathcal{U}(\widehat{f}_\alpha) = \alpha \mathcal{U}(f^*) = \alpha \mathsf{Var}(V) \ .$$

Therefore, with our notations we have

$$\mathcal{U}(\widehat{f}_\alpha) - \alpha \mathcal{U}(f^*) = \alpha \left(\mathsf{Var}(U) - \mathsf{Var}(V) \right) \ . \tag{7.36}$$

Furthermore, for all $\varepsilon \in (0, 1)$ we have that $\mathsf{Var}(U)$ equals to

$$\begin{aligned} \mathsf{Var}(U - V + V) &= \mathsf{Var}(U - V) + 2\mathbf{E}[(U - V - \mathbf{E}[U] + \mathbf{E}[V])(V - \mathbf{E}[V])] + \mathsf{Var}(V) \\ &\leq \mathsf{Var}(U - V) + 2\sqrt{\mathsf{Var}(U - V)\mathsf{Var}(V)} + \mathsf{Var}(V) \\ &\leq \sum_{s=1}^K w_s (\widehat{b}_s - b^*_s)^2 + 2\sqrt{\mathcal{U}(f^*)} \sqrt{\sum_{s=1}^K w_s (\widehat{b}_s - b^*_s)^2} + \mathsf{Var}(V) \end{aligned} \tag{7.37}$$

Finally, combining Eqs. (7.36) and (7.37) we deduce

$$\mathcal{U}(\widehat{f}_\alpha) \leq \alpha \left(\sum_{s=1}^K w_s (\widehat{b}_s - b^*_s)^2 + 2\sqrt{\mathcal{U}(f^*)} \sqrt{\sum_{s=1}^K w_s (\widehat{b}_s - b^*_s)^2} + \mathcal{U}(f^*) \right) \ .$$

The proof of Eq. (7.34) is concluded after factorizing the square of the *r.h.s.* of the above bound. To prove Eq. (7.35), we set $\alpha = 1$ in Eq. (7.34) to get

$$\mathcal{U}^{1/2}(\widehat{f}_1) \leq \mathcal{U}^{1/2}(f^*) + \left\{ \sum_{s=1}^K w_s (\widehat{b}_s - b^*_s)^2 \right\}^{1/2} \ .$$

The converse bound is derived in a similar fashion using

$$\mathsf{Var}(V) \leq \mathsf{Var}(U - V) + 2\sqrt{\mathsf{Var}(U - V)\mathsf{Var}(U)} + \mathsf{Var}(U).$$

□

Lemma 7.11.5 (General risk control)**.** *Under notation of Lemma 7.6.3 it holds that*

$$\begin{aligned} \mathcal{R}(\widehat{f}_\alpha) \leq & \sum_{s=1}^K w_s \mathbb{E}(\langle \boldsymbol{X}, \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}} \rangle + (b^*_s - \widehat{b}_s))^2 \\ & + 2(1 - \sqrt{\alpha}) \sqrt{\sum_{s=1}^K w_s (b^*_s - \widehat{b}_s)^2} \sqrt{\sum_{s=1}^K w_s \left(\widehat{b}_s - \sum_{s'=1}^K w_{s'} \widehat{b}_{s'} \right)^2} \\ & + (1 - \sqrt{\alpha})^2 \sum_{s=1}^K w_s \left(\widehat{b}_s - \sum_{s'=1}^K w_{s'} \widehat{b}_{s'} \right)^2 \ . \end{aligned}$$

Proof. Recall the expression for $\widehat{f}_\alpha$

$$\widehat{f}_\alpha(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \widehat{\boldsymbol{\beta}} \rangle + \sqrt{\alpha}\widehat{b}_s + (1-\sqrt{\alpha}) \sum_{s=1}^{K} w_s \widehat{b}_s \ .$$

Using this expression, we can write for the risk of $\widehat{f}_\alpha$

$$\begin{aligned}
\mathcal{R}(\widehat{f}_\alpha) &= \sum_{s=1}^{K} w_s \mathbb{E}\left(\langle \boldsymbol{X}, \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}} \rangle + (b_s^* - \widehat{b}_s) + (1-\sqrt{\alpha}) \left(\widehat{b}_s - \sum_{s'=1}^{K} w_{s'} \widehat{b}_{s'} \right) \right)^2 \\
&\overset{(a)}{=} \sum_{s=1}^{K} w_s \mathbb{E}(\langle \boldsymbol{X}, \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}} \rangle + (b_s^* - \widehat{b}_s))^2 + 2(1-\sqrt{\alpha}) \sum_{s=1}^{K} w_s (b_s^* - \widehat{b}_s) \left(\widehat{b}_s - \sum_{s'=1}^{K} w_{s'} \widehat{b}_{s'} \right) \\
&\quad + (1-\sqrt{\alpha})^2 \sum_{s=1}^{K} w_s \left(\widehat{b}_s - \sum_{s'=1}^{K} w_{s'} \widehat{b}_{s'} \right)^2 \\
&\overset{(b)}{\leq} \sum_{s=1}^{K} w_s \mathbb{E}(\langle \boldsymbol{X}, \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}} \rangle + (b_s^* - \widehat{b}_s))^2 + (1-\sqrt{\alpha})^2 \sum_{s=1}^{K} w_s \left(\widehat{b}_s - \sum_{s'=1}^{K} w_{s'} \widehat{b}_{s'} \right)^2 \\
&\quad + 2(1-\sqrt{\alpha}) \sqrt{\sum_{s=1}^{K} w_s (b_s^* - \widehat{b}_s)^2} \sqrt{\sum_{s=1}^{K} w_s \left(\widehat{b}_s - \sum_{s'=1}^{K} w_{s'} \widehat{b}_{s'} \right)^2}
\end{aligned}$$

where (a) follows from the fact that $\boldsymbol{X}$ is centered and (b) is due to the Cauchy-Schwarz inequality. □

Theorem 7.11.6 (Risk-unfairness bound for any τ)**.** *Recall the definition of* $\delta_n(p, K, t)$

$$\delta_n(p, K, t) = 8\left(\frac{p}{n} + \frac{K}{n}\right) + 16\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{K}{n}}\right)\sqrt{\frac{t}{n}} + \frac{32t}{n} \ .$$

On the event

$$\mathcal{A} = \left\{ \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_2^2 + \sum_{s=1}^{K} w_s (b_s^* - \widehat{b}_s)^2 \leq \sigma^2 \delta_n(p, K, t) \right\} \ ,$$

it holds that

$$\begin{aligned}
\mathcal{R}(\widehat{f}_\tau) &\leq \left(\sigma \delta_n^{1/2}(p, K, t) + (1 - \sqrt{\tau}) \mathcal{U}^{1/2}(\widehat{f}_1) \right)^2 \ . \\
\mathcal{U}(\widehat{f}_\tau) &\leq \tau \mathcal{U}(f^*) \left(1 + \mathrm{NUR}\, \delta_n^{1/2}(p, K, t) \right)^2 \ .
\end{aligned}$$

Proof. We recall that Lemma 7.6.3 gives, for any $\tau \in [0, 1]$,

$$\mathcal{U}(\widehat{f}_\tau) = \tau \sum_{s=1}^{K} w_s \left(\widehat{b}_s - \sum_{s'=1}^{K} w_{s'} \widehat{b}_{s'} \right)^2 \ . \tag{7.38}$$

Let us start by proving the first part of the statement. Using Lemma 7.11.5 to upper bound the risk $\mathcal{R}(\widehat{f}_\tau)$ and the definition of $\mathcal{A}$ to control this upper bound, we obtain

$$\begin{aligned}\mathcal{R}(\widehat{f}_\tau) &\leq \sigma^2\delta_n(p,K,t) + 2\sigma\sqrt{\delta_n(p,K,t)}\left((1-\sqrt{\tau})\sqrt{\mathcal{U}(\widehat{f}_1)}\right) + (1-\sqrt{\tau})^2\mathcal{U}(\widehat{f}_1) \\ &= \left(\sigma\sqrt{\delta_n(p,K,t)} + (1-\sqrt{\tau})\sqrt{\mathcal{U}(\widehat{f}_1)}\right)^2 \ .\end{aligned}$$

The second part of the statement follow by applying Lemma 7.11.4 and Theorem 7.11.2 to get

$$\mathcal{U}(\widehat{f}_\tau) \leq \tau\mathcal{U}(f^*)\left(1 + \mathrm{NUR}\sqrt{\delta_n(p,K,t)}\right)^2 \ . \tag{7.39}$$

□

7.11.3 Proof of Theorem 7.6.4

We set $\widehat{f} := \widehat{f}_1$ and $\delta_n = \delta_n(p,K,t)$. The proof relies on Eq. (7.35) of Lemma 7.6.3. Using notations of Theorem 7.11.6 we also define the event

$$\mathcal{A} = \left\{\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_2^2 + \sum_{s=1}^K w_s(b_s^* - \widehat{b}_s)^2 \leq \sigma^2\delta_n(p,K,t)\right\} \tag{7.40}$$

which holds with probability at least $1 - 4\exp(-t)$.

Case 1. Assume that $\mathcal{U}^{1/2}(\widehat{f}) > \sigma\delta_n^{1/2}(p,K,t)$. Note that thanks to Theorem 7.11.6, and the definition of $\widehat{\tau}$ we derive on the event $\mathcal{A}$ that

$$\begin{aligned}\mathcal{U}(\widehat{f}_{\widehat{\tau}}) \leq \widehat{\tau}\mathcal{U}(f^*)\left(1+\sigma\sqrt{\frac{\delta_n}{\mathcal{U}(f^*)}}\right)^2 &= \alpha\mathcal{U}(f^*)\left(1+\sigma\sqrt{\frac{\delta_n}{\mathcal{U}(f^*)}}\right)^2\left(1+\frac{\sigma\delta_n^{1/2}}{\mathcal{U}^{1/2}(\widehat{f})-\sigma\delta_n^{1/2}}\right)^{-2} \\ &\overset{(a)}{\leq} \alpha\mathcal{U}(f^*)\left(1+\sigma\sqrt{\frac{\delta_n}{\mathcal{U}(f^*)}}\right)^2\left(1+\frac{\sigma\delta_n^{1/2}}{\mathcal{U}^{1/2}(f^*)}\right)^{-2} \\ &= \alpha\mathcal{U}(f^*) \ .\end{aligned}$$

In the last equation, inequality (a) follows from Eq. (7.35) of Lemma 7.6.3 and thanks to the fact that on the event $\mathcal{A}$ it holds that $\mathcal{U}^{1/2}(\widehat{f}) \leq \mathcal{U}^{1/2}(f^*) + \left\{\sum_{s=1}^K w_s(\widehat{b}_s - b_s^*)^2\right\}^{1/2} \leq \mathcal{U}^{1/2}(f^*) + \sigma\delta_n^{1/2}$. For the risk we have thanks to Theorems 7.11.6 that

$$\mathcal{R}(\widehat{f}_{\widehat{\tau}}) \leq \left(\sigma\sqrt{\delta_n} + (1-\sqrt{\widehat{\tau}})\sqrt{\mathcal{U}(\widehat{f})}\right)^2 \ . \tag{7.41}$$

Furthermore, we note that

$$\begin{aligned}\sqrt{\widehat{\tau}\mathcal{U}(\widehat{f})} = \sqrt{\alpha}\frac{\sqrt{\mathcal{U}(\widehat{f})}}{1+\frac{\sigma\sqrt{\delta_n}}{\sqrt{\mathcal{U}(\widehat{f})}-\sigma\sqrt{\delta_n}}} &= \sqrt{\alpha}\left(\sqrt{\mathcal{U}(\widehat{f})} - \sigma\sqrt{\delta_n}\right) \\ &\overset{(b)}{\geq} \sqrt{\alpha}\left(\sqrt{\mathcal{U}(f^*)} - 2\sigma\sqrt{\delta_n}\right) \ ,\end{aligned} \tag{7.42}$$

where inequality (b) again follows from Eq. (7.35) of Lemma 7.6.3 and thanks to the fact that on the event $\mathcal{A}$ it holds that $\mathcal{U}^{1/2}(\widehat{f}) \geq \mathcal{U}^{1/2}(f^*) - \left\{\sum_{s=1}^K w_s(\widehat{b}_s - b_s^*)^2\right\}^{1/2} \geq \mathcal{U}^{1/2}(f^*) - \sigma\delta_n^{1/2}$. Recall, that we have already shown that on the event $\mathcal{A}$ we have

$$\mathcal{U}^{1/2}(\widehat{f}) \leq \mathcal{U}^{1/2}(f^*) + \sigma\delta_n^{1/2} \ . \tag{7.43}$$

Combining Eqs. (7.42) and (7.43) we obtain

$$\begin{aligned}(1-\sqrt{\widehat{\tau}})\sqrt{\mathcal{U}(\widehat{f})} &\leq \sqrt{\mathcal{U}(f^*)} + \sigma\sqrt{\delta_n} - \sqrt{\alpha}\left(\sqrt{\mathcal{U}(f^*)} - 2\sigma\sqrt{\delta_n}\right) \\ &= (1-\sqrt{\alpha})\sqrt{\mathcal{U}(f^*)} + (1+2\sqrt{\alpha})\sigma\sqrt{\delta_n}\end{aligned}$$

Thus since the function $(\sigma\delta_n^{1/2} + \cdot)^2$ is increasing on $[-\sigma\sqrt{\delta_n}, \infty)$ we get from Eq. (7.41) that

$$\mathcal{R}(\widehat{f}_{\widehat{\tau}}) \leq \left(2(1+\sqrt{\alpha})\sigma\sqrt{\delta_n} + (1-\sqrt{\alpha})\sqrt{\mathcal{U}(f^*)}\right)^2 \ ,$$

which concludes the proof of the first case.

Case 2. if $\mathcal{U}^{1/2}(\widehat{f}) \leq \sigma\delta_n^{1/2}(p, K, t)$, then

$$\widehat{f}_0(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \widehat{\boldsymbol{\beta}}\rangle + \sum_{s=1}^K w_s \widehat{b}_s \ .$$

Furthermore, on the event $\mathcal{A}$ thanks to Theorem 7.11.6 it holds that $0 = \mathcal{U}(\widehat{f}_0) \leq \alpha\mathcal{U}(f^*)$ and

$$\begin{aligned}\mathcal{R}(\widehat{f}_0) \leq \left(\sigma\delta_n^{1/2} + \mathcal{U}^{1/2}(\widehat{f})\right)^2 &= \left(\sigma\delta_n^{1/2} + \sqrt{\alpha}\mathcal{U}^{1/2}(\widehat{f}) + (1-\sqrt{\alpha})\mathcal{U}^{1/2}(\widehat{f})\right)^2 \\ &\leq \left((1+\sqrt{\alpha})\sigma\delta_n^{1/2} + (1-\sqrt{\alpha})\mathcal{U}^{1/2}(\widehat{f})\right)^2 \\ &\leq \left((1+\sqrt{\alpha})\sigma\delta_n^{1/2} + (1-\sqrt{\alpha})\left(\mathcal{U}^{1/2}(f^*) + \sigma\delta_n^{1/2}\right)\right)^2 \\ &= \left(2\sigma\delta_n^{1/2} + (1-\sqrt{\alpha})\mathcal{U}^{1/2}(f^*)\right)^2 \ .\end{aligned}$$

The proof is concluded by application of Theorem 7.11.2 to control the probability of event $\mathcal{A}$.

7.11.4 Auxiliary results for Theorem 7.6.5

Let us first present auxiliary results used for the proof of Theorem 7.6.5. The next lemma is known as Varshamov-Gilbert Lemma (Varshamov 1957; Gilbert 1952), its statement is taken from Rigollet and Hütter 2015, Lemma 4.12, see also Tsybakov 2009, Lemma 2.9.

Lemma 7.11.7. *Let $d \geq 1$ be an integer. There exist binary vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_M \in \{0,1\}^d$ such that*

1. $\rho(\boldsymbol{\omega}_j, \boldsymbol{\omega}_{j'}) \geq d/4$ *for all* $j \neq j'$,
2. $M = \lfloor e^{d/16} \rfloor \geq e^{d/32}$,

where $\rho(\cdot,\cdot)$ is the Hamming's distance on binary vectors.

The next lemmas can be found in Bellec 2017, Lemma 5.1, see also Kerkyacharian et al. 2014, Lemma 3.

Lemma 7.11.8. *Let $(\Omega, \mathcal{A})$ be a measurable space and $M \geq 1$. Let $A_0, \dots, A_M$ be disjoint measurable events. Assume that $\mathbf{Q}_0, \dots, \mathbf{Q}_M$ are probability measures on $(\Omega, \mathcal{A})$ such that*

$$\frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbf{Q}_j, \mathbf{Q}_0) \leq \kappa < \infty \ .$$

Then,

$$\max_{j=0,\dots,M} \mathbf{Q}_j(A_j^c) \geq \frac{1}{12}\min(1, M\exp(-3\kappa)) \ .$$

Define the diagonal matrix $\mathbf{W} = \mathrm{diag}(w_1, \dots, w_K)$.

Lemma 7.11.9. *Let $n \geq 1$ be an integer and $s > 0$ be a positive number. Let $M \geq 1$ and $(\boldsymbol{\beta}_j, \boldsymbol{b}_j) \in \mathbb{R}^p \times \mathbb{R}^K$, $j = 0, \dots, M$, such that $\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k)\|_2^2 + \|\mathbf{W}^{1/2}(\boldsymbol{b}_j - \boldsymbol{b}_k)\|_2^2 \geq 4s$ for $j \neq k$. Assume that*

$$\frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbf{P}_{(\boldsymbol{\beta}_j, \boldsymbol{b}_j)}, \mathbf{P}_{(\boldsymbol{\beta}_0, \boldsymbol{b}_0)}) \leq \kappa < \infty.$$

Then, for any estimator $\widehat{f}$,

$$\max_{j=0,\dots,M} \mathbf{P}_{(\boldsymbol{\beta}_j, \boldsymbol{b}_j)}\left(\mathcal{R}(\widehat{f}) \geq s\right) \geq \frac{1}{12}\min\left(1, M\exp(-3\kappa)\right) \ .$$

Proof. Denote by A_j the event $\mathcal{R}_j(\widehat{f}) < s$ for $j = 1, \dots, M$. Note that the events $A_0, \dots, A_M$ are pair-wise disjoint. Indeed, if they were not there would exist indices j and j', with $j \neq j'$, such that, on the non-empty event $A_j \cap A_{j'}$,

$$\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'})\|_2^2 + \|\mathbf{W}^{1/2}(\boldsymbol{b}_j - \boldsymbol{b}_{j'})\|_2^2 \leq 2\mathcal{R}_j(\widehat{f}) + 2\mathcal{R}_{j'}(\widehat{f}) < 4s \tag{7.44}$$

contradicting our assumption on the $(\boldsymbol{\beta}_j, \boldsymbol{b}_j)$ and $(\boldsymbol{\beta}_{j'}, \boldsymbol{b}_{j'})$. We conclude applying Lemma 7.11.8. □

7.11.5 Proof of Theorem 7.6.5

Define the $(p+K) \times (p+K)$ matrix

$$\boldsymbol{\Psi} = \left[\begin{array}{c|c} \boldsymbol{\Sigma} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{W} \end{array}\right] \ , \tag{7.45}$$

Apply Lemma 7.11.7 to obtain $\boldsymbol{\omega}_0, \dots, \boldsymbol{\omega}_M$ with $M + 1 \geq e^{(p+K)/32}$ and such that $\rho(\boldsymbol{\omega}_j, \boldsymbol{\omega}_k) \geq (p+K)/4$. Let $\boldsymbol{B}_0 = (\boldsymbol{\beta}_0, \boldsymbol{b}_0), \dots, \boldsymbol{B}_M = (\boldsymbol{\beta}_M, \boldsymbol{b}_M)$ be such that

$$\boldsymbol{B}_j = \varphi\sqrt{\frac{\sigma^2}{n}}\left(1 + \sqrt{t/(p+K)}\right)\boldsymbol{\Psi}^{-1/2}\boldsymbol{\omega}_j, \tag{7.46}$$

with $p + K \leq 32 \log(M)$ and $\varphi > 0$ to be determined later.

On the one hand we have

$$\begin{aligned}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k)\|_2^2 + \|\mathbf{W}^{1/2}(\boldsymbol{b}_j - \boldsymbol{b}_k)\|_2^2 &= \frac{\varphi^2\sigma^2}{n}(1 + \sqrt{t/(p+K)})^2 \rho(\boldsymbol{\omega}_j, \boldsymbol{\omega}_{j'}) \\ &\geq \frac{\varphi^2\sigma^2}{n}(1 + \sqrt{t/(p+K)})^2 (p+K)/4 \\ &= \frac{\varphi^2\sigma^2}{4n}(\sqrt{p+K} + \sqrt{t})^2 \ .\end{aligned}$$

On the other hand, recall that $\mathbb{P}_{Y|\boldsymbol{X},S=s} = \mathcal{N}(\langle \boldsymbol{X}, \boldsymbol{\beta}\rangle + b_s, \sigma^2)$ and $\mathbb{P}_{\boldsymbol{X}} = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, then, for a given $(\boldsymbol{\beta}, \boldsymbol{b}) \in \mathbb{R}^p \times \mathbb{R}^K$ the joint distribution of observations is

$$\mathbf{P}_{(\boldsymbol{\beta},\boldsymbol{b})} = \bigotimes_{s=1}^{K} \Big(\mathcal{N}(\langle \boldsymbol{X}, \boldsymbol{\beta}\rangle + b_s, \sigma^2) \otimes \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})\Big)^{\otimes n_s} \ .$$

Given $\boldsymbol{B} = (\boldsymbol{\beta}, \boldsymbol{b}), \boldsymbol{B}' = (\boldsymbol{\beta}', \boldsymbol{b}')$ in $\mathbb{R}^p \times \mathbb{R}^K$ we can write

$$\begin{aligned}\mathrm{KL}\left(\mathbf{P}_{(\boldsymbol{\beta},\boldsymbol{b})}, \mathbf{P}_{(\boldsymbol{\beta}',\boldsymbol{b}')}\right) &= \sum_{s=1}^{K} n_s \mathbb{E}_{\boldsymbol{X}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}\left[\widetilde{\mathrm{KL}}\left(\mathcal{N}(\langle \boldsymbol{X}, \boldsymbol{\beta}\rangle + b_s, \sigma^2), \mathcal{N}(\langle \boldsymbol{X}, \boldsymbol{\beta}'\rangle + b'_s, \sigma^2)\right)\right] \\ &= \sum_{s=1}^{K} n_s \mathbb{E}_{\boldsymbol{X}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}\left(\frac{(\langle \boldsymbol{X}, \boldsymbol{\beta} - \boldsymbol{\beta}'\rangle + b_s - b'_s)^2}{2\sigma^2}\right) \\ &= \sum_{s=1}^{K} n_s \left(\frac{\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}')\|_2^2}{2\sigma^2} + \frac{(b_s - b'_s)^2}{2\sigma^2}\right) \\ &= n\left(\frac{\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}')\|_2^2}{2\sigma^2} + \frac{\|\mathbf{W}^{1/2}(\boldsymbol{b} - \boldsymbol{b}')\|_2^2}{2\sigma^2}\right) \\ &= \frac{n}{2\sigma^2}\|\boldsymbol{\Psi}^{1/2}(\boldsymbol{B} - \boldsymbol{B}')\|_2^2 \\ &\leq \frac{\varphi^2}{2}(\sqrt{p+K} + \sqrt{t})^2 \leq \varphi^2(p+K) + \varphi^2 t \leq 32\varphi^2 \log(M) + \varphi^2 t \ .\end{aligned}$$

Let $\widehat{f}$ be any estimator and define the risks

$$\mathcal{R}_j(\widehat{f}) = \sum_{s=1}^{K} w_s \mathbb{E}\left[(\widehat{f}(\boldsymbol{X}, S) - \langle \boldsymbol{X}, \boldsymbol{\beta}_j\rangle - (b_j)_S)^2 \mid S = s\right], \quad j = 1, \ldots, M \ .$$

Set $u_n(p, K, t, \varphi, \sigma) = \frac{\varphi^2\sigma^2}{16n}(\sqrt{p+K} + \sqrt{t})^2$. Applying Lemma 7.11.9 after reducing the supremum to a finite number of hypothesis, we get for all estimators $\widehat{f}$ that

$$\begin{aligned}\sup_{(\boldsymbol{\beta}^*,\boldsymbol{b}^*)\in\mathbb{R}^p\times\mathbb{R}^K} \mathbf{P}_{(\boldsymbol{\beta}^*,\boldsymbol{b}^*)}\left(\mathcal{R}(\widehat{f}) \geq u_n(p, K, t, \varphi, \sigma)\right) &\geq \max_{j=0,\ldots,M} \mathbf{P}_{(\boldsymbol{\beta}_j,\boldsymbol{b}_j)}\left(\mathcal{R}_j(\widehat{f}) \geq u_n(p, K, t, \varphi, \sigma)\right) \\ &\geq \frac{1}{12}\min\left(1, M\exp\left(-96\varphi^2\log(M) - 3\varphi^2 t\right)\right)\end{aligned}$$

Setting $\varphi = 1/\sqrt{96}$, we obtain

$$\sup_{(\boldsymbol{\beta}^*,\boldsymbol{b}^*)\in\mathbb{R}^p\times\mathbb{R}^K} \mathbf{P}_{(\boldsymbol{\beta}^*,\boldsymbol{b}^*)}\left(\mathcal{R}(\widehat{f}) \geq \frac{\sigma^2}{1536n}\left(\sqrt{p+K}+\sqrt{t}\right)^2\right) \geq \frac{1}{12}\exp\left(-\frac{t}{32}\right) \ .$$

The proof is concluded.

7.12 Relation between $\mathcal{U}_{\mathrm{KS}}$ and $\mathcal{U}$

Lemma 7.12.1. *Let μ, ν be two univariate measures such that μ admits a density* w.r.t. *the Lebesgue measure bounded by C_μ, then*

$$\mathrm{KS}(\mu,\nu) \leq 2\sqrt{C_\mu W_1(\mu,\nu)} \ .$$

Proposition 7.12.2. *Fix some measurable $f : \mathbb{R}^p \times [K] \to \mathbb{R}$. Assume that $a_s = f(\cdot, s)\#\mu_s \in \mathcal{P}_2(\mathbb{R})$ and it admits density bounded by $C_{f,s}$ for all $s \in [K]$, then*

$$\mathcal{U}_{\mathrm{KS}}(f) \leq \|1/\boldsymbol{w}\|_\infty \sqrt{8\bar{C}_f}\cdot \mathcal{U}^{1/4}(f) \ ,$$

where $\bar{C}_f = \sum_{s=1}^K w_s C_{f,s}$.

Proof. We set $a_s = \mathrm{Law}(f(\boldsymbol{X}, S) \mid S{=}s)$ and $a = \sum_{s=1}^K w_s a_s$. Therefore, thanks to assumption of the proposition and Lemma 7.12.1 we can write

$$\mathcal{U}_{\mathrm{KS}}(f) := \sum_{s=1}^K \mathrm{KS}(a_s, a) \leq \|1/\boldsymbol{w}\|_\infty \sum_{s=1}^K w_s \,\mathrm{KS}(a_s, a) \leq 2\|1/\boldsymbol{w}\|_\infty \sum_{s=1}^K w_s C_{f,s}^{1/2} W_1^{1/2}(a_s, a) \ .$$

Furthermore we can write for any measure $\nu \in \mathcal{P}_2(\mathbb{R})$ that

$$\begin{aligned}
\mathcal{U}_{\mathrm{KS}}(f) &\overset{(a)}{\leq} 2\|1/\boldsymbol{w}\|_\infty \sum_{s=1}^K w_s C_{f,s}^{1/2}\left\{\sum_{s'=1}^K w_{s'} W_1(a_s, a_{s'})\right\}^{1/2} \\
&\overset{(b)}{\leq} 2\|1/\boldsymbol{w}\|_\infty \sum_{s=1}^K w_s C_{f,s}^{1/2}\left\{W_1(a_s,\nu) + \sum_{s'=1}^K w_{s'} W_1(a_{s'},\nu)\right\}^{1/2} \ .
\end{aligned}$$

In the above inequalities (a) follows from the convexity of $W_1(a_s, \cdot)$ see *e.g.,* Bobkov and Ledoux 2019, Section 4.1 and (b) uses the triangle inequality. Applying the Cauchy–Schwarz inequality we obtain

$$\begin{aligned}
\mathcal{U}_{\mathrm{KS}}(f) &\leq 2\|1/\boldsymbol{w}\|_\infty \left\{\sum_{s=1}^K w_s C_{f,s}\right\}^{1/2}\left\{\sum_{s=1}^K w_s\left(W_1(a_s,\nu) + \sum_{s'=1}^K w_{s'} W_1(a_{s'},\nu)\right)\right\}^{1/2} \\
&= 2^{3/2}\|1/\boldsymbol{w}\|_\infty \left\{\sum_{s=1}^K w_s C_{f,s}\right\}^{1/2}\left\{\sum_{s=1}^K w_s W_1(a_s,\nu)\right\}^{1/2} \\
&\overset{(c)}{\leq} 2^{3/2}\|1/\boldsymbol{w}\|_\infty \left\{\sum_{s=1}^K w_s C_{f,s}\right\}^{1/2}\left\{\sum_{s=1}^K w_s W_1^2(a_s,\nu)\right\}^{1/4} \ ,
\end{aligned}$$

where (c) uses the Cauchy–Schwarz inequality one more time. Finally, setting ν as the Wasserstein-2 barycenter of $a_1, \ldots, a_K$ and using the fact that $W_1(\mu, \nu) \leq W_2(\mu, \nu)$ we deduce that

$$\mathcal{U}_{\mathrm{KS}}(f) \leq 2^{3/2} \|1/\boldsymbol{w}\|_\infty \left\{ \sum_{s=1}^{K} w_s C_{f,s} \right\}^{1/2} \mathcal{U}^{1/4}(f) \ .$$

The proof is concluded. □

CHAPTER 8

An example of prediction which complies with Demographic Parity and equalizes group-wise risks in the context of regression

Let $(\boldsymbol{X}, S, Y) \in \mathbb{R}^p \times \{1, 2\} \times \mathbb{R}$ be a triplet following some joint distribution $\mathbb{P}$ with feature vector $\boldsymbol{X}$, sensitive attribute S , and target variable Y. The Bayes optimal prediction f^* which does not produce Disparate Treatment is defined as $f^*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$. We provide a non-trivial example of a prediction $\boldsymbol{x} \to f(\boldsymbol{x})$ which satisfies two common group-fairness notions: Demographic Parity and Equal Group-Wise Risks

$$(f(\boldsymbol{X}) \mid S = 1) \stackrel{d}{=} (f(\boldsymbol{X}) \mid S = 2) \ ,$$
$$\mathbb{E}[(f^*(\boldsymbol{X}) - f(\boldsymbol{X}))^2 \mid S = 1] = \mathbb{E}[(f^*(\boldsymbol{X}) - f(\boldsymbol{X}))^2 \mid S = 2] \ .$$

To the best of our knowledge this is the first explicit construction of a non-constant predictor satisfying the above. We discuss several implications of this result on better understanding of mathematical notions of algorithmic fairness.

Based on Evgenii Chzhen and Nicolas Schreuder (2020b). "An example of prediction which complies with Demographic Parity and equalizes group-wise risks in the context of regression". In: *arXiv preprint arXiv:2011.07158*.

Contents

8.1 Introduction

Designing methods that satisfy group-fairness requirements has received a lot of theoretical and empirical attention in recent years (Barocas, Hardt, and Narayanan 2019; Calmon et al. 2017; Chierichetti et al. 2017; Donini et al. 2018; Dwork et al. 2018; Hardt, Price, and Srebro 2016; Dwork et al. 2012; Kilbertus et al. 2017; Lum and Johndrow 2016; Zafar et al. 2017; Zemel et al. 2013; Agarwal, Dudík, and Wu 2019; Lipton, McAuley, and Chouldechova 2018; Chiappa et al. 2020; Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c). Most of the contributions in this direction are concerned with the problem of binary classification, while the regression setup receiving much less attention to this date (Agarwal, Dudík, and Wu 2019). However, even if the underlying problem at hand has a structure of binary classification, a continuous regression-type output might be more informative in real-world scenarios.

In the literature on algorithmic fairness, it is a standard practice to consider two distinct types of predictions: *fairness through awareness* (Dwork et al. 2012) and *fairness through unawareness (without Disparate Treatment)* (Gajane and Pechenizkiy 2017; Lipton, McAuley, and Chouldechova 2018). The former type of prediction allows one to build separate model for each sensitive attribute, while the latter obliges one to fix a single model which is later applied across all groups. In the infinite sample regime, assuming that the joint distribution of the observations is known, recent works showed that the problem of regression with fairness through awareness under the *Demographic Parity* constraint shares a strong connection with the problem of Wasserstein barycenters (Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c). In particular, Le Gouic, Loubes, and Rigollet (2020) derives a closed form expression of fair optimal prediction in the sense of Demographic Parity. However, very little is known about the predictions which avoid Disparate Treatment and achieve Demographic Parity even in the infinite sample regime. Actually, even the existence of non-trivial regression prediction strategies satisfying the two constraints is unclear.

In this work we make progress towards the mathematical understanding of the latter problem. We make the following contributions: we propose a large family of prediction functions which achieve Demographic Parity without producing Disparate Treatment; we identify a specific function within this class which additionally equalizes the group-wise risks. Even though the proposed prediction rule achieves several desirable formal group-fairness notions, we argue that this prediction is not suitable for real-world scenarios. In contrast, a prediction that is allowed to produce Disparate Treatment can alleviate these drawbacks. In the context of binary classification, similar conclusions were reached by Lipton, McAuley, and Chouldechova 2018.

Organization The rest of this chapter is organised as follows. We present in Section 8.2 our setup and general goal. In Section 8.3 we provide a description of a family of prediction rules satisfying the fairness constraints of interest. Finally in Section 8.4 we discuss a critical flaw of those prediction rules from individual level fairness viewpoint and provide some open questions. Proofs can be found in Section 8.7.

Notation For a distribution μ defined on a measurable space $(X, \mathscr{X})$ and a measurable map $T : X \mapsto Y$, where Y is another space endowed with a σ-algebra $\mathscr{Y}$, we denote by $T\sharp\mu$ the push-forward measure defined by $(T\sharp\mu)(A) = \mu(T^{-1}(A))$ for all $A \in \mathscr{Y}$. For two

random variables U, V we write $U \overset{\mathrm{d}}{=} V$ to denote their equality in distribution. The standard Euclidean inner product and Euclidean norm in $\mathbb{R}^p$ are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_2$ respectively.

8.2 Setup and general goal

Let $(\boldsymbol{X}, S, Y) \in \mathbb{R}^p \times \{1, 2\} \times \mathbb{R}$ be a triplet following some joint distribution $\mathbb{P}$ where $\boldsymbol{X}$ is a feature vector, S a binary sensitive attribute (*e.g.*, gender or race) and Y is a target variable. For $s \in \{1, 2\}$, let $\mu_{\boldsymbol{X}|s}$ denote the distribution of the features inside the group $S = s$. We are interested in finding a mapping between the feature vector and the target variable which is fair in a sense we specify in this section.

The first notion of fairness that we consider restricts the class of predictors to those which do not take as input the sensitive attribute S.

Definition 8.2.1 (Disparate Treatment)**.** *Any measurable function $f : \mathbb{R}^p \to \mathbb{R}$ that cannot receive the sensitive attribute S in its functional form does not produce* Disparate Treatment*.*

We note that Gajane and Pechenizkiy (2017) refer to the latter as fairness through unawareness. This property might be desirable for obvious legal and/or privacy reasons (Primus 2003; Barocas and Selbst 2016; Gajane and Pechenizkiy 2017). However it does not guarantee the prediction to be *statistically independent* from the sensitive attribute S because of correlations between the sensitive attribute S and the feature vector $\boldsymbol{X}$. Indeed, consider the Bayes optimal predictor $\boldsymbol{x} \mapsto f^*(\boldsymbol{x})$ defined as

$$f^*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] \ .$$

It does not take as input the sensitive attribute and achieves the lowest possible squared risk among predictions avoiding Disparate Treatment. Yet, the predictor f^* might still promote disparity between sensitive groups if the distributions of features $\boldsymbol{X}$ differ between groups.

To address the above shortcoming, we further restrict the space of possible predictions to those satisfying Demographic Parity (DP) (Calders, Kamiran, and Pechenizkiy 2009; Calders et al. 2013).

Definition 8.2.2 (Demographic Parity)**.** *A predictor $f : \mathbb{R}^p \to \mathbb{R}$ achieves* Demographic Parity *if*

$$(f(\boldsymbol{X}) \mid S = 1) \overset{d}{=} (f(\boldsymbol{X}) \mid S = 2) \ .$$

Such predictors are also said to avoid *Disparate Impact*. This notion of fairness is quite intuitive since it asks the group-wise distributions of the predictions to be the same across all groups. However this probabilistic constraint is not particularly nice to handle and describing explicitly all the functions satisfying this constraint is not an easy task. Obviously, any constant function satisfies this constraint; but what about functions depending on the feature vector $\boldsymbol{X}$? It is not obvious that one can design a non-trivial function f which does not depend on the sensitive attribute S while achieving Demographic Parity. We give two simple scenarios for which we can explicit the class of functions satisfying Demographic Parity.

Example 8.2.3 (Simple case 1)**.** *Assume that distributions of the features* $\boldsymbol{X}$ *is the same within each group,* i.e.,

$$(\boldsymbol{X} \mid S = 1) \stackrel{d}{=} (\boldsymbol{X} \mid S = 2) \ .$$

In this case any *function* $f : \mathbb{R}^p \to \mathbb{R}$ *achieves Demographic Parity. In particular, one can use the Bayes optimal prediction* f^**.*

Example 8.2.4 (Simple case 2)**.** *Assume that the sensitive attribute* S *is a deterministic function of the features* $\boldsymbol{X}$ *and that the supports of* $\mu_{\boldsymbol{X}|1}$ *and* $\mu_{\boldsymbol{X}|2}$ *are non-intersecting. Then we can construct two different functions* $f_1, f_2 : \mathbb{R}^p \to \mathbb{R}$ *such that*

$$(f_1(\boldsymbol{X}) \mid S = 1) \stackrel{d}{=} (f_2(\boldsymbol{X}) \mid S = 2) \ .$$

Then we define $f(\boldsymbol{x}) := f_1(\boldsymbol{x})$ *for all* $\boldsymbol{x}$ *in the support of* $\mu_{\boldsymbol{X}|1}$ *and* $f(\boldsymbol{x}) := f_2(\boldsymbol{x})$ *for all* $\boldsymbol{x}$ *in the support of* $\mu_{\boldsymbol{X}|2}$*. In this way Demographic Parity is achieved by* one *function* f*, which avoids Disparate Impact. In other words, when the sensitive attribute* S *is a deterministic function of features* $\boldsymbol{X}$*, using* S *or not using* S *in the functional form of the prediction does not change anything.*

Those two toy examples enable us to get a better understanding of Demographic Parity; however they are far from sufficient since assuming that the features are distributed the same across groups or that the sensitive attribute is a deterministic function of $\boldsymbol{X}$ is clearly unrealistic in practice. Thus, we would like to be able to cover more scenarios than those listed above. More formally, the main question that we would like to address is:

Main question: Is there a non trivial prediction strategy f which
1) avoids Disparate Treatment;
2) achieves Demographic Parity;
under minimal assumptions on the distribution of $(\boldsymbol{X}, S, Y)$?

Let us emphasize that the main mathematical challenge of this question comes from the fact that the sensitive attribute cannot be used in the functional form of the prediction while the prediction must satisfy a constraint depending on the sensitive attribute. We elaborate more on this issue in the next example.

Example 8.2.5 (Gaussian features)**.** *Assume that the feature vector* $\boldsymbol{X} \mid S$ *is distributed as*

$$(\boldsymbol{X} \mid S = 1) \sim \mathcal{N}(\boldsymbol{m}_1, \mathbf{I}), \qquad (\boldsymbol{X} \mid S = 2) \sim \mathcal{N}(\boldsymbol{m}_2, 2\mathbf{I}) \ ,$$

with $\boldsymbol{m}_1 \neq \boldsymbol{m}_2$*. It is very easy to find a function* $g : \mathbb{R}^p \times \{1, 2\} \to \mathbb{R}$ *so that*

$$(g(\boldsymbol{X}, S) \mid S = 1) \stackrel{d}{=} (g(\boldsymbol{X}, S) \mid S = 2) \ .$$

In particular, one can consider group-wise affine predictions: $g(\boldsymbol{x}, 1) = \langle \boldsymbol{\beta}_1, \boldsymbol{x} \rangle + b_1$ *and* $g(\boldsymbol{x}, 2) = \langle \boldsymbol{\beta}_2, \boldsymbol{x} \rangle + b_2$ *with* $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$ *satisfying*

$$\langle \boldsymbol{\beta}_1, \boldsymbol{m}_1 \rangle + b_1 = \langle \boldsymbol{\beta}_2, \boldsymbol{m}_2 \rangle + b_2, \qquad \|\boldsymbol{\beta}_1\|_2 = \sqrt{2}\|\boldsymbol{\beta}_2\|_2 \ .$$

Moreover, the risk-optimal choice of g is also group-wise affine (we elaborate on it later in the text). However, there is no non-trivial (i.e., *nonconstant) affine function $f : \mathbb{R}^p \to \mathbb{R}$ which achieves Demographic Parity and avoids Disparate Treatment.*

Indeed, assume that there exist $\boldsymbol{\beta} \in \mathbb{R}^p, b \in \mathbb{R}$ such that $f(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + b$ achieves Demographic Parity. Since the features are group-wise Gaussians, then

$$(f(\boldsymbol{X}) \mid S = 1) \sim \mathcal{N}(\langle \boldsymbol{\beta}, \boldsymbol{m}_1 \rangle + b, \|\boldsymbol{\beta}\|_2^2), \qquad (f(\boldsymbol{X}) \mid S = 2) \sim \mathcal{N}(\langle \boldsymbol{\beta}, \boldsymbol{m}_2 \rangle + b, 2\|\boldsymbol{\beta}\|_2^2) \ .$$

For the above two distributions to be equal we must set $\boldsymbol{\beta} = \mathbf{0}$ and the prediction $f \equiv b$ reduces to a trivial constant.

Example 8.2.5 highlights the intrinsic difficulty of the considered question – even if the distribution of the covariates is group-wise Gaussian and even if the Bayes optimal prediction $f^*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$ is affine, there is *no* non-trivial affine prediction rule $f(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + b$ achieving Demographic Parity *and* avoiding Disparate Treatment. Besides, this example demonstrates that learning prediction function without Disparate Impact and Disparate Treatment in an agnostic learning manner might be a bad idea. Indeed, assume the same model as in Example 8.2.5 and define $f^{\mathrm{DP}}_{\mathrm{affine}}$ as a solution of

$$\min_{f:\mathbb{R}^p \to \mathbb{R}} \left\{ \mathbb{E}(Y - f(\boldsymbol{X}))^2 \ : \ (f(\boldsymbol{X}) \mid S = 1) \stackrel{\mathrm{d}}{=} (f(\boldsymbol{X}) \mid S = 2), \quad f \in \mathcal{F}_{\mathrm{affine}} \right\} \ ,$$

where $\mathcal{F}_{\mathrm{affine}} = \{f : \boldsymbol{x} \mapsto \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + b; \boldsymbol{\beta} \in \mathbb{R}^p, b \in \mathbb{R}\}$ – a recurrent prediction class restriction in the learning literature (Vapnik and Chervonenkis 1968). Following Example 8.2.5 we know that $f^{\mathrm{DP}}_{\mathrm{affine}}$ is a trivial constant prediction thus building a data-driven method which performs as well as $f^{\mathrm{DP}}_{\mathrm{affine}}$ is not that relevant. Due to these observations we believe that current fairness definitions should be first examined without the restriction of the predictors.

In this work we provide a large family of prediction functions $\mathcal{F}_{\nearrow}$, which are parametrized by non-decreasing continuous functions $Q : [0, 1] \to \mathbb{R}$. Every function $f_Q \in \mathcal{F}_{\nearrow}$ avoid Disparate Treatment and achieves Demographic Parity. Furthermore, we show that the family $\mathcal{F}_{\nearrow}$ contains a special prediction function f_{Q^*}, which achieves an additional fairness criterion. Namely, it achieves *Equality of Group-Wise Risks* defined below.

Definition 8.2.6. *A predictor $f : \mathbb{R}^p \to \mathbb{R}$ achieves the* Equality of Group-Wise Risks *(EGWR) constraint if*

$$\mathbb{E}[(f^*(\boldsymbol{X}) - f(\boldsymbol{X}))^2 \mid S = 1] = \mathbb{E}[(f^*(\boldsymbol{X}) - f(\boldsymbol{X}))^2 \mid S = 2] \ .$$

Similar notion of fairness in its relaxed formulation was considered in the context of regression by Agarwal, Dudík, and Wu 2019.

Despite three fruitful properties of formal group-fairness requirements achieved by f_{Q^*}, we argue that this function fails to satisfy basic principles of fairness and justice. The main reason for its failure is the avoidance of Disparate Treatment, which forces a prediction to "guess" the sensitive attribute of a given feature vector $\boldsymbol{x} \in \mathbb{R}^p$. Such guessing leads to undesirable predictions for individuals $\boldsymbol{x} \in \mathbb{R}^p$ with sensitive attribute $S = 1$ but who are more likely to have $S = 2$ and vice-versa.

Fairness through awareness: a reminder Before going further into the problem, let us provide a short theoretical reminder for the situation when the sensitive attribute S is allowed to be used in the functional form of the prediction, that is, the Disparate Treatment is allowed. More formally, we are interested in finding a prediction $g^* : \mathbb{R}^p \times \{1,2\} \to \mathbb{R}$ which is a solution of

$$\min_{g:\mathbb{R}^p\times\{1,2\}\to\mathbb{R}} \left\{ \mathbb{E}(Y - g(\boldsymbol{X}, S))^2 \,:\, (g(\boldsymbol{X}, S) \mid S = 1) \stackrel{\mathrm{d}}{=} (g(\boldsymbol{X}, S) \mid S = 2) \right\} \,.$$

Le Gouic, Loubes, and Rigollet (2020) and Chzhen et al. (2020c) showed that under mild additional assumptions[1] on the distribution $\mathbb{P}$, the optimal fair prediction g^* can be obtained for all $(\boldsymbol{x}, s) \in \mathbb{R}^p \times \{1,2\}$ as

$$g^*(\boldsymbol{x}, s) = \left(p_1 G_1^{-1} + p_2 G_2^{-1}\right) \circ G_s(\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, S = s]) \,, \tag{8.1}$$

where $p_s = \mathbb{P}(S = s)$, $G_s(t) = \mathbb{P}\left(\mathbb{E}[Y \mid \boldsymbol{X}, S] \le t \mid S = s\right)$, and G_s^{-1} is the generalized inverse of G_s for all $s \in \{1,2\}$. In particular, returning to Example 8.2.5, one can show that if $\mathbb{E}[Y \mid \boldsymbol{X}, S] = \langle \boldsymbol{\beta}^*_S, \boldsymbol{X}\rangle + b_S$, then the fair optimal prediction g^* is also group-wise affine. Indeed, we note that under the assumptions of Example 8.2.5 it holds that

$$(\mathbb{E}[Y \mid \boldsymbol{X}, S] \mid S = 1) \sim \mathcal{N}(b_1, \|\boldsymbol{\beta}^*_1\|_2^2), \qquad (\mathbb{E}[Y \mid \boldsymbol{X}, S] \mid S = 2) \sim \mathcal{N}(b_2, 2\|\boldsymbol{\beta}^*_2\|_2^2) \,.$$

Denoting by Φ the cumulative distribution function of the standard Gaussian we can write that $G_1(t) = \Phi(^{(t-b_1)}/_{\|\boldsymbol{\beta}^*_1\|_2})$ and $G_2(t) = \Phi(^{(t-b_2)}/_{\sqrt{2}\|\boldsymbol{\beta}^*_1\|_2})$. Their inverses can be respectively written as

$$G_1^{-1}(t) = b_1 + \|\boldsymbol{\beta}^*_1\|_2 \Phi^{-1}(t), \qquad G_2^{-1}(t) = b_2 + \sqrt{2}\|\boldsymbol{\beta}^*_2\|_2 \Phi^{-1}(t) \,.$$

Substituting these expressions into Eq. (8.1) and simplifying we get that

$$g^*(\boldsymbol{x}, 1) = \langle \boldsymbol{\beta}^*_1, \boldsymbol{x}\rangle \left(p_1 + p_2 \frac{\sqrt{2}\|\boldsymbol{\beta}^*_2\|_2}{\|\boldsymbol{\beta}^*_1\|_2}\right) + p_1 b_1 + p_2 b_2 \,,$$
$$g^*(\boldsymbol{x}, 2) = \langle \boldsymbol{\beta}^*_2, \boldsymbol{x}\rangle \left(p_2 + p_1 \frac{\|\boldsymbol{\beta}^*_1\|_2}{\sqrt{2}\|\boldsymbol{\beta}^*_2\|_2}\right) + p_1 b_1 + p_2 b_2 \,.$$

The above highlights that in the case of linear regression model, the predictor $g^* : \mathbb{R}^p \times \{1,2\} \to \mathbb{R}$ which minimizes the risk under the Demographic Parity constraint remains affine. We again emphasize that the situation is changed drastically if a prediction is not allowed to produce Disparate Treatment.

8.3 Description of the family

In this section we present a family of prediction rules, indexed by the set of continuous non-decreasing functions $Q : [0,1] \to \mathbb{R}$, which achieve Demographic Parity and explicit a function from this family which also satisfies the Equality of Group-Wise Risks constraint.

[1] They assume that for all $s \in \{1,2\}$ the measure $g(\cdot, s)\sharp\mu_{\boldsymbol{X}|s}$ is continuous and has finite second moment.

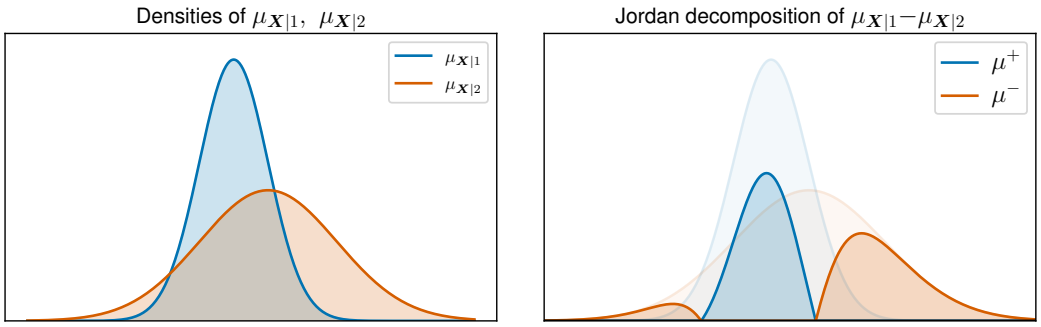


Figure 8.1: Jordan decomposition of a signed measure. (`Left`) An example of feature distributions within two groups. (`Right`) Jordan decomposition of the difference $\mu_{\boldsymbol{X}|1} - \mu_{\boldsymbol{X}|2}$.

Jordan decomposition Consider the signed measure $\mu \coloneqq \mu_{\boldsymbol{X}|1} - \mu_{\boldsymbol{X}|2}$, and let μ^+, μ^- be its Jordan decomposition, that is $\mu = \mu^+ - \mu^-$ and $\text{supp}(\mu^+) \cap \text{supp}(\mu^-) = \emptyset$. Unless the supports of $\mu_{\boldsymbol{X}|1}$ and $\mu_{\boldsymbol{X}|2}$ are disjoint, the measures μ^+ and μ^- do not integrate to one, *i.e.,* they are not probability measures. However, both μ^+ and μ^- have the same total mass. We define $P\mu^\pm = \mu^\pm / \mu^\pm(\mathbb{R})$ the projection of $\mu^\pm$ on the space of probability measures. We also define

$$F_\pm(t) := P\mu^\pm \left(\{ \boldsymbol{x} \in \mathbb{R}^p \,:\, f^*(\boldsymbol{x}) \leq t \} \right) \quad ,$$

the cumulative distribution function of $f^* \sharp P\mu^\pm$.

The supports of measures μ^+ and μ^- have a simple and intuitive interpretation. Note that if $\boldsymbol{x} \in \text{supp}(\mu^+)$, then $\boldsymbol{x}$ is more likely to be a member of the group $S = 1$ and vice versa. Meanwhile, if $\boldsymbol{x} \in \mathbb{R}^p \setminus (\text{supp}(\mu^+) \cup \text{supp}(\mu^-))$ then $\boldsymbol{x}$ can be equally likely coming from $S = 1$ or from $S = 2$. See Figure 8.1 for an illustration with univariate covariates.

The rational behind the introduction of the Jordan decomposition of μ into μ^+ and μ^- comes from the following simple insight. It says that in order to check Demographic Parity for predictions without Disparate Treatment one only needs to know μ^+ and μ^- instead of the whole distribution of the covariates $\mu_{\boldsymbol{X}|s}$. This idea is formalized in the next lemma.

Lemma 8.3.1. *A prediction without Disparate Treatment $f : \mathbb{R}^p \to \mathbb{R}$ achieves Demographic Parity iff*

$$f \sharp \mu^+ = f \sharp \mu^- \quad . \tag{8.2}$$

Proof. Set $A_\square = \text{supp}(\mu^\square)$ for $\square \in \{\pm\}$ and $A_0 = \mathbb{R}^p \setminus (A_+ \cup A_-)$.

($\Rightarrow$) If f achieves Demographic Parity, then $f \sharp \mu_{\boldsymbol{X}|1} = f \sharp \mu_{\boldsymbol{X}|2}$. Note that, for all $t \in \mathbb{R}$ it holds that

$$\begin{aligned} \mu_{\boldsymbol{X}|\square} \{ \boldsymbol{x} \in \mathbb{R}^p \,:\, f(\boldsymbol{x}) \leq t \} = \mu_{\boldsymbol{X}|\square} \{ \boldsymbol{x} \in A_+ \,:\, f(\boldsymbol{x}) \leq t \} + \mu_{\boldsymbol{X}|\square} \{ \boldsymbol{x} \in A_- \,:\, f(\boldsymbol{x}) \leq t \} \\ + \mu_{\boldsymbol{X}|\square} \{ \boldsymbol{x} \in A_0 \,:\, f(\boldsymbol{x}) \leq t \} \quad . \end{aligned}$$

Note that by the definition of μ^+ and μ^- it holds that

$$\mu_{\boldsymbol{X}|1} \{ \boldsymbol{x} \in A_0 \,:\, f(\boldsymbol{x}) \leq t \} = \mu_{\boldsymbol{X}|2} \{ \boldsymbol{x} \in A_0 \,:\, f(\boldsymbol{x}) \leq t \} \quad ,$$

and thus, the condition $f\sharp\mu_{\boldsymbol{X}|1} = f\sharp\mu_{\boldsymbol{X}|2}$ implies that for all $t \in \mathbb{R}$

$$\mu_{\boldsymbol{X}|1}\{\boldsymbol{x} \in A_+ \,:\, f(\boldsymbol{x}) \leq t\} - \mu_{\boldsymbol{X}|2}\{\boldsymbol{x} \in A_+ \,:\, f(\boldsymbol{x}) \leq t\} = \\ \mu_{\boldsymbol{X}|2}\{\boldsymbol{x} \in A_- \,:\, f(\boldsymbol{x}) \leq t\} - \mu_{\boldsymbol{X}|1}\{\boldsymbol{x} \in A_- \,:\, f(\boldsymbol{x}) \leq t\} \quad . \tag{8.3}$$

The latter is equivalent to Eq. (8.2).

($\Leftarrow$) Recall that Eq. (8.2) is equivalent to Eq. (8.3) and that for any $f : \mathbb{R}^p \to \mathbb{R}$ it holds that

$$\mu_{\boldsymbol{X}|1}\{\boldsymbol{x} \in A_0 \,:\, f(\boldsymbol{x}) \leq t\} = \mu_{\boldsymbol{X}|2}\{\boldsymbol{x} \in A_0 \,:\, f(\boldsymbol{x}) \leq t\} \quad .$$

Combining both concludes the proof. □

Note that such an argument would not work if f was allowed to depend on the sensitive attribute.

Prediction rules Let $Q : [0,1] \to \mathbb{R}$ be any continuous non-decreasing function. Define the following prediction rule

$$f_Q(\boldsymbol{x}) = \begin{cases} Q \circ F_+ \circ f^*(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in \operatorname{supp}(\mu^+) \\ Q \circ F_- \circ f^*(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in \operatorname{supp}(\mu^-) \\ f^*(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in \mathbb{R}^p \setminus (\operatorname{supp}(\mu^+) \cup \operatorname{supp}(\mu^-)) \end{cases} \quad . \tag{$*$}$$

One should think of Q as a quantile function of some continuous univariate probability measure λ. In the language of optimal transport the function $Q \circ F_\square \circ f^*(\cdot)$ is the optimal transport map from $f^*\sharp P\mu^\square$ to λ. An exact theoretical motivation to introduce function Q will be clarified later in the text.

There are three cases in the above prediction rule:

1. $\boldsymbol{x} \in \operatorname{supp}(\mu^+)$, in this case $\boldsymbol{x}$ is more likely associated with $S = 1$.
2. $\boldsymbol{x} \in \operatorname{supp}(\mu^-)$, in this case $\boldsymbol{x}$ is more likely associated with $S = 2$.
3. $\boldsymbol{x} \in \mathbb{R}^p \setminus (\operatorname{supp}(\mu^+) \cup \operatorname{supp}(\mu^-))$, in this case $\boldsymbol{x}$ can be equally likely associated with group $S = 1$ and $S = 2$ and the decision is made in accordance with the Bayes optimal prediction by the analogy with Example 8.2.3.

Fairness of prediction rules Note that since the prediction rules are not allowed to depend on the sensitive attribute in its functional form, they do not produce Disparate Treatment. In order to show that the prediction rules defined in ($*$) satisfy other fairness constraints, we make one standard technical assumption about particular distributions induced by the Bayes rule.

Assumption 8.3.2. *The measures $f^*\sharp\mu^+$, $f^*\sharp\mu^-$ are non-atomic with finite second moments.*

The following proposition states that, under the previous assumption, the defined prediction rules achieve Demographic Parity.

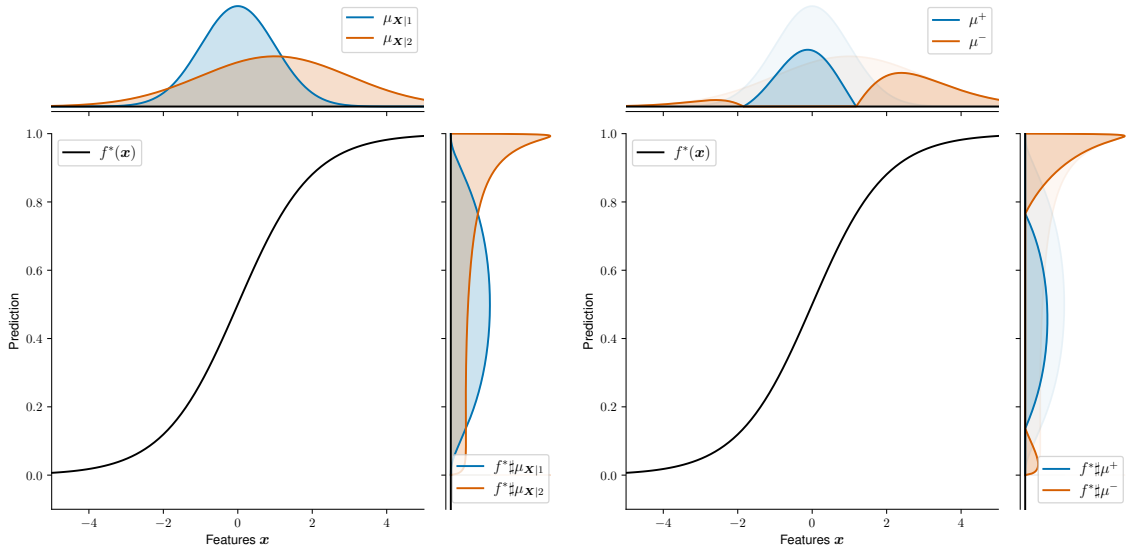


Figure 8.2: (`Left`) The Bayes optimal prediction f^* is illustrated in the center. The group-wise distributions of features is illustrated on the top. The group-wise distribution of the Bayes optimal prediction is illustrated on the right. (`Right`) Jordan decomposition of $\mu_{\boldsymbol{X}|1} - \mu_{\boldsymbol{X}|2}$ is illustrated on the top and their push-forward (through f^*) measures are on the right.

Proposition 8.3.3. *Let Assumption 8.3.2 hold. Let $Q : [0,1] \to \mathbb{R}$ be any continuous non-decreasing function, then the prediction rule f_Q is fair in the sense of Demographic Parity.*

The proof of Proposition 8.3.3 is postponed to Section 8.7. The result becomes rather intuitive following the interpretation of Q as a quantile function of some continuous univariate probability measure λ and of $Q \circ F_{\square} \circ f^*(\cdot)$ as the optimal transport map from $f^*\sharp P\mu^{\square}$ to λ in combination with Lemma 8.3.1.

We have a large class of prediction rules which avoid Disparate Treatment and achieves Demographic Parity. Can we find a subset of this class such that its elements also satisfy Equality of Group-Wise Risks? The next proposition explicitly gives a continuous non-decreasing function Q^* such that the resulting prediction rule f_{Q^*} satisfies the Equality of Group-Wise Risks constraint.

Proposition 8.3.4. *Let Assumption 8.3.2 hold. For the choice $Q^* = (F_+^{-1} + F_-^{-1})/2$, the prediction rule f_{Q^*} is fair in the sense of Equality of Group-Wise Risks.*

8.4 Discussion and open questions

In the previous section we have proved that the prediction rules defined in (∗) achieve Demographic Parity and that for a specific choice of continuous non-decreasing function Q^*, the prediction rule f_{Q^*} also satisfies the Equality of Group-Wise Risks. The latter prediction rule is represented in Figure 8.3 for a particular problem: the features are assumed to be group-wise Gaussian random variables with different means and variances. We set the Bayes optimal predictor as $f^*(\boldsymbol{x}) = 1/(1 + e^{a\boldsymbol{x}})$ for some positive real $a > 0$.

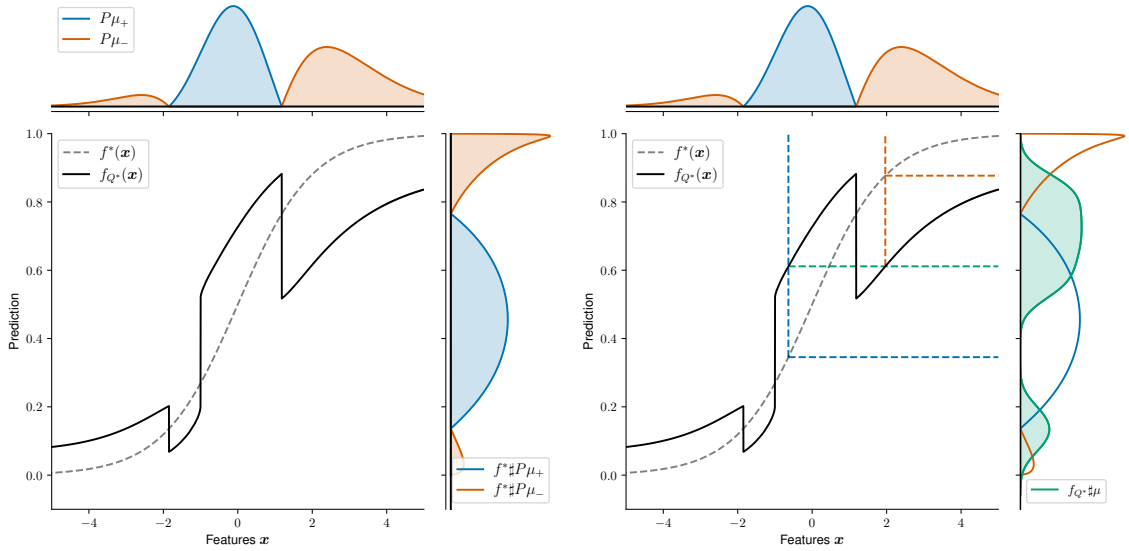


Figure 8.3: (`Left`) Prediction f_{Q^*} which achieves DP and EGWR. (`Right`) Concrete examples of predictions.

In both plots the dashed grey curve corresponds to the Bayes prediction rule $f^*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$ while the black solid curve represents the prediction rule f_{Q^*} defined in Proposition 8.3.4. On top of the plots are the densities from the (normalized) Jordan decomposition of $\mu_{\boldsymbol{X}|1} - \mu_{\boldsymbol{X}|2}$ (see also Figure 8.1) while on the right side are the densities corresponding to the predictions.

In the right plot, the green horizontal dashed line corresponds to the prediction by f_{Q^*} for the points whose axis correspond to the vertical blue and orange dashed lines. The horizontal blue and orange dashed lines correspond to the prediction by f^*. We notice that the prediction curve corresponding to f_{Q^*} looks like a piece-wise translation of the Bayes decision rule in which the predicted value is increased for features which seem to come from the group corresponding to $S = 1$ and lowered for the other features.

The prediction rule f_{Q^*} could be formally considered as a good fair predictor since it simultaneously satisfies several formal group-fairness constraints and avoids Disparate Treatment. However, Demographic Parity and EGWR only define fairness on the group level and inspecting the individual level reveals a critical flow of this prediction rule. We have constrained our predictors to those that do not produce Disparate Treatment by prohibiting them from having the sensitive variable as direct input. Nevertheless, enforcing group level fairness constraints (such as DP and EGWR) forces the prediction rule to guess the sensitive attribute corresponding to a given feature vector $\boldsymbol{x}$. The idea of our prediction rules is simple: if a feature vector $\boldsymbol{x}$ is more likely to belong to some group then it is treated as a member of this group. A critical resulting issue of this is that an individual from the minority (*i.e.,* the group which gets discriminated) which "looks like" an individual from the majority will be treated as the latter and thus might potentially receive a negative discrimination, worsening their position in the population and in the society. This is clearly contrary to what one would expect from a fair decision-making system and should therefore be avoided. We remark that a simple remedy from the above flaw is to allow to construct a separate prediction rule for each sensitive group – wave away the Disparate Treatment requirement. Indeed, making separate

predictions for separate groups erases the effect of group guessing and allows to make a more informed decision (Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020c; Lipton, McAuley, and Chouldechova 2018).

An interesting open question concerns the optimality of the derived prediction rules: is it possible to find a prediction rule which avoids Disparate Treatment while achieving Demographic Parity and which has smaller squared risk than those of the prediction rules in (∗) ? An answer to this question would yield an important step towards understanding the limits of predictions under fairness constraint without having access to the sensitive attribute. Establishing the optimality would also allow to address relaxed notions of fairness in this context and provide a statistical study similar to Chzhen and Schreuder (2020a).

8.5 Conclusion

In this work we proposed a large family of prediction rules which simultaneously avoid Disparate Treatment and achieve Demographic Parity. In addition, we also showed that a particular member of the proposed family equalizes the group-wise risks. However, despite these fruitful formal fairness properties, none of the above predictions are able to comply with the intuitive understanding of fairness. We attribute this effect to the avoidance of Disparate Treatment. An interesting mathematical challenge which remains unsolved is connected with the risk optimality of the proposed prediction rules.

8.6 Extension to non-binary sensitive attribute

Previous parts were dealing with the construction of a DP fair prediction $f : \mathbb{R}^d \to \mathbb{R}$ in case when there are only two sensitive attributes. Here we provide an extension.

Assume that the measures $\mu_{\boldsymbol{X}|1}, \dots, \mu_{\boldsymbol{X}|K}$ admit density. For any permutation $\boldsymbol{\sigma}$ of $[K]$, any $K-1$ tuple of univariates measures with densities $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{K-1})$ and any measurable functions g with $\sup_{c\in\mathbb{R}} \mathrm{Leb}\left\{\boldsymbol{x} \in \mathbb{R}^d \,:\, g(\boldsymbol{x}) = c\right\} = 0$, set

$$\nu_s^1 = g\sharp\mu_{\boldsymbol{X}|\boldsymbol{\sigma}(s)}, \qquad \forall s \in [K] \ .$$

For any $j = 1, \dots, K-1$ set

$$\nu_s^{j+1} = T_j^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}\sharp\nu_s^j \ , \tag{8.4}$$

where $T_j^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}$ is defined as

$$T_j^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}(y) = \begin{cases} T_j^+(y) & y \in \mathrm{supp}(\pi_j^+) \\ T_j^-(y) & y \in \mathrm{supp}(\pi_j^-) \\ y & \text{otherwise} \end{cases} \ ,$$

with $\pi_j^\pm$ defined as the re-scaled Jordan decomposition of $\nu_j^j - \nu_{j+1}^j$ and $T_j^\pm$ being the optimal transport maps from $\pi_j^\pm$ to λ_j. Finally, define

$$f_{\boldsymbol{\sigma},\boldsymbol{\lambda},g} := \left(T_{K-1}^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ \dots \circ T_1^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ g\right) \ . \tag{8.5}$$

Lemma 8.6.1. *Under the assumptions of the above construction any function $f_{\boldsymbol{\sigma},\boldsymbol{\lambda},g}$ defined in Eq. (8.5) achieves Demographic Parity.*

Proof. Fix some real-valued function $g : \mathbb{R}^d \to \mathbb{R}$, some non-atomic univariate measures $\lambda_1, \dots, \lambda_{K-1}$.

The proof goes by induction on K with the base case $K = 2$. If $K = 2$, then

$$f_{\boldsymbol{\sigma},\boldsymbol{\lambda},g} := \left(T_1^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ g\right) \ .$$

Note that since ν_1^1, ν_2^1, admit density, then the re-scaled Jordan decomposition π_1^+, π_1^- of $\nu_1^1 - \nu_2^1$ also admit a density. Thus, there exist optimal transport maps T_1^+, T_1^- from π_1^+, π_1^- to λ_1, implying that $T_1^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}$ is well-defined. Moreover, by construction

$$T_1^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}\sharp\nu_1^1 - T_1^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}\sharp\nu_2^1 = T_1^+\sharp\pi_1^+ - T_1^-\sharp\pi_1^- = 0 \ ,$$

which proves the base case.

Fix some permutation $\boldsymbol{\sigma}$ of $[K+1]$. For $K+1$ by the induction assumption it holds for any $s, s' \in \{\boldsymbol{\sigma}(1), \dots, \boldsymbol{\sigma}(K)\}$ that

$$\left(T_{K-1}^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ \dots \circ T_1^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ g\right)\sharp\mu_{\boldsymbol{X}|s} = \left(T_{K-1}^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ \dots \circ T_1^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ g\right)\sharp\mu_{\boldsymbol{X}|s'} \ ,$$

implying that $\forall s, s' \in \{\boldsymbol{\sigma}(1), \dots, \boldsymbol{\sigma}(K)\}$

$$T_K^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}\sharp\left(T_{K-1}^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ \dots \circ T_1^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ g\right)\sharp\mu_{\boldsymbol{X}|s} = T_K^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}\sharp\left(T_{K-1}^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ \dots \circ T_1^{(\boldsymbol{\sigma},\boldsymbol{\lambda})} \circ g\right)\sharp\mu_{\boldsymbol{X}|s'} \ .$$

The above is equivalent to

$$f_{\boldsymbol{\sigma},\boldsymbol{\lambda},g}\sharp\mu_{\boldsymbol{X}|s} = f_{\boldsymbol{\sigma},\boldsymbol{\lambda},g}\sharp\mu_{\boldsymbol{X}|s'} \qquad \forall s, s' \in \{\boldsymbol{\sigma}(1), \dots, \boldsymbol{\sigma}(K)\} \ .$$

To complete the proof it is sufficient to show that

$$f_{\boldsymbol{\sigma},\boldsymbol{\lambda},g}\sharp\mu_{\boldsymbol{X}|\boldsymbol{\sigma}(K+1)} = f_{\boldsymbol{\sigma},\boldsymbol{\lambda},g}\sharp\mu_{\boldsymbol{X}|\boldsymbol{\sigma}(K)} \ .$$

Observing that

$$f_{\boldsymbol{\sigma},\boldsymbol{\lambda},g}\sharp\mu_{\boldsymbol{X}|\boldsymbol{\sigma}(K+1)} = T_K^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}\sharp\nu_{K+1}^K \ ,$$
$$f_{\boldsymbol{\sigma},\boldsymbol{\lambda},g}\sharp\mu_{\boldsymbol{X}|\boldsymbol{\sigma}(K+1)} = T_K^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}\sharp\nu_K^K \ ,$$

we use the fact that by construction

$$T_K^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}\sharp\nu_{K+1}^K - T_K^{(\boldsymbol{\sigma},\boldsymbol{\lambda})}\sharp\nu_K^K = T_K^+\sharp\pi_K^+ - T_K^-\sharp\pi_K^- = 0 \ .$$

The proof is concluded. □

8.7 Proofs

We recall that the *Wasserstein-2* distance between probability distributions μ and ν in $\mathcal{P}_2(\mathbb{R}^d)$, the space of measures on $\mathbb{R}^d$ with finite second moment, is defined as

$$\mathsf{W}_2^2(\mu,\nu) \coloneqq \inf_{\gamma\in\Gamma(\mu,\nu)} \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 d\gamma(\boldsymbol{x},\boldsymbol{y}) \right\} \ , \tag{8.6}$$

where $\Gamma(\mu,\nu)$ denotes the collection of measures on $\mathbb{R}^d\times\mathbb{R}^d$ with marginals μ and ν. See Santambrogio 2015; Villani 2003 for more details about Wasserstein distances and optimal transport.

Proof of Proposition 8.3.3. In order to prove that f_Q satisfies the Demographic Parity constraint we examine the following quantity

$$\Delta_{f_Q}(t) = \mu_{\boldsymbol{X}|1}\left\{\boldsymbol{x}\in\mathbb{R}^p \,:\, f_Q(\boldsymbol{x})\le t\right\} - \mu_{\boldsymbol{X}|2}\left\{\boldsymbol{x}\in\mathbb{R}^p \,:\, f_Q(\boldsymbol{x})\le t\right\} \ ,$$

for any $t\in\mathbb{R}$. Fix some $t\in\mathbb{R}$. Let us fix some $Q:[0,1]\to\mathbb{R}$ continuous non-decreasing function. For simplicity we drop the subscript Q from f_Q and write f instead. We can write by the definition of f, μ and μ^+,μ^- that

$$\begin{aligned}\Delta_f(t) &= \int_{f(\boldsymbol{x})\le t} d\mu(\boldsymbol{x}) = \int_{f(\boldsymbol{x})\le t} d\mu^+(\boldsymbol{x}) - \int_{f(\boldsymbol{x})\le t} d\mu^-(\boldsymbol{x}) \\ &= \int_{Q(\mu^+\{\boldsymbol{x}'\in\mathbb{R}^p\,:\,f^*(\boldsymbol{x}')\le f^*(\boldsymbol{x})\})\le t} d\mu^+(\boldsymbol{x}) - \int_{Q(\mu^-\{\boldsymbol{x}'\in\mathbb{R}^p\,:\,f^*(\boldsymbol{x}')\le f^*(\boldsymbol{x})\})\le t} d\mu^-(\boldsymbol{x}) \ .\end{aligned}$$

Let Q^{-1} be the generalized inverse of Q. Hence, since Q is assumed to be continuous we can write

$$\Delta_f(t) = \int_{\mu^+\{\boldsymbol{x}'\in\mathbb{R}^p\,:\,f^*(\boldsymbol{x}')\le f^*(\boldsymbol{x})\}\le Q^{-1}(t)} d\mu^+(\boldsymbol{x}) - \int_{\mu^-\{\boldsymbol{x}'\in\mathbb{R}^p\,:\,f^*(\boldsymbol{x}')\le f^*(\boldsymbol{x})\}\le Q^{-1}(t)} d\mu^-(\boldsymbol{x}) \ .$$

Introduce $F_\square(\cdot) = \mu^\square\{\boldsymbol{x}'\in\mathbb{R}^p \,:\, f^*(\boldsymbol{x}')\le\cdot\}$ for $\square\in\{\pm\}$ and note that thanks to Assumption 8.3.2 both F_+ and F_- are non-decreasing continuous. Thus,

$$\begin{aligned}\Delta_f(t) &= \int_{F_+(f^*(\boldsymbol{x}))\le Q^{-1}(t)} d\mu^+(\boldsymbol{x}) - \int_{F_-(f^*(\boldsymbol{x}))\le Q^{-1}(t)} d\mu^-(\boldsymbol{x}) \\ &= \int_{f^*(\boldsymbol{x})\le F_+^{-1}\circ Q^{-1}(t)} d\mu^+(\boldsymbol{x}) - \int_{f^*(\boldsymbol{x})\le F_-^{-1}\circ Q^{-1}(t)} d\mu^-(\boldsymbol{x}) \\ &= F_+\circ F_+^{-1}\circ Q^{-1}(t) - F_-\circ F_-^{-1}\circ Q^{-1}(t) = 0 \ .\end{aligned}$$

The proof is concluded since $\sup_{t\in\mathbb{R}}|\Delta_f(t)| = 0$ implies that f satisfies the Demographic Parity constraint. □

Proof of Proposition 8.3.4. In this proof we consider the prediction rule f_{Q^*} defined in (∗) with the specific choice $Q^* \coloneqq (F_+^{-1}+F_-^{-1})/2$. Let $p_1=\mathbb{P}(S=1)$ and $p_2=\mathbb{P}(S=2)=1-p_1$. Since Q^* is fixed in throughout this proof, we drop the subscript Q^* and write f instead of f_{Q^*} for compactness.

Recall that we defined the signed measure $\mu = \mu_{\boldsymbol{X}|1} - \mu_{\boldsymbol{X}|2}$. Using its Hahn decomposition, $\mu = \mu^+ - \mu^-$, we can write $\mu_{\boldsymbol{X}|1} = \mu^+ - \mu^- + \mu_{\boldsymbol{X}|2}$ and express the risk of the predictor f as

$$\begin{aligned}\mathcal{R}(f) &= p_1 \int (f^*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\mu_{\boldsymbol{X}|1}(\boldsymbol{x}) + p_2 \int (f^*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\mu_{\boldsymbol{X}|2}(\boldsymbol{x}) \\ &= \int (f^*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\mu_{\boldsymbol{X}|2}(\boldsymbol{x}) \\ &\quad + p_1 \left(\int (f^*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\mu^+(\boldsymbol{x}) - \int (f^*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\mu^-(\boldsymbol{x}) \right) \ . \end{aligned} \tag{8.7}$$

Since $f\sharp\mu^\square = T_\square\sharp(f^*\sharp\mu^\square)$ for $\square \in \{\pm\}$, where $T_\square = Q \circ F_\square$ is a monotone non-decreasing function, Santambrogio 2015, Theorem 2.9 implies

$$\int (f^*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\mu^\square(\boldsymbol{x}) = \mathsf{W}_2^2(f^*\sharp\mu^\square, f\sharp\mu^\square), \text{ for } \square \in \{\pm\} \ .$$

Following Agueh and Carlier 2011, Section 6.1, the solution to the Wasserstein-2 barycenter problem

$$\min_\nu \left(\frac{1}{2}\mathsf{W}_2^2(\nu, f^*\sharp\mu^+) + \frac{1}{2}\mathsf{W}_2^2(\nu, f^*\sharp\mu^-) \right)$$

is given by the measure

$$\bar{\nu} = \frac{1}{2}\left(F_+^{-1} + F_-^{-1}\right) \circ F_+ \circ f^*\sharp\mu^+ \tag{8.8}$$

$$= \frac{1}{2}\left(F_+^{-1} + F_-^{-1}\right) \circ F_- \circ f^*\sharp\mu^- \ . \tag{8.9}$$

Indeed, observe that $\frac{1}{2}\left(F_+^{-1} + F_-^{-1}\right) \circ F_+$ is the optimal transportation plan from $f^*\sharp\mu^+$ to the barycenter of $f^*\sharp\mu^+, f^*\sharp\mu^-$. Since Eq. (8.8) corresponds to $f\sharp\mu^+$ on $\text{supp}(\mu^+)$ and Eq. (8.9) to $f\sharp\mu^-$ on $\text{supp}(\mu^-)$, the distances to the barycenter being equal, we have

$$\mathsf{W}_2^2(f^*\sharp\mu^+, f\sharp\mu^+) = \mathsf{W}_2^2(f^*\sharp\mu^-, f\sharp\mu^-) \ . \tag{8.10}$$

Plugging (8.10) in (8.7) yields

$$\mathcal{R}(f) = \int (f^*(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\mu_{\boldsymbol{X}|2}(\boldsymbol{x}) \ ,$$

and concludes the proof. □

CHAPTER 9

Classification with abstention but without disparities

Classification with abstention has gained a lot of attention in recent years as it allows to incorporate human decision-makers in the process. Yet, abstention can potentially amplify disparities and lead to discriminatory predictions. The goal of this work is to build a general purpose classification algorithm, which is able to abstain from prediction, while avoiding disparate impact. We formalize this problem as risk minimization under fairness and abstention constraints for which we derive the form of the optimal classifier. Building on this result, we propose a post-processing classification algorithm, which is able to modify any off-the-shelf score-based classifier using only unlabeled sample. We establish finite sample risk, fairness, and abstention guarantees for the proposed algorithm. In particular, it is shown that fairness and abstention constraints can be achieved independently from the initial classifier as long as sufficiently many unlabeled data is available. The risk guarantee is established in terms of the quality of the initial classifier. Our post-processing scheme reduces to a sparse linear program allowing for an efficient implementation, which we provide. Finally, we validate our method empirically showing that moderate abstention rates allow to bypass the risk-fairness trade-off.

Based on Nicolas Schreuder and Evgenii Chzhen (2021). "Classification with abstention but without disparities". In: *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*. Proceedings of Machine Learning Research. PMLR.

Contents

9.1 Introduction

In recent years classification with abstention or with reject option has gained a considerable amount of attention from both statistical and machine learning communities. Probably the earliest appearance of classification with reject option can be found in the works of Chow (1957) and Chow (1970) in the context of information retrieval and an initial statistical treatment was given in (Györfi, Györfi, and Vajda 1979). Much later, Herbei and Wegkamp (2006) provided non-parametric analysis for the problem of binary classification with a fixed rejection cost in the spirit of Audibert and Tsybakov (2007). Several extensions followed later, all working with fixed cost of rejection (Yuan and Wegkamp 2010; Wegkamp and Yuan 2011; Bartlett and Wegkamp 2008).

Following the conformal prediction literature (see, *e.g.,* Vovk, Gammerman, and Shafer 2005), Lei (2014a) considers a framework where ones wants to minimize the reject rate under a pre-specified accuracy constraint, meanwhile Denis and Hebiri (2020) target its reversed formulation. Both derive finite sample guarantees for plug-in type classification procedures and instanciate their analysis to standard non-parametric class of distributions. In a similar direction, several practical methods (Grandvalet et al. 2008; Nadeem, Zucker, and Hanczar 2009) have been proposed in the machine learning community to address the problem of classification with abstention. Recently, Bousquet and Zhivotovskiy (2019), Neu and Zhivotovskiy (2020), and Puchkin and Zhivotovskiy (2021) show that abstention can significantly improve regret bounds and convergence rates for the problems of online and batch classification.

Crucially, in our work we view abstention as a mechanism to lighten the burden of fairness constraints and bypass the risk-fairness trade-off (Agarwal et al. 2018; Menon and Williamson 2018b; Chzhen and Schreuder 2020a): one can enjoy the best of both worlds – a simultaneously fair and accurate classifier – at the cost of rejection. A majority of observations are still classified in an automatic manner, while the rejected ones can be handled by, *e.g.,* human experts. Importantly, in our setting, the rejection rate is rigorously controlled by the practitioner depending on the number of available experts. In addition, since it is illusory to assume that a data-dependent classifier can make error-less and trustworthy decisions, it is desirable to put human experts back in the loop for sensitive tasks. The rejection mechanism partially transfers the burden of optimizing those conflicting quantities to human experts, who can eventually have access to more information to make a better informed decision (*e.g.,* a doctor can ask for extra medical examination for its final diagnosis).

Fairness in binary classification is a very popular topic with various types of algorithmic and statistical contributions (see, *e.g.,* Hardt, Price, and Srebro 2016; Barocas, Hardt, and Narayanan 2019). However, abstention framework has not yet received a lot of attention in the context of fair learning. Notable exceptions are work of Madras, Pitassi, and Zemel (2018) and Jones et al. (2020). The latter demonstrates that an imprudent use of abstention might amplify potential disparities already present in the data. In particular, they show that in the framework of prediction without disparate treatment (Zafar et al. 2017) the use of the same rejection threshold across sensitive groups might result in a large group-wise risks disparities. As a potential remedy, our work offers a theoretically grounded way to enforce fairness constraints as well as a desired group-dependent reject rates. The idea of relying on a reject mechanism to enforce fairness has only been explored once, in Madras, Pitassi, and Zemel (2018). The authors introduce "learning to defer" framework – an extension of classification with abstention – where the cost of rejection is allowed to depend on the prediction of an external decision-maker (*e.g.,* a human expert). The authors argue that by making the automated model aware of the potential biases and weaknesses of the external decision-maker, it can globally optimize for accuracy and fairness. The authors enforce Equalized Odds (Hardt, Price, and Srebro 2016) through regularization of the risk and thus cannot control explicitly the reject rate, which might potentially lead to a huge external decision-maker costs. While the authors provide empirical evidences of their claims, theoretical justification of their results remains open. Our work offers a completely theory-driven way to enforce both fairness and rejection constraints while optimizing for accuracy, leading to a computationally efficient post-processing algorithm.

Contributions. Our work combines and extends previous results in abstention framework with recent results on fair binary classification. Namely, similarly to (Denis and Hebiri 2020), we aim at minimizing misclassification risk under a control over *group-wise* reject rates. As we would like to avoid disparate impact, we explicitly add this as a constraint to our framework. We derive the optimal form of a reject classifier, which minimizes the misclassification risk under the discussed constraints. Our explicit characterization of the optimal reject classifier provides a better understating of the interplay between, on one side, the fairness and rejection constraints and, on the other side, the accuracy. We propose a data-driven post-processing algorithm which enjoys generic plug-and-play finite sample guarantees. An appealing feature of our post-processing algorithm is that it can be used on top of *any* pre-trained classifier, thus avoiding the – potentially high – cost of re-fitting a classifier from scratch. From numerical perspective, the proposed method reduces to a solution of a sparse linear program, allowing us to leverage efficient LP solvers. Numerical experiments validate our theoretical result demonstrating that the proposed method successfully enforces fairness and rejection constraints in practice, while achieving a high level of accuracy.

Notation. For each $K \in \mathbb{N}$ we denote by $[K]$ the set of the first K positive integers. The standard Euclidean inner product is denoted by $\langle \cdot, \cdot \rangle$. For a real number $a \in \mathbb{R}$ we write $(b)_+$ (*resp.* $(a)_-$) to denote the positive (*resp.* the negative) part of a. For two real numbers a, b we denote by $a \vee b$ (*resp.* $a \wedge b$) the maximum (*resp.* the minimum) between the two. We denote by $\mathbf{1} \in \mathbb{R}^K$ the vector composed of ones and by $\boldsymbol{e}_s \in \mathbb{R}^K$ the s^{th} basis vector of $\mathbb{R}^K$.

9.2 Problem presentation

Consider a triplet $(\boldsymbol{X}, S, Y) \sim \mathbb{P}$, where $\boldsymbol{X} \in \mathbb{R}^d$ is the feature vector, $S \in [K]$ is the sensitive attribute, and $Y \in \{0,1\}$ is the binary label to be predicted. A classifier is a mapping $g : \mathbb{R}^d \times [K] \to \{0,1,r\}$. That is, any classifier g is able to provide a prediction in $\{0,1\}$, or to abstain from prediction by outputting r. With any classifier g, we associate the following quantities:

$$\begin{aligned}
&\mathcal{R}(g) := \mathbb{P}(Y \neq g(\boldsymbol{X}, S) \mid g(\boldsymbol{X}, S) \neq r) \ , \\
&\mathrm{NAb}_s(g) := \mathbb{P}(g(\boldsymbol{X}, S) \neq r \mid S = s) \ , \\
&\mathrm{NAb}(g) := \mathbb{P}(g(\boldsymbol{X}, S) \neq r) \ , \\
&\mathrm{PT}_s(g) := \mathbb{P}(g(\boldsymbol{X}, S) = 1 \mid S = s, g(\boldsymbol{X}, S) \neq r) \ , \\
&\mathrm{PT}(g) := \mathbb{P}(g(\boldsymbol{X}, S) = 1 \mid g(\boldsymbol{X}, S) \neq r)) \ .
\end{aligned} \tag{9.1}$$

The first one is the risk of a classifier, which measures the probability of incorrect prediction, given that an actual prediction was issued. The second two quantities measure the group-wise and marginal prediction rates. The last two quantities describe the group-wise and marginal rates of positive predictions given that the prediction was made. Intuitively, a good classifier has low risk $\mathcal{R}$, high NAb_s, and low disparities between $\mathrm{PT}_s(g)$.

Fairness constraint. We formalize fairness through the notion of Demographic Parity (see for instance, Barocas, Hardt, and Narayanan 2019). A predictor g is said to satisfy Demographic Parity (or, equivalently, to avoid Disparate Impact) if the distribution of its prediction is independent from the sensitive attribute. Formally, in the standard binary classification framework it means that for any $z \in \{0,1\}$ and for any $s, s' \in [K]$,

$$\mathbb{P}(g(X, S) = z \mid S = s) = \mathbb{P}(g(X, S) = z \mid S = s') \ .$$

In the setting of classification with abstention, we naturally want to condition on the fact that the classifier issues a prediction, that is, $g(X, S) \neq r$. Using the quantities introduced in Eq. (9.1), the latter reduces to

$$\forall s \in [K], \quad \mathrm{PT}_s(g) = \mathrm{PT}(g) \ .$$

Penalized version. There are various trade-offs that one can consider between the quantities in Eq. (9.1). For instance, adapting the approach of Herbei and Wegkamp (2006) to the context of fairness, one can target a prediction which avoids disparate impact and minimizes penalized risk. Formally, it amounts to solving the following problem:

$$\begin{aligned}
&\min_{g:\mathbb{R}^d \times [K] \to \{0,1,r\}} \mathcal{R}(g) + \sum_{s=1}^{K} \lambda_s \, \mathrm{NAb}_s(g) \\
&\text{s.t. } \forall s \in [K], \quad \mathrm{PT}_s(g) = \mathrm{PT}(g)
\end{aligned} \ , \tag{\textbf{P-DPWA}}$$

for some $\lambda_s \geq 0$, $s \in [K]$. This approach also resembles the one employed by Madras, Pitassi, and Zemel 2018, who additionally penalized for fairness violation instead of directly controlling it. The main issue with the formulation (**P-DPWA**) is connected with the choice of the penalization parameters $\lambda_s \geq 0$, $s \in [K]$, which do not have simple and intuitive interpretation.

Indeed, it is impossible to know beforehand which $\lambda_s \geq 0$, $s \in [K]$ will result in a usable reject rate, forcing the practitioner to explore the whole space of the hyperparameters $\lambda_s \geq 0$, $s \in [K]$. Instead of the above formulation, we consider the problem in which one is able to *explicitly* control the rejection rate. In particular, such an approach allows us to develop a *parameter-free* post-processing method.

Explicit control of reject. Given $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top \in [0,1]^K$, our goal is to find a solution of the following problem

$$\begin{aligned} &\min_{g:\mathbb{R}^d \times [K] \to \{0,1,r\}} \mathcal{R}(g) \\ &\text{s.t. } ,\forall s \in [K], \begin{cases} \mathrm{NAb}_s(g) = \alpha_s \\ \mathrm{PT}_s(g) = \mathrm{PT}(g) \end{cases} . \end{aligned} \tag{DPWA}$$

It will be shown later that, under a mild assumption on the distribution of the conditional expectation $\mathbb{E}[Y \mid \boldsymbol{X}, S]$, the above problem admits a global minimizer written in the form of group-wise thresholding.

The first constraint in (DPWA) specifies the abstention level accepted for each class while the second constraint, as before, demands the classifier g to avoid disparate impact. Notably, in this formulation, the parameter vector $\boldsymbol{\alpha} \in [0,1]^K$ has a simple and intuitive interpretation – it allows to fix precisely *different* levels of rejects for different groups. This, for instance, can be beneficial, if $g(\boldsymbol{x}, s) = r$ is followed by the intervention of a human decision-maker, who replaces the classifier. One can force a higher rejection rate (*i.e.,* a higher rate of human intervention) for disadvantaged groups by lowering the corresponding $\alpha_s \in [0,1]$. Crucially, we implicitly assume that the practitioner is able to treat unclassified instances in an accurate and fair manner. While this assumption is void for the theoretical contributions of this chapter, we *warn* the practitioner that it must not be overlooked once our method is deployed in real world.

This formulation allows to bypass the usual trade-off between fairness and accuracy at the price of rejection. Indeed, note that a classifier that solves (DPWA) is fair for any parameters $(\alpha_s)_{s \in [K]}$. At the same time, setting $\alpha_1 = \ldots = \alpha_K = \tilde{\alpha}$ for some $\tilde{\alpha} \in (0,1]$, one can observe that by varying $\tilde{\alpha}$ we can recover the accuracy of a classifier without constraints while still satisfying Demographic Parity. This will be later empirically confirmed in Section 9.7. We again emphasize that the accuracy gain comes at a price of a possible reject region, which, depending on the application at hand might or might not constitute a reasonable price.

9.3 Optimal classifier

Our first theoretical contribution is the derivation of a classification strategy g^*, which is a solution of (DPWA). We define the conditional expectation of the label Y knowing $(\boldsymbol{X}, S)$ as

$$\eta(\boldsymbol{X}, S) = \mathbb{E}[Y \mid \boldsymbol{X}, S] \ .$$

It is known that the Bayes optimal rule for the problem of binary classification with misclassification risk is given by the point-wise thresholding of $\eta(\boldsymbol{X}, S)$ on the level $1/2$ (Devroye,

Györfi, and Lugosi 2013). In our case the classifier does not correspond to the Bayes decision. Instead, it is a solution of a constrained optimization problem with constraints that depend on the unknown data distribution $\mathbb{P}$. In several frameworks, which are also formulated as risk minimization under distribution dependent constraints, it is possible to obtain a closed form expression of a minimizer under fairly mild assumptions. In particular, it is the case for the classification with reject option (Chow 1970; Lei 2014a; Denis and Hebiri 2020) as well as classification under various fairness constraints (Hardt, Price, and Srebro 2016; Chzhen et al. 2019; Barrio, Gordaliza, and Loubes 2020). Similarly to the above contributions, we will make a mild assumption on the behaviour of $\eta(\boldsymbol{X}, S)$, which is, for instance, naturally satisfied whenever $\eta(\boldsymbol{X}, S)$ admits a density *w.r.t.* the Lebesgue measure.

Assumption 9.3.1. *The random variables* $(\eta(\boldsymbol{X}, S) \mid S = s)$ *are non-atomic for all* $s \in [K]$.

One can actually get rid of this assumption, as explained in Lei (2014a), by switching from deterministic classification strategies, which are valued in $\{0, 1, r\}$, to randomized classifiers, which output a distribution over $\{0, 1, r\}$.

To present the main result of this section, we introduce the notations $p_s := \mathbb{P}(S = s)$, $\bar{\alpha} := \sum_{s\in[K]} p_s\alpha_s$ and we define the following function

$$G(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \left|\frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{x}, s) - \langle\boldsymbol{\gamma}, \mathbf{1}\rangle) + \frac{\boldsymbol{\gamma}_s}{2\alpha_s}\right| - \frac{p_s}{2\bar{\alpha}}(1 - \langle\boldsymbol{\gamma}, \mathbf{1}\rangle) - \boldsymbol{\lambda}_s - \frac{\boldsymbol{\gamma}_s}{2\alpha_s} \ ,$$

which plays a key role in the derivation of an optimal classifier for (**DPWA**). We now state the first result of this work, which provides a form of g^* – solution for (**DPWA**).

Theorem 9.3.2. *Under Assumption 9.3.1, an optimal classifier for* (**DPWA**) *is given for all* $(\boldsymbol{x}, s) \in \mathbb{R}^d \times [K]$ *by*

$$g^*(\boldsymbol{x}, s) = \begin{cases} r & \text{if } G(\boldsymbol{x}, s, \boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*) \leq 0 \\ \mathbb{1}\left(\eta(\boldsymbol{x}, s) \geq \frac{1}{2} + c_{\boldsymbol{\gamma}^*, s}\right) & \text{otherwise} \end{cases} \ ,$$

where $(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$ *are solutions of*

$$\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left\{ \langle\boldsymbol{\lambda}, \boldsymbol{\alpha}\rangle + \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s}[(G(\boldsymbol{X}, S, \boldsymbol{\lambda}, \boldsymbol{\gamma}))_+] \right\} \ ,$$

and $c_{\boldsymbol{\gamma}^*,s} := \frac{1}{2}\left(\frac{\bar{\alpha}\boldsymbol{\gamma}^*_s}{\alpha_s p_s} - \langle\mathbf{1}, \boldsymbol{\gamma}^*\rangle\right)$.

Let us mention that unlike other similar results described above, the main difficulty in the proof of Theorem 9.3.2 lies in the fact the misclassification risk in our case involves conditioning on the event which itself depends on the classifier that we want to find. Theorem 9.3.2 is instructive and allows to develop an intuition which is similar to that of the original rule derived by Chow 1957; Chow 1970. To be more precise, denoting by

$$t_{\boldsymbol{\gamma}^*,s} := (1 - \langle\boldsymbol{\gamma}, 1\rangle) + \frac{\bar{\alpha}\boldsymbol{\gamma}_s}{p_s\alpha_s} \ ,$$

the reject region is expressed as a strip around $t_{\boldsymbol{\gamma},s}$:

$$|\eta(\boldsymbol{x}, s) - t_{\boldsymbol{\gamma},s}| \leq t_{\boldsymbol{\gamma},s} + \frac{\bar{\alpha}\boldsymbol{\lambda}_s}{p_s} \ .$$

We highlight that the center as well as the size of this strip is group-dependent. Interestingly, the position of the strip only depends on the Lagrange multiplier controlling for the fairness constraint, while its width is determined by both constraints.

9.4 Empirical method

The form of the optimal classifier suggests to develop a post-processing algorithm, which receives an estimator $\widehat{\eta}(\boldsymbol{x}, s)$ of $\eta(\boldsymbol{x}, s)$ and an additional *unlabeled* set of samples to estimate $(\boldsymbol{\gamma}^*, \boldsymbol{\lambda}^*)$. Indeed, observe that the optimal classifier g^* is know up to the quantities $\eta(\boldsymbol{x}, s), \boldsymbol{\gamma}^*, \boldsymbol{\lambda}^*$.

Remark 9.4.1. *For simplicity of exposition we assume that the marginal distribution of S is known, that is, we have access to $p_s := \mathbb{P}(S = s)$. Note that S follows multinomial distribution, and, in practice, we can estimate these probabilities by their empirical counterparts, which is the direction that we take in our experimental section. Our proofs generalize straightforwardly for the case of unknown p_s, but such modification results in additional, unnecessary, complications.*

We denote by $\widehat{\eta}(\boldsymbol{X}, S)$ any off-the-shelf estimator of $\eta(\boldsymbol{X}, S)$. For instance, one can take k-NN (Stone 1977; Devroye, Györfi, and Lugosi 2013), locally polynomial estimator (Korostelev and Tsybakov 2012), logistic regression Bühlmann and Van de Geer 2011, random forest (Breiman 2001; Biau and Scornet 2016; Mourtada, Gaïffas, and Scornet 2020) to name a few. Our theoretical guarantees on the misclassification risk will explicitly depend on the quality of this off-the-shelf estimator, hence it is advisable to use those methods which are supported by statistical guarantees. Yet, our algorithm remains valid even for inconsistent estimators $\widehat{\eta}$ in the sense that the resulting classifier after post-processing will (nearly) satisfy the prescribed constraints independently from $\widehat{\eta}$.

Remark 9.4.2. *In what follows we assume that the estimator $\widehat{\eta}(X, S)$ is independent from the* unlabeled *sample (introduced below) and is valued in $[0, 1]$. In other words, we require a new* unseen *unlabeled sample for the post-processing. As it will be seen from our bound, the assumption that $\widehat{\eta}(X, S)$ is valued in $[0, 1]$ is not restrictive, since we can always perform clipping without damaging statistical properties. On a more technical note, we require that $\mathbb{P}(\widehat{\eta}(X, S) = c \mid \widehat{\eta}) = 0$ almost surely for any $c \in [0, 1]$. Again, this assumption is not restrictive, since we can always randomize the output of $\widehat{\eta}(X, S)$ by adding a negligible noise coming from a continuous distribution. In Algorithm 2 we use uniformly distributed noise supported on $[0, \sigma]$, with σ being a small parameter. One can take this parameter σ arbitrarily small, preserving the statistical properties of $\widehat{\eta}$.*

As mentioned before, to build the post-processing scheme, we will use only *unlabeled* sample. We also do not restrict ourselves to sampling from $\mathbb{P}_{(\boldsymbol{X},S)}$. Instead, we assume that for all $s \in [K]$ we observe $\{X_i\}_{i \in \mathcal{I}_s}$ sampled i.i.d. from $\mathbb{P}_{\boldsymbol{X}|S=s}$. In the above notation, $\mathcal{I}_s$ have cardinality n_s and they form a partition of $[n]$. That is, we have that $n_1 + \ldots + n_K = n$. The described sampling scheme is potentially appealing in situations when it is possible to gather a lot of data about the minority group without the need of labeling them. In particular, this sampling scheme allows to set $n_1 = \ldots = n_K$, which, since we do not require labeling, is more realistic. The conditional expectation $\mathbb{E}_{\boldsymbol{X}|S=s}$ is estimated based on the following empirical

Procedure 2 `Post-processing`

Input: base estimator $\widehat{\eta}$, unlabeled data $\{\boldsymbol{X}_i\}_{i\in\mathcal{I}_s}$ for $s \in [K]$, noise magnitude σ
`Randomize`:
for $i \in \mathcal{I}_s, s \in [K]$ **do**
  Sample independently $\zeta_i \sim \mathcal{U}([0, \sigma])$
  Set $\widehat{\eta}(\boldsymbol{X}_i, s) \leftarrow \widehat{\eta}(\boldsymbol{X}_i, s) + \zeta_i$
end for
`Solve`: Eq. (9.3) based on LP formulation to get $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}})$
Output: $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}})$

measure

$$\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} = \frac{1}{n_s} \sum_{i\in\mathcal{I}_s} \delta_{\boldsymbol{X}_i} \ .$$

Before providing the proposed post-processing method, we define the empirical counterpart to the function G as

$$\widehat{G}(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \left| \frac{p_s}{2\bar{\alpha}}(1 - 2\widehat{\eta}(\boldsymbol{x}, s) - \langle \boldsymbol{\gamma}, \mathbf{1}\rangle) + \frac{\gamma_s}{2\alpha_s} \right| - \frac{p_s}{2\bar{\alpha}}(1 - \langle \boldsymbol{\gamma}, \mathbf{1}\rangle) - \boldsymbol{\lambda}_s - \frac{\gamma_s}{2\alpha_s}$$

The post-processing classifier with abstention is given by

$$\widehat{g}(\boldsymbol{x}, s) = \begin{cases} r & \text{if } \widehat{G}(\boldsymbol{x}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) \leq 0 \\ \mathbb{1}\left(\widehat{\eta}(\boldsymbol{x}, s) > \frac{1}{2} + c_{\widehat{\boldsymbol{\gamma}},s}\right) & \text{otherwise} \end{cases} , \tag{9.2}$$

where $c_{\widehat{\boldsymbol{\gamma}},s} := \frac{1}{2}\left(\frac{\bar{\alpha}\widehat{\gamma}_s}{\alpha_s p_s} - \langle \mathbf{1}, \widehat{\boldsymbol{\gamma}}\rangle\right)$ and $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}})$ is a solution of

$$\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left\{ \langle \boldsymbol{\lambda}, \boldsymbol{\alpha}\rangle + \sum_{s=1}^{K} \widehat{\mathbb{E}}_{\boldsymbol{X}|S=s}(\widehat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}))_+ \right\} \ . \tag{9.3}$$

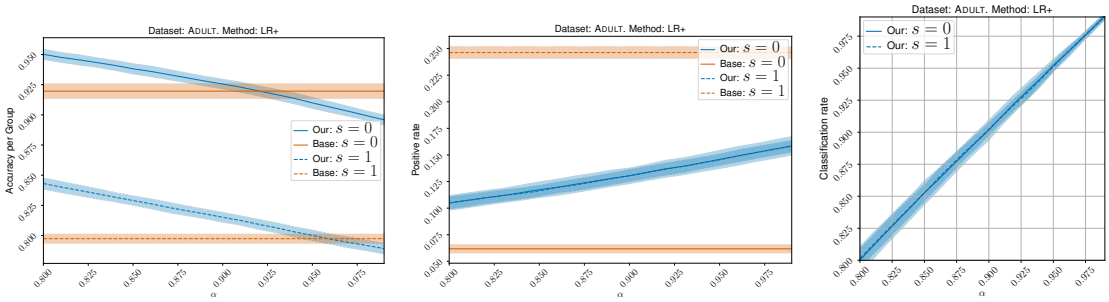
We summarize the proposed procedure in Algorithm 2 incorporating the randomization step. Note that there is a clear analogy between the result of Theorem 9.3.2 and the constructed algorithm. Indeed, the latter is an empirical version of the former built via the plug-in approach.

Lemma 9.4.3. *The minimization problem in Eq. (9.3) is convex and it admits a global minimizer.*

In Section 9.6 we will actually prove a stronger statement. Namely, it will be shown that the minimization problem in Eq. (9.3) is equivalent to a linear program with sparse constraints, which will allow us to provide an efficient implementation of the proposed procedure.

9.5 Finite sample guarantees

In this section we provide finite sample guarantees on the behavior of the post-processing classifier with abstention regarding its performance, its reject rate and its fairness. In order to



(a) Accuracy per group (b) Positive rate (c) Classification rate

Figure 9.1: Results on Adult dataset with Logistic Regression (LR) as the base estimator. Blue lines correspond to our post-processing method; Orange lines correspond to the base classifier. Dashed line correspond to $s = 1$ and solid line to $s = 0$. Shaded areas correspond to the variance of the result over 20 repetitions.

lighten the presentation of our results, let us define now the sequence

$$u_n^{\delta,K} := \sqrt{\frac{2\log(4K/\delta)}{2n}} + \frac{2}{n} \quad , \quad \forall n \geq 1 \ .$$

The sequence $u_n^{\delta,K}$ behaves as $O(\sqrt{\log(K/\delta)/n})$, that is, it depends logarithmically on the number of sensitive attributes K, on the confidence parameter δ and goes to zero as $n^{-1/2}$ with the growth of n. Our goal in this section is to derive constraint and risk guarantees. Namely, we would like to show that when $n_s \to \infty$ we have for all $s \in [K]$ that

$$\begin{aligned} &|\mathrm{NAb}_s(\widehat{g}) - \alpha_s| \to 0 \\ &|\mathrm{PT}_s(\widehat{g}) - \mathrm{PT}(\widehat{g})| \to 0 \end{aligned} \qquad \text{and} \qquad \mathcal{E}(\widehat{g}) := \mathcal{R}(\widehat{g}) - \mathcal{R}(g^*) \to 0 \ .$$

The first part ensures satisfaction of reject and fairness constraints, while the second part shows that the risk of the proposed method is similar to that of g^*. Importantly, both guarantees will be derived in the finite-sample regime and with high probability.

The next proposition provides a quantitative control on the violation of the reject and Demographic Parity constraints in the finite sample regime.

Proposition 9.5.1. *Let $\delta \in (0,1)$. The violation of the constraints by the post-processing classifier with abstention $\widehat{g}$ defined in Eq. (9.2) can be controlled, with probability at least $1-\delta$, for any $s \in [K]$, as*

$$|\mathrm{NAb}_s(\widehat{g}) - \alpha_s| \leq u_{n_s}^{\delta/2,K}, \quad \textit{and} \quad |\mathrm{PT}_s(\widehat{g}) - \mathrm{PT}(\widehat{g})| \leq \frac{6}{\alpha_s} u_{n_s}^{\delta,K} + \frac{6}{\bar{\alpha}} \sum_{s=1}^{K} p_s u_{n_s}^{\delta,K} \ .$$

The proof for the control of the reject rate is postponed to Section 9.9.3 while the proof for the control of the Demographic Parity constraint can be found in Section 9.9.4.

Remarkably Proposition 9.5.1 is assumption-free. In particular it does not depend on the conditional expectation η as well as it does not depend on the initial estimator $\widehat{\eta}$. If one has

enough *unlabeled* data than one can get arbitrarily close to exact satisfaction of the constraints. Intuitively, this is the case because the fairness and reject constraints only depend on the conditional distribution of the feature vector $\boldsymbol{X}$ given the sensitive attribute S, not on the relation between the features and the label Y.

We also remark that both bounds of Proposition 9.5.1 depend on the amount of observation available for each group $s \in [K]$ – it is easier to satisfy constraints for well-represented groups. In particular, it is advisable to collect an unlabeled sample which is balanced in terms of the sensitive attributes. Note that it is explicitly allowed in our framework, since we require samples from $\mathbb{P}_{X|S=s}$ and not from $\mathbb{P}_{(X,S)}$.

The next result establishes excess risk guarantees for the proposed method.

Proposition 9.5.2. *Assume that* $2u_{n_s}^{\delta,K} < \alpha_s < 1 - 2/n_s$ *for any* $s \in [K]$ *and that Assumption 9.3.1 holds. Then, for any* $\delta \in (0,1)$*, the excess risk of the post-processing classifier with abstention* $\widehat{g}$ *defined in Eq. (9.2) satisfies with probability at least* $1-\delta$*,*

$$\mathcal{E}(\widehat{g}) \leq \frac{3}{\bar{\alpha}} \|\eta - \widehat{\eta}\|_1 + 6 \sum_{s=1}^{K} \left(\frac{p_s}{\bar{\alpha}} + \frac{1}{\alpha_s} \right) u_{n_s}^{\delta,K} \quad . \tag{9.4}$$

For convenience and clarity of exposition we stated separately the control on the constraint and on the excess risk. However, we remark that both Proposition 9.5.1 and Proposition 9.5.2 hold on the same high-probability event.

We naturally conclude from Proposition 9.5.2 that if one has access to a consistent estimator $\widehat{\eta}$ of η, *i.e.,* such that $\|\eta - \widehat{\eta}\|_1$ goes to 0 as the sample sizes $(n_s)_{s=1}^K$ go to infinity, then the excess risk can be made arbitrarily small by getting more labeled and unlabeled data.

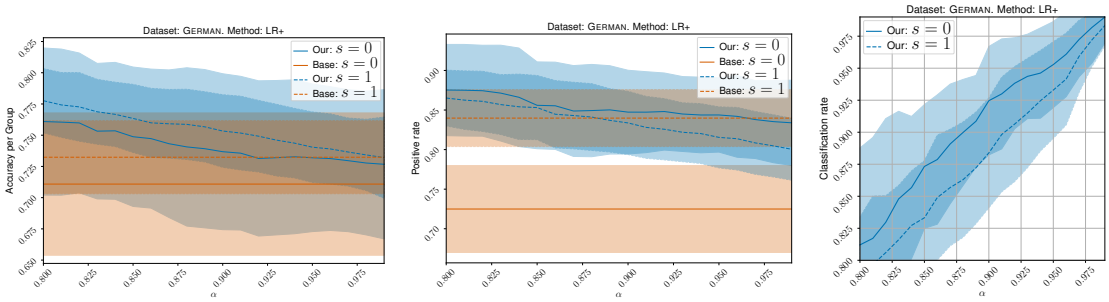
The only assumption, constraining the reject rates $(\alpha_s)_{s=1}^K$, is quite benign. Recall that α_s is the rate at which the classifier is asked to give a prediction thus, in practice, it is expected to be at least greater than a half. Furthermore, note that it only depends on the size of the unlabeled dataset thus, if one has enough samples, this assumption essentially holds for free. If the sample size is small, than one has to allow the classifier to reject more often in order to satisfy the constraints. Similar constraints are present in other contributions (**agarwal2018reductions**; see *e.g.,* Agarwal, Dudík, and Wu 2019).

Our theoretical analysis is inspired by that of Chzhen et al. 2020b. However, their results hold only in expectation while ours hold with high-probability. Moreover, due to the interplay of the reject and demographic parity constraints, their proof technique requires a non-trivial adaptation to our context.

9.6 LP reduction

We recall that the proposed post-processing scheme involves solving convex non-smooth minimization problem in Eq. (9.3). While for low values of K (few sensitive attributes) this problem can be solved via simple grid-search, which would be faster than sub-gradient methods, large values of K can pose significant computational difficulties.

It turns out that the minimization problem in Eq. (9.3) is equivalent to Linear Programming (LP) (Matousek and Gärtner 2007) with sparse constraint matrix. For any matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ we denote by $\mathrm{nnz}(\mathbf{A})$ the number of non-zero elements of $\mathbf{A}$.



(a) Accuracy per group (b) Positive rate (c) Classification rate

Figure 9.2: Results on German dataset with Logistic Regression (LR) as the base estimator. Blue lines correspond to our post-processing method; Orange lines correspond to the base classifier. Dashed line correspond to $s = 1$ and solid line to $s = 0$. Shaded areas correspond to the variance of the result over 20 repetitions.

Proposition 9.6.1. *There exist* $\boldsymbol{c} \in \mathbb{R}^{n+2K}$, $\boldsymbol{b} \in \mathbb{R}^{2n}$, $\mathbf{A} \in \mathbb{R}^{2n \times (n+2K)}$ *with* $\operatorname{nnz}(\mathbf{A}) \leq 4n + nK$, *such that the minimization problem in Eq.* (9.3), *is equivalent to*

$$\begin{aligned} &\min_{\boldsymbol{y} \in \mathbb{R}^{n+2K}} \langle \boldsymbol{c}, \boldsymbol{y} \rangle \\ &s.t. \quad \begin{cases} \mathbf{A}\boldsymbol{y} \leq \boldsymbol{b} \\ y_i \geq 0 \qquad i \in [n] \end{cases} . \end{aligned} \tag{LP}$$

Due to the space considerations, the previous result is stated in existential form, however, all the parameters of the LP are explicit and are provided in the supplementary material. Seminal works of (Khachiyan 1979; Karmarkar 1984) confirmed that LP with rational coefficients can be solved in weakly polynomial time. Since then, extremely efficient solvers were developed based on the interior-point and simplex methods. The fact that the post-processing reduces to an LP problem allows us to use these fast solvers. In particular, most of the computational burden lies on the training of the base estimator $\widehat{\eta}$ while the post-processing can be performed almost instantly. From theoretical perspective, one can leverage the sparse structure of the problem using, for instance, the result of (Lee and Sidford 2015) who provide an efficient solver to find an ε solution of an LP in $\tilde{O}((\operatorname{nnz}(\mathbf{A}) + n^2)\sqrt{n}\log(\varepsilon^{-1}))$ time. In particular, the previous guarantee scales only linearly with the number of sensitive attributes and logarithmically with the precision ε. However, in our practical implementation of the proposed method, we use interior point method available as a part of `scipy.optimize.linprog` (Virtanen et al. 2020).

9.7 Experiments

We provide an implementation of the proposed post-processing procedure described in Algorithm 2 using `scipy.optimize.linprog` (Virtanen et al. 2020), which implements interior point method for solving problem (**LP**). The source code is available at `https://github.com/evgchz/dpabst`. We consider Adult (Kohavi 1996) and German (Dua and Graff 2017) datasets, which are standard benchmark datasets in the fairness literature.

ADULT dataset is fetched via `fairlearn.datasets` (Bird et al. n.d.). This dataset contains 14 features and around $48,000$ observations. We dropped those observations that contain missing values. This dataset consists of 1994 US Census entries. Each entry of this dataset corresponds to an individual who is described by 14 characteristics, the binary target variable is equal to 1 if the individual earns more than $\$50K$ per year and it is set to 0 otherwise. In our experiments we take sex as a sensitive attribute.

GERMAN dataset is hosted on the UCI Machine Learning Repository (Dua and Graff 2017). Each of the $1,000$ entries represents a person who takes a credit by a bank. The binary target variable is equal to one if the individual is considered as good credit risks based on 20 categorical/symbolic attributes and is set to 0 otherwise. We use ordinal-encoding for ordinal variables and one-hot-encoding for other categorical variables which yields 46 features in total. In our experiments we take sex as sensitive attribute.

We consider the following off-the-shelf methods: Random Forest (RF) and Logistic Regression (LR). We used the `sklearn` (Pedregosa et al. 2011b) implementation of the aforementioned methods.

Each dataset of size N we partition in three parts. The first labeled part (60% of N) is used to train the base classifier, the second unlabeled part (20% of N) is used to apply the proposed post-processing, and the third part (20% of N) is used for evaluation of various statistics, which describe performance of the algorithm.

The hyperparameters of each base algorithm are tuned via 5-fold cross validation with accuracy as the performance measure. The regularization parameter of LR is searched among 30 values, equally spaced in logarithmic scale between 10^{-4} and 10^{4}. For RF the number of trees has been set to 1000 and the size of the subset of features optimized at each node has been searched in $\{d, \lceil d^{15/16} \rceil, \lceil d^{7/8} \rceil, \lceil d^{3/4} \rceil, \lceil d^{1/2} \rceil, \lceil d^{1/4} \rceil, \lceil d^{1/8} \rceil, \lceil d^{1/16} \rceil, 1\}$ where d is the number of features in the dataset. Recall that our post-processing algorithm is parameter-free, thus, the second step is performed without any tuning. Our setup allows to set different reject rates for different groups. However, the exact values heavily depend on the domain specific knowledge and on the problem itself. Because of that, in our experiments, we set $\alpha_1 = \ldots = \alpha_K = \alpha$ for 20 values of α taking values in the uniform grid over $[.8, .99]$, which correspond to reject rate ranging from 20% to 1%.

Given a classifier with reject option g and a test data $\mathcal{T} = \{(\boldsymbol{x}_i, s_i, y_i)\}_{i=1}^{n_{\text{test}}}$, we evaluate the following statistics

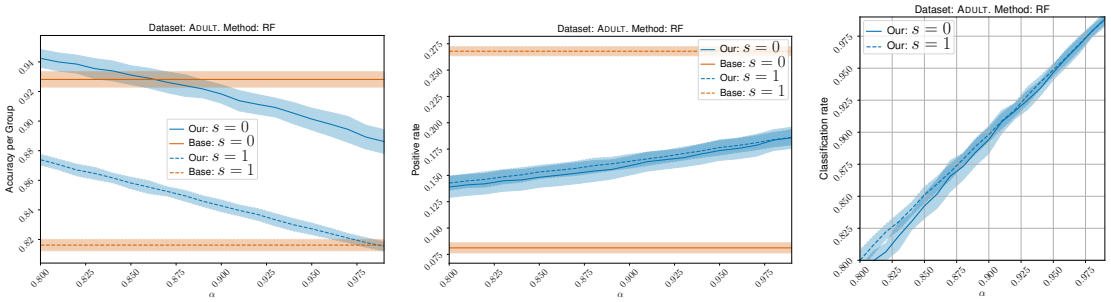
$$\widehat{\text{acc}}_s(g) = \frac{\sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{g(\boldsymbol{x}_i, s_i) = y_i\}\mathbb{I}\{s_i = s\}}{\sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{g(\boldsymbol{x}_i, s_i) \neq r\}\mathbb{I}\{s_i = s\}} \quad , \qquad s = 1, \ldots, K \quad ,$$

$$\widehat{\text{clf}}_s(g) = \frac{\sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{g(\boldsymbol{x}_i, s_i) \neq r\}\mathbb{I}\{s_i = s\}}{\sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{s_i = s\}} \quad , \qquad s = 1, \ldots, K \quad ,$$

$$\widehat{\text{pos}}_s(g) = \frac{\sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{g(\boldsymbol{x}_i, s_i) = 1\}\mathbb{I}\{s_i = s\}}{\sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{g(\boldsymbol{x}_i, s_i) \neq r\}\mathbb{I}\{s_i = s\}} \quad , \qquad s = 1, \ldots, K \quad .$$

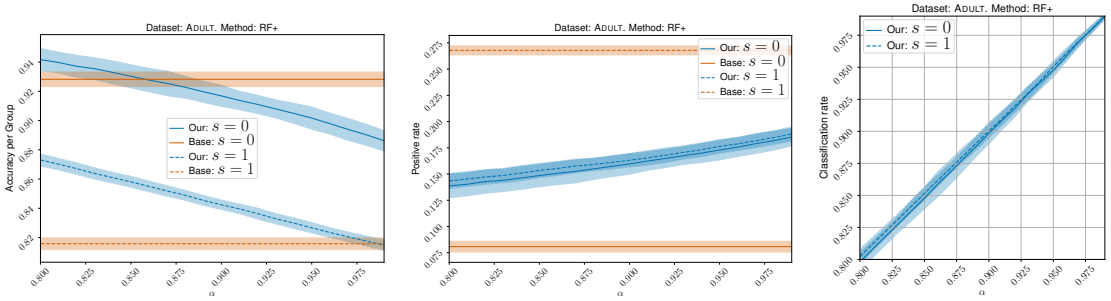
The first statistic measures the accuracy of g, the second the group-wise classification rate of g, and the third one measures the group-wise predicted positive rate of g. It is important to keep in mind that a classifier g which never rejects achieves $\text{clf}_s(g) = 1$ on any dataset.

Figure 9.1 presents results on ADULT dataset. First of all we observe that the proposed post-processing is effective in imposing reject and fairness constraints as illustrated on Figures 9.1(b)-



(a) Accuracy per group (b) Positive rate (c) Classification rate

Figure 9.3: Results on Adult dataset with Random Forest (RF) **without** additional randomization as the base estimator. Blue lines correspond to our post-processing method; Orange lines correspond to the base classifier. Dashed line correspond to $s = 1$ and solid line to $s = 0$. Shaded areas correspond to the variance of the result over 20 repetitions.



(a) Accuracy per group (b) Positive rate (c) Classification rate

Figure 9.4: Results on Adult dataset with Random Forest (RF) **with** additional randomization as the base estimator. Blue lines correspond to our post-processing method; Orange lines correspond to the base classifier. Dashed line correspond to $s = 1$ and solid line to $s = 0$. Shaded areas correspond to the variance of the result over 20 repetitions.

9.1(c). Looking at Figure 9.1(a), we observe that for already moderately low values of rejection our classification algorithm equalizes and even exceeds the accuracy per groups and overall of the base classifier. Figure 9.2 presents result on German dataset. Overall conclusions remain the same as for the Adult dataset. The main difference is an increase in variance of the result. This effect should not be attributed to the method itself but rather to the size of the two datasets. Indeed, Adult contains around $40,000$ observation, while German contains only $1,000$ observations. Hence, it is simply a more difficult task to learn stable classification algorithms on the German dataset. Remarkably, already 1% of reject rate allows to maintain the accuracy of the base classifier while significantly improving its fairness as illustrated on Figure 9.2(a).

We would also like to highlight the importance of the additive noise perturbation present in Algorithm 2. To this end, we consider RF classifier, which naturally does not lead to continuous estimator $\widehat{\eta}(\boldsymbol{X}, S)$ due to its partitioning nature. On Figure 9.3 we display the performance of our algorithm without any additional randomization and on Figure 9.4 follow

Algorithm 2 with $\sigma = 10^{-3}$. One can see that on Figure 9.3(c) the behaviour of our procedure fails to satisfy rejection rate constraints for lower values of α, even, considering the fact, that we have a rather large dataset. In contrast, this phenomenon disappears once the noise is added (see Figure 9.4(c)), confirming our theoretical findings. It is important to emphasize that this additional randomization has only a little impact on the group-wise accuracy, which suggest that the randomization step is always advisable in practice.

9.8 Conclusion

We proposed a classification with abstention algorithm which is able to satisfy Demographic Parity and whose reject rate is controlled explicitly. Our procedure is based on a post-processing scheme of any base estimator and can be computed efficiently using LP solvers. We derived distribution-free finite-sample guarantees demonstrating that the proposed method is able to achieve the prescribed constraints with high probability. Under additional mild assumption, we showed the risk of the proposed procedure nearly matches that of the theoretical minimum, provided the initial estimator is consistent. Our experimental results support the developed theory and suggest that by allowing small reject rate it is possible to avoid the accuracy-fairness trade-off.

9.9 Proofs

Appendix 9.9.1 is devoted to the proof of Theorem 9.3.2. Appendix 9.9.2 reminds and proves auxiliary results that are used in the rest of the supplementary material. The proof of Proposition 9.5.1 is split across Appendix 9.9.3 for the control of the reject rate and Appendix 9.9.4 for the control of the demographic parity violation. Appendix 9.9.5 contains the proof of Proposition 9.5.2. Finally, Appendix 9.9.6 provides a constructive proof of Proposition 9.6.1.

9.9.1 Derivation of the optimal prediction

Recall that we are interested in solving the following problem

$$\min_{g:\mathbb{R}^d\times[K]\to\{0,1,r\}} \mathbb{P}(g(\boldsymbol{X},S)\neq Y \mid g(\boldsymbol{X},S))\neq r)$$
$$\text{s.t.}, \forall s\in[K], \begin{cases}\mathbb{P}(g(\boldsymbol{X},S))\neq r \mid S=s)=\alpha_s,\\ \mathbb{P}(g(\boldsymbol{X},S)=1 \mid S=s, g(\boldsymbol{X},S)\neq r)=\mathbb{P}(g(\boldsymbol{X},S)=1 \mid g(\boldsymbol{X},S))\neq r).\end{cases}.$$

Simplifications

First we simplify the quantities involved in the above problem. Set $\bar{\alpha}=\sum_{s=1}^K p_s\alpha_s$ and recall that we defined the random variable $\eta(\boldsymbol{X},S)=\mathbb{E}[Y \mid \boldsymbol{X},S]$. Observe that for any g such that

$\mathbb{P}(g(\boldsymbol{X},S) \neq r \mid s=s) = \alpha_s$, we can write

$$\mathbb{P}(g(\boldsymbol{X},S) \neq Y \mid g(\boldsymbol{X},S) \neq r) = \sum_{s=1}^{K} \frac{p_s}{\bar{\alpha}} \mathbb{E}_{\boldsymbol{X}|S=s}\left[(1-\eta(\boldsymbol{X},S)\mathbb{1}_{g(\boldsymbol{X},S)=1} + \eta(\boldsymbol{X},S)\mathbb{1}_{g(\boldsymbol{X},S)=0}\right] \ ,$$

$$\mathbb{P}(g(\boldsymbol{X},S) \neq r|S=s) = \mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1} + \mathbb{1}_{g(\boldsymbol{X},S)=0}\right] \ ,$$

$$\mathbb{P}(g(\boldsymbol{X},S) = 1 \mid g(\boldsymbol{X},S) \neq r) = \sum_{s=1}^{K} \frac{p_s}{\bar{\alpha}} \mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}\right] \ ,$$

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

Lagrangian

We introduce the Lagrangian $\mathcal{L}$ of the constrained minimization problem as

$$\begin{aligned}\mathcal{L}(g,\boldsymbol{\lambda},\boldsymbol{\gamma}) = \mathbb{P}(g(\boldsymbol{X},S) \neq Y \mid g(\boldsymbol{X},S) \neq r) + \sum_{s=1}^{K} \lambda_s(\mathbb{P}(g(\boldsymbol{X},S) \neq r|S=s) - \alpha_s) \\ + \sum_{s=1}^{K} \gamma_s(\mathbb{P}(g(\boldsymbol{X},S) = 1|S=s, g(\boldsymbol{X},S) \neq r) - \mathbb{P}(g(\boldsymbol{X},S) = 1|g(\boldsymbol{X},S) \neq r)) \ .\end{aligned}$$

Using the simpler expressions we derived earlier, the Lagrangian can be expressed as

$$\begin{aligned}\mathcal{L}(g,\boldsymbol{\lambda},\boldsymbol{\gamma}) &= \sum_{s=1}^{K} \frac{p_s}{\bar{\alpha}} \mathbb{E}_{\boldsymbol{X}|S=s}\left[(1-\eta(\boldsymbol{X},S))\mathbb{1}_{g(\boldsymbol{X},S)=1} + \eta(\boldsymbol{X},S)\mathbb{1}_{g(\boldsymbol{X},S)=0}\right] \\ &+ \sum_{s=1}^{K} \lambda_s \left\{\mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1} + \mathbb{1}_{g(\boldsymbol{X},S)=0}\right] - \alpha_s\right\} \\ &+ \sum_{s=1}^{K} \frac{\gamma_s}{\alpha_s} \mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}\right] - \left(\sum_{s'=1}^{K} \gamma_{s'}\right)\left(\sum_{s=1}^{K} \frac{p_s}{\bar{\alpha}} \mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}\right]\right) \ .\end{aligned}$$

After straightforward algebraic manipulations, the Lagrangian can be simplified to

$$\mathcal{L}(g,\boldsymbol{\lambda},\boldsymbol{\gamma}) = \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s}\left[H_{(\boldsymbol{X},s)}(g,\boldsymbol{\lambda},\boldsymbol{\gamma})\right] - \sum_{s=1}^{K} \lambda_s \alpha_s \ ,$$

where, setting $\bar{\gamma} := \sum_{s=1}^{K} \gamma_s$, we defined the function

$$H_{(\boldsymbol{x},s)}(g,\boldsymbol{\lambda},\boldsymbol{\gamma}) = \begin{cases} 0, & \text{if } g(\boldsymbol{x},s) = r \\ \frac{p_s}{\bar{\alpha}}\eta(\boldsymbol{x},s) + \lambda_s, & \text{if } g(\boldsymbol{x},s) = 0 \\ \frac{p_s}{\bar{\alpha}}(1-\eta(\boldsymbol{x},s) - \bar{\gamma}) + \lambda_s + \frac{\gamma_s}{\alpha_s}, & \text{if } g(\boldsymbol{x},s) = 1 \end{cases} \ .$$

Using this Langrangian, our initial problem can be expressed as

$$\min_{g} \max_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K} \mathcal{L}(g,\boldsymbol{\lambda},\boldsymbol{\gamma}) \ .$$

Weak duality then implies that

$$\min_{g} \max_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K} \mathcal{L}(g,\boldsymbol{\lambda},\boldsymbol{\gamma}) \geq \max_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K} \min_{g} \mathcal{L}(g,\boldsymbol{\lambda},\boldsymbol{\gamma}) \ .$$

Dual problem. We first solve the inner minimization problem of the max min formulation for any $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$,

$$\min_{g} \mathcal{L}(g, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \ , \tag{9.5}$$

and then show that strong duality holds under our assumptions. The problem in Eq. (9.5) can be solved point-wise, that is, it is sufficient to solve

$$\min_{z \in \{0,1,r\}} H_{(\boldsymbol{x},s)}(z, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \ ,$$

for any $s \in [K]$ and any $\boldsymbol{x} \in \mathbb{R}^d$. One can easily check that, for any given couple $(\boldsymbol{x}, s)$, the minimizer of the above expression is given by

$$\tilde{g}(\boldsymbol{x}, s) = \begin{cases} r, & \text{if } 0 \leq \lambda_s + \min(\frac{p_s}{\bar{\alpha}}\eta(\boldsymbol{x}, s), \frac{p_s}{\bar{\alpha}}(1-\eta(\boldsymbol{x}, s)-\bar{\gamma})+\frac{\gamma_s}{\alpha_s}) \\ \mathbb{1}(\frac{p_s}{\bar{\alpha}}(1-2\eta(\boldsymbol{x}, s)-\bar{\gamma})+\frac{\gamma_s}{\alpha_s} < 0), & \text{otherwise} \end{cases} .$$

Note that, using the fact that $2\min(a, b) = a + b - |a - b|$, the previous expression simplifies to

$$\tilde{g}(\boldsymbol{x}, s) = \begin{cases} r, & \text{if } \left|\frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{x}, s)-\bar{\gamma})+\frac{\gamma_s}{2\alpha_s}\right| \leq \lambda_s + \frac{p_s}{2\bar{\alpha}}(1-\bar{\gamma})+\frac{\gamma_s}{2\alpha_s} \\ \mathbb{1}(\frac{p_s}{\bar{\alpha}}(1-2\eta(\boldsymbol{x}, s)-\bar{\gamma})+\frac{\gamma_s}{\alpha_s} < 0), & \text{otherwise} \end{cases} .$$

Plugging back the expression for $\tilde{g}$ in the function H we get

$$H_{(\boldsymbol{x},s)}(\tilde{g}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \left(\frac{p_s}{2\bar{\alpha}}(1-\bar{\gamma}) + \lambda_s + \frac{\gamma_s}{2\alpha_s} - \left|\frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{x}, s)-\bar{\gamma}) + \frac{\gamma_s}{2\alpha_s}\right|\right)_- \ ,$$

where $(a)_- := \min(a, 0)$. Substituting this expression into the Lagrangian, we can derive the dual optimization problem as

$$\max_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left\{ \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s} \left(\frac{p_s}{2\bar{\alpha}}(1-\bar{\gamma}) + \lambda_s + \frac{\gamma_s}{2\alpha_s} - \left|\frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{x}, s)-\bar{\gamma}) + \frac{\gamma_s}{2\alpha_s}\right|\right)_- - \sum_{s=1}^{K} \lambda_s \alpha_s \right\} \ .$$

Writing this optimization problem as a minimization problem in vector form, the optimal Lagrange multipliers $(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$ are a solution of

$$\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left\{ \sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s} \left(\left|\frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{X}, S)-\langle\boldsymbol{\gamma}, \mathbf{1}\rangle)+\frac{\boldsymbol{\gamma}_s}{2\alpha_s}\right| - \frac{p_s}{2\bar{\alpha}}(1-\langle\boldsymbol{\gamma}, \mathbf{1}\rangle)-\boldsymbol{\lambda}_s-\frac{\boldsymbol{\gamma}_s}{2\alpha_s}\right)_+ + \langle\boldsymbol{\lambda}, \boldsymbol{\alpha}\rangle \right\} \ , \tag{9.6}$$

where for any real number y, $(y)_+ := \max(x, 0)$.

Let us check that the objective function of the above optimization problem is jointly convex in $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$. First of all, the mappings

$$(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \mapsto \frac{p_s}{2\bar{\alpha}}(1 - 2\eta(\boldsymbol{x}, s) - \langle\boldsymbol{\gamma}, \mathbf{1}\rangle) + \frac{\boldsymbol{\gamma}_s}{2\alpha_s} \ ,$$
$$(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \mapsto -\frac{p_s}{2\bar{\alpha}}(1 - \langle\boldsymbol{\gamma}, \mathbf{1}\rangle) - \boldsymbol{\lambda}_s - \frac{\boldsymbol{\gamma}_s}{2\alpha_s} \ ,$$

are clearly affine mappings. Since taking the absolute value of an affine mapping gives a convex mapping (as a maximum between two affine, hence convex, functions), the sum of the absolute value of the first mapping with the second mapping is a convex function. Furthermore, the composition with the positive part function preserves convexity since this operation can be expressed as taking the maximum between two convex functions. Finally, by linearity of expectation, we notice that the objective is expressed as a finite sum of convex functions and conclude that it is jointly convex in $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$.

The objective function is not smooth everywhere due to the presence of absolute values and positive part functions. However, thanks to Assumption 9.3.1, the set of points at which the objective function is not differentiable has zero Lebesgue measure and can thus be ignored. The First-Order Optimality Conditions (FOOC) on the optimal Lagrange multipliers $(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$ then read as, for any $s \in [K]$,

$$\begin{aligned} \alpha_s &= \mathbb{P}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{X},s)-\langle\boldsymbol{\gamma}^*,\mathbf{1}\rangle)+\frac{\boldsymbol{\gamma}_s^*}{2\alpha_s}\right| \geq \frac{p_s}{2\bar{\alpha}}(1-\langle\boldsymbol{\gamma}^*,\mathbf{1}\rangle)+\boldsymbol{\lambda}_s^*+\frac{\boldsymbol{\gamma}_s^*}{2\alpha_s}\right) \\ 0 &= \sum_{s=1}^{K}\left(\frac{p_s}{\bar{\alpha}}\mathbf{1}-\frac{\boldsymbol{e}_s}{\alpha_s}\right)\mathbb{P}_{\boldsymbol{X}|S=s}\left(\min\left(2\eta(\boldsymbol{X},S),\eta(\boldsymbol{X},S)-\frac{\bar{\alpha}\lambda_s}{p_s}\right) \geq \frac{\bar{\alpha}\gamma_s}{p_s\alpha_s}+1-\bar{\gamma}\right) \quad , \end{aligned} \tag{FOOC}$$

where, for any $s \in [K]$, $\boldsymbol{e}_s$ is the s-basis vector of $\mathbb{R}^K$.

Feasibility of $\tilde{g}$ for the primal problem Let us check that $\tilde{g}$ is feasible for the primal problem. Using the definition of $\tilde{g}$ and the first-order optimal condition on $\boldsymbol{\lambda}^*$ we obtain, for any $s \in [K]$,

$$\begin{aligned} &\mathbb{P}(\tilde{g}(\boldsymbol{X},S) \neq r \mid S=s) \\ &\qquad = \mathbb{P}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{X},s)-\langle\boldsymbol{\gamma},\mathbf{1}\rangle)+\frac{\boldsymbol{\gamma}_s}{2\alpha_s}\right| \geq \frac{p_s}{2\bar{\alpha}}(1-\langle\boldsymbol{\gamma},\mathbf{1}\rangle)+\boldsymbol{\lambda}_s+\frac{\boldsymbol{\gamma}_s}{2\alpha_s}\right) \\ &\qquad = \alpha_s \quad , \end{aligned}$$

which proves that $\tilde{g}$ satisfies the first set of constraints. For the Demographic Parity constraints, one easily obtains

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{X}|S=s}(\tilde{g}(\boldsymbol{X},S)=1 \mid \tilde{g}(\boldsymbol{X},S) \neq r) \\ &\qquad = \frac{1}{\alpha_s}\mathbb{P}_{\boldsymbol{X}|S=s}(\tilde{g}(\boldsymbol{X},S)=1) \\ &\qquad = \frac{1}{\alpha_s}\mathbb{P}_{\boldsymbol{X}|S=s}\left(\min\left(2\eta(\boldsymbol{X},S),\eta(\boldsymbol{X},S)-\frac{\bar{\alpha}\lambda_s}{p_s}\right) \geq \frac{\alpha\gamma_s}{p_s\alpha_s}+1-\bar{\gamma}\right) \quad , \\ &\mathbb{P}_{(\boldsymbol{X},S)}(\tilde{g}(\boldsymbol{X},S)=1 \mid \tilde{g}(\boldsymbol{X},S) \neq r) \\ &\qquad = \sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\mathbb{P}_{\boldsymbol{X}|S=s}\left(\min\left(2\eta(\boldsymbol{X},S),\eta(\boldsymbol{X},S)-\frac{\bar{\alpha}\lambda_s}{p_s}\right) \geq \frac{\bar{\alpha}\gamma_s}{p_s\alpha_s}+1-\bar{\gamma}\right) \quad . \end{aligned}$$

The first-order optimality condition for $\boldsymbol{\gamma}^*$ guarantees that for, any $s \in [K]$,

$$\mathbb{P}_{\boldsymbol{X}|S=s}(\tilde{g}(\boldsymbol{X},S)=1 \mid \tilde{g}(\boldsymbol{X},S) \neq r) = \mathbb{P}_{(\boldsymbol{X},S)}(\tilde{g}(\boldsymbol{X},S)=1 \mid \tilde{g}(\boldsymbol{X},S) \neq r) \quad ,$$

i.e. it guarantees that the classifier $\tilde{g}$ satisfies the Demographic Parity constraint.

We conclude that the classifier $\tilde{g}$ is feasible for the primal problem and thus that strong duality holds.

9.9.2 Auxiliary results

We will need a tight control on the sup-norm of the difference between CDF and empirical CDF. The next result is (Massart 1990, Corollary 1).

Theorem 9.9.1. *Let $\boldsymbol{Z}, \boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$ be $n+1$ i.i.d. continuous random variable sampled from $\mathbb{P}$ on $\mathcal{Z}$, then for any $\delta > 0$, with probability at least $1-\delta$,*

$$\sup_{z \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Z_i \leq z\} - \mathbb{P}(Z \leq z) \right| \leq \sqrt{\frac{\log(2/\delta)}{2n}} \ .$$

9.9.3 Control of reject rate

Proposition 9.9.2. *For all $\delta \in (0,1)$, the proposed algorithm satisfies with probability at least $1-\delta$ that*

$$|\mathbb{P}(\widehat{g}(\boldsymbol{X}, S) \neq r \mid S = s) - \alpha_s| \leq \sqrt{\frac{2\log(2K/\delta)}{n_s}} + \frac{2}{n_s}, \quad \forall s \in [K] \ .$$

The rest of this section is devoted to the proof of this result. In what follows, all the derivations should be understood conditionally on $\widehat{\eta}$. In simple words, the estimator $\widehat{\eta}$ is treated as fixed and the only randomness comes from the unlabeled data. According to the definition of our estimator,

$$\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X}, s) \neq r\right) = \mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) > 0\right) \ .$$

Using the triangle inequality we can upper bound $|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) > 0\right) - \alpha_s|$ by two terms

$$\begin{aligned} &\underbrace{\left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) > 0\right) - \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) > 0\right)\right|}_{\mathtt{T}_1} \\ &\qquad + \underbrace{\left|\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) > 0\right) - \alpha_s\right|}_{\mathtt{T}_2} \ , \end{aligned} \tag{9.7}$$

which are treated separately.

Control of $\mathtt{T}_1$. The first term $\mathtt{T}_1$ can be controlled using tools from empirical process theory. One can directly observe that

$$\begin{aligned} \mathtt{T}_1 &\leq \sup_{(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K} \left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) > 0\right) - \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) > 0\right)\right| \\ &\leq \sup_{(a,b) \in \mathbb{R}\times\mathbb{R}} \left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}\widehat{\eta}(\boldsymbol{X}, S) - a\right| - a + b > 0\right) - \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}\widehat{\eta}(\boldsymbol{X}, S) - a\right| - a + b > 0\right)\right| \\ &\leq \sup_{(a,c) \in \mathbb{R}\times\mathbb{R}} \left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}\widehat{\eta}(\boldsymbol{X}, S) - a\right| > c\right) - \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}\widehat{\eta}(\boldsymbol{X}, S) - a\right| > c\right)\right| \\ &\leq 2 \sup_{a \in \mathbb{R}} \left|\mathbb{P}_{\boldsymbol{X}|S=s}(\widehat{\eta}(\boldsymbol{X}, S) \leq a) - \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}(\widehat{\eta}(\boldsymbol{X}, S) \leq a)\right| \ , \end{aligned} \tag{9.8}$$

where we used the triangle inequality and the fact that $(\widehat{\eta}(\boldsymbol{X}, S) \mid S = s)$ is a continuous random variable to obtain the last inequality.

By our assumption (see Remark 9.9.4), the random variables $\widehat{\eta}(\boldsymbol{X}_i, s), (\widehat{\eta}(\boldsymbol{X}, S) \mid S = s)$ for $i \in \mathcal{I}_s$ are i.i.d. continuous conditionally on $\widehat{\eta}$. Thus, applying Theorem 9.9.1 we conclude that with probability at least $1 - \delta$ it holds that

$$\mathtt{T}_1 \leq \sqrt{\frac{2\log(2/\delta)}{n_s}} \ . \tag{9.9}$$

Control of $\mathtt{T}_2$. The control of the second term $\mathtt{T}_2$ requires a more involved analysis. Since $\widehat{\boldsymbol{\lambda}}$ is a minimizer of (9.3), the first order optimality condition for convex non-smooth minimization problems state that for any $s \in [K]$, there exists $\rho_s \in [0, 1]$ such that

$$\alpha_s = \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) > 0\right) + \rho_s \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) = 0\right)$$

Thus, the second term of Eq. (9.7) can be bounded as

$$\left|\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) > 0\right) - \alpha_s\right| \leq \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) = 0\right) \ . \tag{9.10}$$

The control of $\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) = 0\right)$ is provided by the following result.

Lemma 9.9.3. *Assume that $(\widehat{\eta}(\boldsymbol{X}, S) \mid S = s, \widehat{\eta})$ is almost surely continuous, then for any $s \in [K]$, for any $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$,*

$$\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0\right) \leq \frac{2}{n_s} \ , \qquad a.s.$$

Proof. We recall that by definition of $\widehat{\mathbb{P}}_{\boldsymbol{X}|s}$ we have

$$\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0\right) = \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbb{1}(\widehat{G}(\boldsymbol{X}_i, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0) \ .$$

The proof goes by contradiction. Assume that the event

$$\frac{1}{n_s}\sum_{i=1}^{n_s} \mathbb{1}(\widehat{G}(\boldsymbol{X}_i, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0) \geq \frac{3}{n_s} \ ,$$

happens with positive probability. Then, there exist three indexes i_1, i_2, i_3 such that

$$\widehat{G}(\boldsymbol{X}_{i_j}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0 \ , \quad j = 1, 2, 3 \ .$$

However, $\widehat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = 0$ implies that either

$$\frac{\widehat{p}_s}{\bar{\alpha}}\widehat{\eta}(\boldsymbol{X}, s) + \boldsymbol{\lambda}_s = 0 \quad \text{or} \quad \frac{\widehat{p}_s}{\bar{\alpha}}(\widehat{\eta}(\boldsymbol{X}, s) + \langle \boldsymbol{\gamma}, 1\rangle - 1) - \boldsymbol{\lambda}_s + \frac{\boldsymbol{\gamma}_s}{\alpha_s} = 0 \ .$$

By the pigeonhole principle, there exist $i, j \in \{i_1, i_2, i_3\}, i \neq j$ such that

$$\widehat{\eta}(\boldsymbol{X}_i, s) = \widehat{\eta}(\boldsymbol{X}_j, s) \ ,$$

which contradicts our assumption that $(\widehat{\eta}(\boldsymbol{X}, S) \mid S = s, \widehat{\eta})$ is continuous almost surely.

Remark 9.9.4. *Recall that the assumption of continuity of* $(\widehat{\eta}(\boldsymbol{X}, S) \mid S = s, \widehat{\eta})$ *can always be fulfilled with the help of additional randomization. More formally, one needs to replace* $\widehat{\eta}$ *by its smoothed version using additional randomization present in Algorithm 2. To keep things simple, we avoid this technicality in our proof and simply assume that* $(\widehat{\eta}(\boldsymbol{X}, S) \mid S = s, \widehat{\eta})$ *is indeed continuous. The statement of this result is straightforwardly adapted to the perturbed version of* $\widehat{\eta}$*.*

□

Lemma 9.9.3 allows to control the second term in Eq. (9.7) yielding

$$\mathtt{T}_2 \leq \frac{2}{n_s} \ . \tag{9.11}$$

Putting together. Substituting Eqs. (9.9) and (9.11) into Eq. (9.8), we deduce that for all $s \in [K]$ we have, with probability $1 - \delta$,

$$\left| \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) > 0 \right) - \alpha_s \right| \leq \sqrt{\frac{2 \log(2/\delta)}{n_s}} + \frac{2}{n_s} \ .$$

Finally, taking the union bound we deduce that, with probability at least $1 - \delta$, we have for all $s \in [K]$

$$\left| \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{G}(\boldsymbol{X}, s, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) > 0 \right) - \alpha_s \right| \leq \sqrt{\frac{2 \log(2K/\delta)}{n_s}} + \frac{2}{n_s} \ .$$

The proof of Proposition 9.9.2 is concluded.

9.9.4 Control of Demographic Parity violation

Proposition 9.9.5. *For any* $\delta \in (0, 1)$*, the proposed algorithm satisfies with probability at least* $1 - \delta$*, for any* $s \in [K]$*,*

$$\left| \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, s) = 1 \mid \widehat{g}(\boldsymbol{X}, s) \neq r \right) - \mathbb{P}_{(\boldsymbol{X},S)} \left(\widehat{g}(\boldsymbol{X}, S) = 1 \mid \widehat{g}(\boldsymbol{X}, S) \neq r \right) \right| \leq \frac{1}{\alpha_s} v_{n_s}^{\delta, K} + \frac{1}{\bar{\alpha}} \sum_{s=1}^{K} p_s v_{n_s}^{\delta, K} \ ,$$

where

$$v_n^{\delta, K} := \left(3 \sqrt{\frac{\log(4K/\delta)}{n}} + \frac{4}{n} \right) \ .$$

Remark 9.9.6. *It is easy to see in the proof that the high-probability event on which Proposition 9.9.5 holds is contained in the high-probability event on which Proposition 9.9.2 holds.*

The rest of this section is devoted to the proof of this result.

Problem splitting. Similarly to the control of the reject rate we start by splitting our problem in several parts. Recall that our goal here is to control

$$(\mathrm{DP}^s) := \left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)=1 \mid \widehat{g}(\boldsymbol{X},s)\neq r\right) - \mathbb{P}_{(\boldsymbol{X},S)}\left(\widehat{g}(\boldsymbol{X},S)=1 \mid \widehat{g}(\boldsymbol{X},S)\neq r\right)\right| \ ,$$

for all $s \in [K]$. Triangle inequality yields that

$$\begin{aligned}
(\mathrm{DP}^s) \leq & \left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)=1 \mid \widehat{g}(\boldsymbol{X},s)\neq r\right) - \alpha_s^{-1}\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)=1\right)\right| \\
& + \left|\alpha_s^{-1}\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)=1\right) - \alpha_s^{-1}\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)=1\right)\right| \\
& + \left|\alpha_s^{-1}\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)=1\right) - \bar{\alpha}^{-1}\sum_{s\in[K]} p_s\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)=1\right)\right| \\
& + \left|\bar{\alpha}^{-1}\sum_{s\in[K]} p_s\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)=1\right) - \bar{\alpha}^{-1}\sum_{s\in[K]} p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)=1\right)\right| \\
& + \left|\bar{\alpha}^{-1}\sum_{s\in[K]} p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)=1\right) - \mathbb{P}_{(\boldsymbol{X},S)}\left(\widehat{g}(\boldsymbol{X},S)=1 \mid \widehat{g}(\boldsymbol{X},S)\neq r\right)\right| \ .
\end{aligned}$$

The second and the fourth terms will be controlled using empirical process theory. We can get a bound on the first and fifth terms through our control of the reject rate. The third term is controlled via the first-order optimality condition on $\boldsymbol{\gamma}$.

High-probability event. Let us describe in details the high-probability event on which we will place ourselves for controlling all the terms, uniformly over the classes $s \in [K]$.

Proposition 9.9.2 states that there exists an event $\mathbf{R}$ that holds with probability at least $1 - K\delta$ and on which, for any $\delta \in (0, 1/K)$, the proposed algorithm satisfies with probability at least $1 - K\delta$ that

$$|\mathbb{P}(\widehat{g}(\boldsymbol{X},S)\neq r \mid S=s) - \alpha_s| \leq u^\delta_{n_s}, \quad \forall s \in [K] \ ,$$

where

$$u^\delta_n := \sqrt{\frac{2\log(2/\delta)}{n}} + \frac{2}{n} \ , \quad \forall n \geq 1 \ .$$

Furthermore, for any class $s \in [K]$, using the fact that the random variable $(\eta(\boldsymbol{X},S) \mid S=s)$ is continuous, the event

$$\mathrm{EP}_s := \left\{\sup_{a\in\mathbb{R}}\left|\mathbb{P}_{\boldsymbol{X}|S=s}\left(\eta(\boldsymbol{X},s)>a\right) - \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\eta(\boldsymbol{X},s)>a\right)\right| \leq \sqrt{\frac{\log(2/\delta)}{2n_s}}\right\} \ ,$$

holds with probability at least $1-\delta$ (see Theorem 9.9.1). By a simple union bound argument, the intersection of those events, denoted by $\mathrm{EP} := \cap_{s\in[K]}\mathrm{EP}_s$, then holds with probability at least $1 - 2K\delta$.

In what follows we place ourselves on the event $\mathbf{A} := \mathbf{R} \cap \mathrm{EP}$ which holds with probability at least $1 - 2K\delta$.

First-order optimality condition for $\widehat{\boldsymbol{\gamma}}$. Recall that $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}})$ is a solution of

$$\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left[\langle \boldsymbol{\lambda}, \boldsymbol{\alpha} \rangle + \widehat{\mathbb{E}}_{\boldsymbol{X}|S=s}(\widehat{G}(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}))_+ \right] \ ,$$

where the function $\widehat{G}$ is defined as

$$\widehat{G}(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \left| \frac{\widehat{p}_s}{2\bar{\alpha}}(1 - 2\widehat{\eta}(\boldsymbol{x}, s) - \langle \boldsymbol{\gamma}, \mathbf{1} \rangle) + \frac{\boldsymbol{\gamma}_s}{2\alpha_s} \right| - \frac{\widehat{p}_s}{2\bar{\alpha}}(1 - \langle \boldsymbol{\gamma}, \mathbf{1} \rangle) - \boldsymbol{\lambda}_s - \frac{\boldsymbol{\gamma}_s}{2\alpha_s} \ .$$

The positive part of $\widehat{G}$ can be expressed as

$$(\widehat{G}(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}))_+ = \max(0, m_+(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}), m_-(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma})) \ ,$$

where

$$m_+(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = -\frac{p_s}{\bar{\alpha}}\widehat{\eta}(\boldsymbol{x}, s) - \lambda_s$$

and

$$m_-(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \frac{p_s}{\bar{\alpha}}(\widehat{\eta}(\boldsymbol{x}, s) + \langle \boldsymbol{\gamma}, 1 \rangle - 1) - \frac{\boldsymbol{\gamma}_s}{\alpha_s} - \lambda_s \ .$$

Noticing that the event $m_-(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) > \max(0, m_+(\boldsymbol{x}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}))$ is the same as the event $\widehat{g}(\boldsymbol{X}, s) = 1$, the first-order optimality condition on $\widehat{\boldsymbol{\gamma}}$ reads as: $\exists (\rho_s)_{s=1}^K \in [0,1]^K$ s.t.

$$\sum_{s=1}^{K} \left(\frac{p_s}{\bar{\alpha}} \mathbf{1} - \frac{1}{\alpha_s} \boldsymbol{e}_s \right) \left(\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, s) = 1 \right) + \rho_s \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\Delta_s(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) \right) \right) = 0 \ ,$$

where we define the event $\Delta_s(\boldsymbol{\lambda}, \boldsymbol{\gamma}) := \{ m_-(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \max(0, m_+(\boldsymbol{X}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma})) \}$. In scalar form the previous condition can be expressed as: for any $s \in [K]$, there exists $\rho_s \in [0,1]$ such that

$$\begin{aligned} \sum_{s=1}^{K} \frac{p_s}{\bar{\alpha}} \Big(\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, s) = 1 \right) + \rho_s \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\Delta_s(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) \right) \Big) \\ = \frac{1}{\alpha_s} \left(\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, s) = 1 \right) + \rho_s \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\Delta_s(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) \right) \right) \ . \end{aligned}$$

Control of the first term. Re-arranging terms and using the fact that

$$\mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, S) = 1 \right) \leq \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, S) \neq r \right),$$

we get

$$\begin{aligned} (\mathrm{DP}_1^s) &:= \left| \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, s) = 1 \mid \widehat{g}(\boldsymbol{X}, s) \neq r \right) - \alpha_s^{-1} \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, S) = 1 \right) \right| \\ &= \left| \frac{1}{\alpha_s} - \frac{1}{\mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, s) \neq r \right)} \right| \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, S) = 1 \right) \\ &\leq \left| \frac{1}{\alpha_s} - \frac{1}{\mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, s) \neq r \right)} \right| \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, S) \neq r \right) \\ &= \frac{1}{\alpha_s} \left| \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, S) \neq r \right) - \alpha_s \right| \ . \end{aligned}$$

Considering that we restrict ourselves to the high-probability event $\mathbf{A}$, we can conclude that

$$(\mathrm{DP}_1^s) \leq \frac{u_{n_s}^{\delta}}{\alpha_s} \ .$$

Control of the second term The second term is given by the empirical process

$$(\mathrm{DP}_2^s) := \alpha_s^{-1} \left| \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, S) = 1 \right) - \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, S) = 1 \right) \right| \ .$$

The event $\{\widehat{g}(\boldsymbol{X}, s) = 1\}$ is the same as the event

$$\left\{ \left| \frac{p_s}{2\bar{\alpha}} (1 - 2\widehat{\eta}(\boldsymbol{X}, s) - \langle \mathbf{1}, \widehat{\boldsymbol{\gamma}} \rangle) + \frac{\widehat{\gamma}_s}{2\alpha_s} \right| > \frac{p_s}{2\bar{\alpha}} (1 - \langle \mathbf{1}, \boldsymbol{\gamma} \rangle) + \widehat{\boldsymbol{\lambda}}_s + \frac{\widehat{\gamma}_s}{2\alpha_s}, \right.$$
$$\left. 2\widehat{\eta}(\boldsymbol{X}, s) \geq 1 + \frac{\bar{\alpha}\widehat{\gamma}_s}{\alpha_s p_s} - \langle \widehat{\boldsymbol{\gamma}}, \mathbf{1} \rangle \right\} \ ,$$

which can be compacted to

$$S(\boldsymbol{\lambda}, \boldsymbol{\gamma}) := \left\{ \widehat{\eta}(\boldsymbol{X}, s) > \max \left(\frac{1}{2} + \frac{\bar{\alpha}\widehat{\gamma}_s}{2\alpha_s p_s} - \frac{1}{2} \langle \widehat{\boldsymbol{\gamma}}, \mathbf{1} \rangle, \frac{\bar{\alpha}}{p_s} \left(\widehat{\boldsymbol{\lambda}}_s + \frac{\widehat{\gamma}_s}{\alpha_s} \right) + 1 - \langle \mathbf{1}, \boldsymbol{\gamma} \rangle \right) \right\} \ .$$

Following this observation, we can express the second term as

$$\begin{aligned}
(\mathrm{DP}_2^s) &= \alpha_s^{-1} \sup_{(\boldsymbol{\lambda}, \boldsymbol{\gamma})} \left| \mathbb{P}_{\boldsymbol{X}|S=s} \left(S(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \right) - \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(S(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \right) \right| \\
&\leq \alpha_s^{-1} \sup_{a \in \mathbb{R}} \left| \mathbb{P}_{\boldsymbol{X}|S=s} \left(\widehat{\eta}(\boldsymbol{X}, s) > a \right) - \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\widehat{\eta}(\boldsymbol{X}, s) > a \right) \right| \ .
\end{aligned}$$

Since we are on the event $\mathbf{A}$ which is contained in the event EP_s, we have

$$(\mathrm{DP}_2^s) \leq \frac{1}{\alpha_s} \sqrt{\frac{\log(2/\delta)}{2n_s}} \ .$$

Control of the third term. The third term can be controlled with the first-order optimality condition on $\widehat{\boldsymbol{\gamma}}$ and multiple triangle inequalities as

$$\begin{aligned}
(\mathrm{DP}_3^s) &:= \left| \alpha_s^{-1} \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, S) = 1 \right) - \bar{\alpha}^{-1} \sum_{s \in [K]} p_s \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\widehat{g}(\boldsymbol{X}, S) = 1 \right) \right| \\
&= \left| \frac{\rho_s}{\alpha_s} \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\Delta_s(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) \right) - \sum_{s=1}^{K} \frac{p_s}{\bar{\alpha}} \rho_s \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\Delta_s(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) \right) \right| \\
&\leq \frac{1}{\alpha_s} \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\Delta_s(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) \right) + \sum_{s=1}^{K} \frac{p_s}{\bar{\alpha}} \widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\Delta_s(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) \right) \ .
\end{aligned}$$

The following lemma gives an almost sure upper bound on $\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\Delta_s(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) \right)$ for any $s \in [K]$.

Lemma 9.9.7. *Assume that $(\widehat{\eta}(\boldsymbol{X}, S) \mid S = s, \widehat{\eta})$ is almost surely continuous, then for any $s \in [K]$, for any $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$,*

$$\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s} \left(\Delta_s(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}) \right) \leq \frac{2}{n_s}, \qquad a.s.$$

Proof. This proof is similar to proof of Lemma 9.9.3. Assume by contradiction that the stated bound is not true. Then, it happens with positive probability that

$$\frac{1}{n_s}\sum_{i=1}^{n_s}\mathbb{1}\left\{m_-(\boldsymbol{X}_i, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \max(0, m_+(\boldsymbol{X}_i, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}))\right\} \geq \frac{3}{n_s} \ ,$$

which implies that there exist a triplet i_1, i_2, i_3 such that

$$m_-(\boldsymbol{X}_{i_j}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \max(0, m_+(\boldsymbol{X}_{i_j}, s, \boldsymbol{\lambda}, \boldsymbol{\gamma})), \quad \text{for } j = 1, 2, 3 \ .$$

By the pigeonhole principle, there must exist a couple $(i, j), i \neq j$ among this triplet such that either

$$m_-(\boldsymbol{X}_i, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = m_-(\boldsymbol{X}_j, s, \boldsymbol{\lambda}, \boldsymbol{\gamma})$$

or

$$m_-(\boldsymbol{X}_i, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) - m_+(\boldsymbol{X}_i, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = m_-(\boldsymbol{X}_j, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) - m_+(\boldsymbol{X}_j, s, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \ .$$

In both cases one must have $\widehat{\eta}(\boldsymbol{X}_i, s) = \widehat{\eta}(\boldsymbol{X}_j, s)$ which happens with probability 0 by the continuity assumption and leads to a contradiction. The proof of lemma is concluded. □

Plugging in the bounds from Lemma 9.9.7 yields

$$(\mathrm{DP}_3^s) \leq \frac{2}{n_s\alpha_s} + \frac{2}{\bar{\alpha}}\sum_{s=1}^{K}\frac{p_s}{n_s} \ .$$

Control of the fourth term. The fourth term can be seen as a sum of empirical processes:

$$\begin{aligned}(\mathrm{DP}_4) &:= \bar{\alpha}^{-1}\left|\sum_{s\in[K]} p_s\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X}, S) = 1\right) - \sum_{s\in[K]} p_s\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X}, S) = 1\right)\right| \\ &\leq \bar{\alpha}^{-1}\sum_{s=1}^{K} p_s\left|\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X}, S) = 1\right) - \mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X}, S) = 1\right)\right| \ .\end{aligned}$$

We can control the fourth term from the bound we have on the second term (which holds uniformly over the classes s) as

$$(\mathrm{DP}_4) \leq \frac{1}{\bar{\alpha}}\sum_{s\in K} p_s\sqrt{\frac{\log(2/\delta)}{2n_s}} \ .$$

Control of the fifth term. Finally, the fifth term can be bounded using the same trick as for the first term.

$$\begin{aligned}
(\mathrm{DP}_5) &:= \left| \bar{\alpha}^{-1} \sum_{s\in[K]} p_s \mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},S)=1\right) - \mathbb{P}_{(\boldsymbol{X},S)}\left(\widehat{g}(\boldsymbol{X},s)=1 \mid \widehat{g}(\boldsymbol{X},s)\neq r\right)\right| \\
&= \left| \frac{1}{\bar{\alpha}} - \frac{1}{\sum_{s=1}^K p_s \mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)\neq r\right)} \right| \sum_{s=1}^K p_s \mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},S)=1\right) \\
&\leq \left| \frac{1}{\bar{\alpha}} - \frac{1}{\sum_{s=1}^K p_s \mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)\neq r\right)} \right| \sum_{s=1}^K p_s \mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},S)\neq r\right) \\
&= \frac{1}{\bar{\alpha}} \left| \sum_{s=1}^K p_s (\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X},s)\neq r\right) - \alpha_s) \right| \leq \frac{1}{\bar{\alpha}} \sum_{s=1}^K p_s u_{n_s}^{\delta}.
\end{aligned}$$

Summary. Putting everything together, we have shown that, on the event $\mathbf{A}$ which holds with probability at least $1-2K\delta$, we have, for any $s\in[K]$,

$$(\mathrm{DP}^s) \leq \frac{1}{\alpha_s}\left(3\sqrt{\frac{\log(2/\delta)}{2n_s}}+\frac{4}{n_s}\right) + \frac{2}{\bar{\alpha}}\sum_{s=1}^K p_s \left(3\sqrt{\frac{\log(2/\delta)}{2n_s}}+\frac{4}{n_s}\right) \quad .$$

9.9.5 Control of the excess risk

Define the sequence

$$u_n^{\delta,K} := \sqrt{\frac{2\log(4K/\delta)}{n}} + \frac{2}{n} \quad , \quad \forall n \geq 1 \quad .$$

We state and prove slightly more precise bound then the one presented in the main body.

Proposition 9.9.8. *Assume that $u_{n_s}^{\delta,K} < \alpha_s < 1 - \frac{2}{n_s}$ for any $s\in[K]$ and that Assumption 9.3.1 holds. Then, for any $\delta\in(0,1)$, the excess risk of the post-processing classifier with abstention $\widehat{g}$ defined in Eq (9.2) satisfies, with probability at least $1-\delta$,*

$$\mathcal{E}(\widehat{g}) \leq \left(\frac{1}{\bar{\alpha}} + \frac{1}{\bar{\alpha} - \sum_s p_s u_{n_s}^{\delta,K}}\right) \|\eta - \widehat{\eta}\|_1 + 6\sum_{s=1}^K \left(\frac{p_s}{\bar{\alpha}} + \frac{1}{\alpha_s}\right) u_{n_s}^{\delta,K} \quad .$$

A quick inspection of the proof shows that the high-probability event on which the stated bound holds is the same as the event on which Proposition 9.9.5 holds, which is contained in the event on which Proposition 9.9.2 holds. Thus we can control the excess risk and the violation of the constraints on the same high-probability event.

Proof. Since, using Assumption 9.3.1 we have established strong duality, the following equality holds

$$\mathcal{R}(g^*) = \max_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left\{ \sum_{s=1}^K \mathbb{E}_{\boldsymbol{X}|S=s}\left(\frac{p_s}{2\bar{\alpha}}(1-\bar{\gamma}) + \lambda_s + \frac{\gamma_s}{2\alpha_s} - \left| \frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{x},s)-\bar{\gamma}) + \frac{\gamma_s}{2\alpha_s} \right| \right)_- - \sum_{s=1}^K \lambda_s \alpha_s \right\}. \tag{9.12}$$

Besides, we can control the risk of any classifier g as

$$\begin{aligned}\mathcal{R}(g) &= \sum_{s=1}^{K} \frac{p_s}{\mathbb{P}(g(\boldsymbol{X},S)\neq r)}\mathbb{E}_{\boldsymbol{X}|S=s}\left[(1-\eta(\boldsymbol{X},s)\mathbb{1}_{g(\boldsymbol{X},s)=1}+\eta(\boldsymbol{X},s)\mathbb{1}_{g(\boldsymbol{X},s)=0}\right] \\ &\leq \sum_{s=1}^{K} \frac{p_s}{\mathbb{P}(g(\boldsymbol{X},S)\neq r)}\mathbb{E}_{\boldsymbol{X}|S=s}\left[(1-\widehat{\eta}(\boldsymbol{X},s)\mathbb{1}_{g(\boldsymbol{X},s)=1}+\widehat{\eta}(\boldsymbol{X},s)\mathbb{1}_{g(\boldsymbol{X},s)=0}\right] \\ &\quad + \frac{\|\eta-\widehat{\eta}\|_1}{\mathbb{P}(g(\boldsymbol{X},S)\neq r)} \ .\end{aligned} \tag{9.13}$$

Setting $\mathtt{A}_s(g) := \frac{p_s}{\bar{\alpha}}\mathbb{E}_{\boldsymbol{X}|S=s}\left[(1-\widehat{\eta}(\boldsymbol{X},s))\mathbb{1}_{g(\boldsymbol{X},s)=1}+\widehat{\eta}(\boldsymbol{X},s)\mathbb{1}_{g(\boldsymbol{X},s)=0}\right]$, we have for any classifier g,

$$\mathcal{R}(g) \leq \sum_{s=1}^{K}\mathtt{A}_s(g) + \frac{\|\eta-\widehat{\eta}\|_1}{\mathbb{P}(g(\boldsymbol{X},S)\neq r)} + \frac{1}{\bar{\alpha}}\left|\mathbb{P}(g(\boldsymbol{X},S)\neq r)-\bar{\alpha}\right| \ .$$

In what follows we bound $\mathrm{r}_1(g) := \sum_{s=1}^{K}\mathtt{A}_s(g)$. Re-arranging terms we trivially have

$$\begin{aligned}\mathrm{r}_1(g) &= \sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\mathbb{E}_{\boldsymbol{X}|S=s}\left[(1-\widehat{\eta}(\boldsymbol{X},S))\mathbb{1}_{g(\boldsymbol{X},S)=1}+\widehat{\eta}(\boldsymbol{X},S)\mathbb{1}_{g(\boldsymbol{X},S)=0}\right] \\ &\pm \sum_{s=1}^{K}\widehat{\lambda}_s\left\{\mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}+\mathbb{1}_{g(\boldsymbol{X},S)=0}\right]-\alpha_s\right\} \\ &\pm \sum_{s=1}^{K}\frac{\widehat{\gamma}_s}{\alpha_s}\mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}\right] - \left(\sum_{s'=1}^{K}\widehat{\gamma}_{s'}\right)\left(\sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}\right]\right) \\ &= \sum_{s=1}^{K}\mathbb{E}_{\boldsymbol{X}|S=s}\left[\widehat{H}_{(\boldsymbol{X},s)}(g,\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\gamma}})\right] \\ &- \sum_{s=1}^{K}\widehat{\lambda}_s\alpha_s - \sum_{s=1}^{K}\widehat{\lambda}_s\left\{\mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}+\mathbb{1}_{g(\boldsymbol{X},S)=0}\right]-\alpha_s\right\} \\ &- \sum_{s=1}^{K}\frac{\widehat{\gamma}_s}{\alpha_s}\mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}\right] - \left(\sum_{s'=1}^{K}\widehat{\gamma}_{s'}\right)\left(\sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\mathbb{E}_{\boldsymbol{X}|S=s}\left[\mathbb{1}_{g(\boldsymbol{X},S)=1}\right]\right) \ ,\end{aligned}$$

where

$$\widehat{H}_{(\boldsymbol{x},s)}(g,\boldsymbol{\lambda},\boldsymbol{\gamma}) = \begin{cases} 0, & \text{if } g(\boldsymbol{x},s)=r \\ \frac{p_s}{\bar{\alpha}}\widehat{\eta}(\boldsymbol{x},s)+\lambda_s, & \text{if } g(\boldsymbol{x},s)=0 \\ \frac{p_s}{\bar{\alpha}}(1-\widehat{\eta}(\boldsymbol{x},s)-\bar{\gamma})+\lambda_s+\frac{\gamma_s}{\alpha_s}, & \text{if } g(\boldsymbol{x},s)=1 \end{cases} ,$$

with $\bar{\gamma} = \sum_{s=1}^{K}\gamma_s$. Note that, by the definition of $\widehat{g}$, it holds that

$$\sum_{s=1}^{K}\mathbb{E}_{\boldsymbol{X}|S=s}\left[\widehat{H}_{(\boldsymbol{X},s)}(\widehat{g},\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\gamma}})\right] = \mathbb{E}(-\widehat{G}(\boldsymbol{X},s,\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\gamma}}))_- \ .$$

Thus, it holds that

$$\begin{aligned}\mathrm{r}_1(\widehat{g}) = &\sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s}\left(\frac{p_s}{2\bar{\alpha}}(1-\bar{\widehat{\gamma}}) + \widehat{\lambda}_s + \frac{\widehat{\gamma}_s}{2\alpha_s} - \left|\frac{p_s}{2\bar{\alpha}}(1-2\widehat{\eta}(\boldsymbol{X},s)-\bar{\widehat{\gamma}}) + \frac{\widehat{\gamma}_s}{2\alpha_s}\right|\right)_- - \sum_{s=1}^{K}\widehat{\lambda}_s\alpha_s \\ &- \sum_{s=1}^{K}\widehat{\lambda}_s\left(\mathbb{P}(\widehat{g}(\boldsymbol{X},S)\neq r \mid S=s) - \alpha_s\right) \\ &- \sum_{s=1}^{K}\widehat{\gamma}_s\left(\frac{\mathbb{P}(\widehat{g}(\boldsymbol{X},S)=1\mid S=s)}{\alpha_s} - \sum_{s'=1}^{K}\frac{p_{s'}}{\bar{\alpha}}\mathbb{P}(\widehat{g}(\boldsymbol{X},S)=1\mid S=s')\right) \quad .\end{aligned} \tag{9.14}$$

Finally, substituting Eq. (9.14) into Eq. (9.13) we obtain the following upper bound on $\mathcal{R}(\widehat{g})$

$$\begin{aligned}\mathcal{R}(\widehat{g}) \leq &\sum_{s=1}^{K} \mathbb{E}_{\boldsymbol{X}|S=s}\left(\frac{p_s}{2\bar{\alpha}}(1-\bar{\widehat{\gamma}}) + \widehat{\lambda}_s + \frac{\widehat{\gamma}_s}{2\alpha_s} - \left|\frac{p_s}{2\bar{\alpha}}(1-2\widehat{\eta}(\boldsymbol{X},s)-\bar{\widehat{\gamma}}) + \frac{\widehat{\gamma}_s}{2\alpha_s}\right|\right)_- - \sum_{s=1}^{K}\widehat{\lambda}_s\alpha_s \\ &- \sum_{s=1}^{K}\widehat{\lambda}_s\left(\mathbb{P}(\widehat{g}(\boldsymbol{X},S)\neq r \mid S=s) - \alpha_s\right) \\ &- \sum_{s=1}^{K}\widehat{\gamma}_s\left(\frac{\mathbb{P}(\widehat{g}(\boldsymbol{X},S)=1\mid S=s)}{\alpha_s} - \sum_{s'=1}^{K}\frac{p_{s'}}{\bar{\alpha}}\mathbb{P}(\widehat{g}(\boldsymbol{X},S)=1\mid S=s')\right) \\ &+ \frac{\|\eta-\widehat{\eta}\|_1}{\mathbb{P}(\widehat{g}(\boldsymbol{X},S)\neq r)} + \frac{1}{\bar{\alpha}}\left|\mathbb{P}(\widehat{g}(\boldsymbol{X},S)\neq r) - \bar{\alpha}\right| \quad ,\end{aligned}$$

which holds almost surely.

Define the excess risk $\mathcal{E}(\widehat{g}) := \mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)$. Note that, using the fact that mapping $x \mapsto (x)_-$ is 1-Lipschitz followed by the triangle inequality, the difference

$$\begin{aligned}\Bigg|\left(\frac{p_s}{2\bar{\alpha}}(1-\bar{\gamma}) + \lambda_s + \frac{\gamma_s}{2\alpha_s} - \left|\frac{p_s}{2\bar{\alpha}}(1-2\widehat{\eta}(\boldsymbol{x},s)-\bar{\gamma}) + \frac{\gamma_s}{2\alpha_s}\right|\right)_- \\ - \left(\frac{p_s}{2\bar{\alpha}}(1-\bar{\gamma}) + \lambda_s + \frac{\gamma_s}{2\alpha_s} - \left|\frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{x},s)-\bar{\gamma}) + \frac{\gamma_s}{2\alpha_s}\right|\right)_-\Bigg| \quad ,\end{aligned}$$

can be upper bounded by $\frac{p_s}{\bar{\alpha}}\left|\widehat{\eta}(\boldsymbol{x},s) - \eta(\boldsymbol{x},s)\right|$, for any $(\boldsymbol{x},s,\boldsymbol{\lambda},\boldsymbol{\gamma})$. Thus, replacing $(\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*)$ by $(\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\gamma}})$ in the expression for $\mathcal{R}(g^*)$ in Eq. (9.12) we obtain

$$\begin{aligned}\mathcal{E}(\widehat{g}) \leq &\frac{\|\eta-\widehat{\eta}\|_1}{\bar{\alpha}} + \frac{1}{\bar{\alpha}}\left|\mathbb{P}(g(\boldsymbol{X},S)\neq r) - \bar{\alpha}\right| + \frac{\|\eta-\widehat{\eta}\|_1}{\mathbb{P}(g(\boldsymbol{X},S)\neq r)} \\ &+ \sum_{s=1}^{K}\left|\widehat{\lambda}_s\right|\left|\mathbb{P}(\widehat{g}(\boldsymbol{X},S)\neq r\mid S=s) - \alpha_s\right| \\ &+ \sum_{s=1}^{K}\left|\widehat{\gamma}_s\right|\left|\frac{\mathbb{P}(\widehat{g}(\boldsymbol{X},S)=1\mid S=s)}{\alpha_s} - \sum_{s'=1}^{K}\frac{p_{s'}}{\bar{\alpha}}\mathbb{P}(\widehat{g}(\boldsymbol{X},S)=1\mid S=s')\right| \quad .\end{aligned} \tag{9.15}$$

In the above inequality we can control all the terms.

Indeed, using the fact that on the event of Proposition 9.9.5 we have, with probability at least $1-2K\delta$, for all $s \in [K]$,

$$\left|\mathbb{P}(\widehat{g}(\boldsymbol{X},S)\neq r\mid S=s) - \alpha_s\right| \leq u^{\delta}_{n_s}, \quad \text{with} \quad u^{\delta}_n := \sqrt{\frac{2\log(2/\delta)}{n}} + \frac{2}{n} \quad , \quad \forall n \geq 1 \quad ,$$

we deduce that with probability at least $1 - 2K\delta$ the following three inequalities hold

$$\begin{aligned}
&\frac{1}{\bar{\alpha}} \left| \mathbb{P}(g(\boldsymbol{X}, S) \neq r) - \bar{\alpha} \right| \leq \frac{1}{\bar{\alpha}} \sum_{s=1}^{K} p_s u_{n_s}^{\delta} \ , \\
&\frac{\|\eta - \widehat{\eta}\|_1}{\mathbb{P}(\widehat{g}(\boldsymbol{X}, S) \neq r)} \leq \frac{\|\eta - \widehat{\eta}\|_1}{\bar{\alpha} - \sum_s p_s u_{n_s}^{\delta}} \ , \\
&\sum_{s=1}^{K} |\widehat{\lambda}_s| \left| \mathbb{P}(\widehat{g}(\boldsymbol{X}, S) \neq r \mid S = s) - \alpha_s \right| \leq \sum_{s=1}^{K} |\widehat{\lambda}_s| u_{n_s}^{\delta} \ .
\end{aligned} \tag{9.16}$$

Note that by the assumption of the proposition, the term $\bar{\alpha} - \sum_s p_s u_{n_s}^{\delta} > 0$.

Furthermore, on the same event, using the notations of the proof of Proposition 9.9.5, we have for any $s \in [K]$

$$\begin{aligned}
\left| \frac{\mathbb{P}_{\boldsymbol{X}|S=s}\left(\widehat{g}(\boldsymbol{X}, S) = 1\right)}{\alpha_s} - \sum_{s'=1}^{K} \frac{p_{s'}}{\bar{\alpha}} \mathbb{P}_{\boldsymbol{X}|S=s'}(\widehat{g}(\boldsymbol{X}, S) = 1) \right| &\leq (\mathrm{DP}_2^s) + (\mathrm{DP}_3^s) + (\mathrm{DP}_4) \\
&\leq \frac{1}{\alpha_s} v_{n_s}^{\delta} + \frac{2}{\bar{\alpha}} \sum_{s=1}^{K} p_s v_{n_s}^{\delta} \ .
\end{aligned} \tag{9.17}$$

where $v_n^{\delta} = \sqrt{\frac{\log(2/\delta)}{2n}} + \frac{2}{n}$. All in all, substituting Eqs. (9.16) and (9.17) into Eq. (9.15) we deduce that

$$\begin{aligned}
\mathcal{E}(\widehat{g}) &\leq \left(\frac{1}{\bar{\alpha}} + \frac{1}{\bar{\alpha} - \sum_s p_s u_{n_s}^{\delta}} \right) \|\eta - \widehat{\eta}\|_1 + \sum_{s=1}^{K} \left(\frac{p_s}{\bar{\alpha}} + \left|\widehat{\lambda}_s\right| \right) u_{n_s}^{\delta} + \sum_{s=1}^{K} \left(\frac{|\widehat{\gamma}_s|}{\alpha_s} + \frac{2p_s}{\bar{\alpha}} (\sum_{s'} |\widehat{\gamma}_{s'}|) \right) v_{n_s}^{\delta} \\
&= \left(\frac{1}{\bar{\alpha}} + \frac{1}{\bar{\alpha} - \sum_s p_s u_{n_s}^{\delta}} \right) \|\eta - \widehat{\eta}\|_1 + \sum_{s=1}^{K} \left(\frac{2p_s}{\bar{\alpha}} + 2\left|\widehat{\lambda}_s\right| + \frac{|\widehat{\gamma}_s|}{\alpha_s} + \frac{2p_s}{\bar{\alpha}} (\sum_{s'} |\widehat{\gamma}_s|) \right) \sqrt{\frac{\log(1/\delta)}{2n_s}} \\
&\quad + \sum_{s=1}^{K} \left(\frac{p_s}{\bar{\alpha}} + |\widehat{\lambda}_s| + \frac{|\widehat{\gamma}_s|}{\alpha_s} + \frac{2p_s}{\bar{\alpha}} (\sum_{s'} |\widehat{\gamma}_{s'}|) \right) \frac{2}{n_s} \ .
\end{aligned}$$

In order to finish the proof it remains to provide a bound on $|\widehat{\lambda}_s|$ and $|\widehat{\gamma}_s|$. Proposition 9.9.9, proven below, establishes this bound and yields

$$\begin{aligned}
\mathcal{E}(\widehat{g}) &\leq \left(\frac{1}{\bar{\alpha}} + \frac{1}{\bar{\alpha} - \sum_s p_s u_{n_s}^{\delta}} \right) \|\eta - \widehat{\eta}\|_1 + \sum_{s=1}^{K} \left[\left(\frac{4p_s}{\bar{\alpha}} + \frac{3}{\alpha_s} \right) \sqrt{\frac{2\log(2/\delta)}{n_s}} + \left(\frac{6}{\alpha_s} + \frac{6p_s}{\bar{\alpha}} \right) \frac{2}{n_s} \right] \\
&\leq \left(\frac{1}{\bar{\alpha}} + \frac{1}{\bar{\alpha} - \sum_s p_s u_{n_s}^{\delta}} \right) \|\eta - \widehat{\eta}\|_1 + 6 \sum_{s=1}^{K} \left(\frac{p_s}{\bar{\alpha}} + \frac{1}{\alpha_s} \right) u_{n_s}^{\delta} \ .
\end{aligned}$$

the proof is concluded after the observation that thanks to our assumption we have $\bar{\alpha} - \sum_s p_s u_{n_s}^{\delta} \geq \bar{\alpha}/2$. □

Boundedness of optimal parameters

Proposition 9.9.9. *The minimization problem in Eq. (9.6) admits a global minimizer* $(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$ *which satisfies*

$$\|\boldsymbol{\gamma}^*\|_1 \leq 2 \quad \textit{and} \quad |\lambda_s^*| \leq \frac{p_s}{\bar{\alpha}} \vee \frac{|\gamma_s^*|}{\alpha_s} \ .$$

Furthermore, if for any s, $n_s > \frac{2}{\alpha_s \wedge (1-\alpha_s)}$ *and* $\widehat{\eta}(\cdot, s) \in [0,1]$, *the same holds for Eq. (9.3), that is,*

$$\|\widehat{\boldsymbol{\gamma}}\|_1 \leq 2 \quad \textit{and} \quad |\widehat{\lambda}_s| \leq \frac{p_s}{\bar{\alpha}} \vee \frac{|\widehat{\gamma}_s|}{\alpha_s} \ .$$

Proof. We denote the conditional expectation of Y given $S = s$ by $\eta(s)$. Denote by $H(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ the objective function of the minimization problem in Eq. (9.6).

Existence of global minizer. Fix arbitrary $(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K$ such that $\sum_{s=1}^K \gamma_s = 0$. Since the function $x \mapsto (|x| - b)_+$ is convex for any $b \in \mathbb{R}$ we can lower bound $H(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ using Jensen's inequality as

$$\begin{aligned} H(\boldsymbol{\lambda}, \boldsymbol{\gamma}) &= \frac{1}{2} \sum_{s=1}^K \frac{1}{\alpha_s} \mathbb{E}_{\boldsymbol{X}|S=s} \left(\left| \frac{\alpha_s p_s}{\bar{\alpha}} (1 - 2\eta(\boldsymbol{X}, s)) + \gamma_s \right| - \frac{p_s \alpha_s}{\bar{\alpha}} - 2\alpha_s \lambda_s - \gamma_s \right)_+ + \sum_{s=1}^K \lambda_s \alpha_s \\ &\geq \frac{1}{2} \sum_{s=1}^K \frac{1}{\alpha_s} \left(\left| \frac{\alpha_s p_s}{\bar{\alpha}} (1 - 2\eta(s)) + \gamma_s \right| - \frac{p_s \alpha_s}{\bar{\alpha}} - 2\alpha_s \lambda_s - \gamma_s \right)_+ + \sum_{s=1}^K \lambda_s \alpha_s \ . \end{aligned}$$

Furthermore, since $\alpha_s \leq 1$ for any s and by assumption, $\bar{\gamma} = 0$, we can further lower bound $H(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ as

$$\begin{aligned} H(\boldsymbol{\lambda}, \boldsymbol{\gamma}) &\geq \frac{1}{2} \sum_{s=1}^K \left(\left| \frac{\alpha_s p_s}{\bar{\alpha}} (1 - 2\eta(s)) + \gamma_s \right| - \frac{p_s \alpha_s}{\bar{\alpha}} - 2\alpha_s \lambda_s - \gamma_s \right)_+ + \sum_{s=1}^K \lambda_s \alpha_s \\ &\geq \frac{1}{2} \left(\|\boldsymbol{\gamma}\|_1 - \sum_{s=1}^K \frac{\alpha_s p_s}{\bar{\alpha}} |(1 - 2\eta(s))| - 1 - 2 \sum_{s=1}^K \lambda_s \alpha_s \right)_+ + \sum_{s=1}^K \lambda_s \alpha_s \\ &\geq \frac{\|\boldsymbol{\gamma}\|_1}{2} - 1 \ , \end{aligned} \tag{9.18}$$

where we used the triangle inequality for the second inequality and we lower bounded the positive part by the number itself and upper bounded $|1 - 2\eta(s)|$ by one.

Besides, notice that

$$\begin{aligned} H(\boldsymbol{\lambda}, \boldsymbol{\gamma}) &= \frac{1}{2} \sum_{s=1}^K \frac{1}{\alpha_s} \mathbb{E}_{\boldsymbol{X}|S=s} \left(\left| \frac{\alpha_s p_s}{\bar{\alpha}} (1 - 2\eta(\boldsymbol{X}, s)) + \gamma_s \right| - \frac{p_s \alpha_s}{\bar{\alpha}} - 2\alpha_s \lambda_s - \gamma_s \right)_+ + \sum_{s=1}^K \lambda_s \alpha_s \\ &\geq \sum_{s=1}^K \frac{1}{\alpha_s} \mathbb{E}_{\boldsymbol{X}|S=s} \left(-\frac{\alpha_s p_s}{\bar{\alpha}} \eta(\boldsymbol{X}, s) - \alpha_s \lambda_s \right)_+ + \sum_{s=1}^K \lambda_s \alpha_s \\ &\geq \sum_{s=1}^K \left\{ \left(-\frac{p_s}{\bar{\alpha}} \eta(s) - \lambda_s \right)_+ + \lambda_s \alpha_s \right\} \end{aligned}$$

One easily observes that

$$\sum_{s=1}^{K}\left\{\left(-\frac{p_s}{\bar{\alpha}}\eta(s)-\lambda_s\right)_+ + \lambda_s\alpha_s\right\} \geq \sum_{s=1}^{K}\{\alpha_s \wedge (1-\alpha_s)\}|\lambda_s| - \sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\eta(s)\{(2\alpha_s)\vee 1\} \ . \quad (9.19)$$

Observe that for any $(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in \mathbb{R}^K \times \mathbb{R}^K$ and for any $c \in \mathbb{R}$ the transformation

$$\gamma_s \mapsto \gamma_s + \frac{p_s\alpha_s}{\bar{\alpha}}c, \quad \text{and} \quad \lambda_s \mapsto \lambda_s \quad s \in [K] \ ,$$

does not change the value of the objective function. Take any minimizing sequence $(\boldsymbol{\lambda}^k,\boldsymbol{\gamma}^k)$ of H. Due to the above observation we transform $(\boldsymbol{\lambda}^k,\boldsymbol{\gamma}^k)$ to another minimizing sequence with the property

$$\sum_{s=1}^{K}\gamma_s^k = 0, \qquad \forall k \in \mathbb{N} \ . \quad (9.20)$$

By an abuse of notation we denote this transformed sequence by $(\boldsymbol{\lambda}^k,\boldsymbol{\gamma}^k)$. By definition of $(\boldsymbol{\lambda}^k,\boldsymbol{\gamma}^k)$, for any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$H(\boldsymbol{\lambda}^k,\boldsymbol{\gamma}^k) \leq H(\mathbf{0},\mathbf{0}) + \epsilon, \qquad \forall k \geq N \ .$$

Since,

$$H(\mathbf{0},\mathbf{0}) = \sum_{s=1}^{K}\frac{p_s}{2\bar{\alpha}}\left(|1-2\eta(\boldsymbol{X},s)|-1\right)_+ = 0 \ ,$$

it holds for all $k \geq N$ that

$$H(\boldsymbol{\lambda}^k,\boldsymbol{\gamma}^k) \leq \epsilon, \qquad \forall k \geq N \ .$$

Furthermore, since for all $k \in \mathbb{N}$ the property in Eq. (9.20) holds, then using Eqs. (9.18) and (9.19) we obtain

$$\begin{aligned}
&\|\boldsymbol{\gamma}^k\|_1 \leq 2(1+\epsilon) \ , \\
&\sum_{s=1}^{K}\{\alpha_s \wedge (1-\alpha_s)\}|\lambda_s^k| \leq \epsilon + \sum_{s=1}^{K}\frac{p_s}{\bar{\alpha}}\eta(s)\{(2\alpha_s)\vee 1\} \ .
\end{aligned}$$

Thus for all $k \geq N$ the minimizing sequence $(\boldsymbol{\lambda}^k,\boldsymbol{\gamma}^k)$ is bounded, extracting convergent sub-sequence and using the fact that H is continuous we conclude that the global minimizer exists.

Refined bound on $\boldsymbol{\lambda}$. Recall that the first-order optimality condition on $\boldsymbol{\lambda}^*$ (see (**FOOC**)) is given by: for all $sin[K]$

$$\alpha_s = \mathbb{P}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{X},s)-\langle\boldsymbol{\gamma}^*,\mathbf{1}\rangle)+\frac{\boldsymbol{\gamma}_s^*}{2\alpha_s}\right| \geq \frac{p_s}{2\bar{\alpha}}(1-\langle\boldsymbol{\gamma}^*,\mathbf{1}\rangle)+\boldsymbol{\lambda}_s^*+\frac{\boldsymbol{\gamma}_s^*}{2\alpha_s}\right) \ .$$

Since $\eta(x,s) \in [0,1]$, then for any $\boldsymbol{x} \in \mathbb{R}^d$ it holds that

$$-\frac{p_s}{\bar{\alpha}} - \frac{(\boldsymbol{\gamma}_s^*)_-}{\alpha_s} \leq \left|\frac{p_s}{2\bar{\alpha}}(1-2\eta(\boldsymbol{x},s)) + \frac{\boldsymbol{\gamma}_s^*}{2\alpha_s}\right| - \frac{p_s}{2\bar{\alpha}} - \frac{\boldsymbol{\gamma}_s^*}{2\alpha_s} \leq -\frac{(\boldsymbol{\gamma}_s^*)_-}{\alpha_s} \ .$$

Therefore, if α_s is not in $\{0,1\}$, we must have that

$$-\frac{p_s}{\bar{\alpha}} \leq \boldsymbol{\lambda}_s^* + \frac{(\boldsymbol{\gamma}_s^*)_-}{\alpha_s} \leq 0 \ ,$$

otherwise the considered probability is either equal to 0 or to 1. In particular, it implies that

$$|\boldsymbol{\lambda}_s^*| \leq \frac{p_s}{\bar{\alpha}} \vee \frac{|\boldsymbol{\gamma}_s^*|}{\alpha_s} \ .$$

Note that the same can be shown for $\widehat{\boldsymbol{\lambda}}$ since Eq. (9.10) and Lemma 9.9.3 imply

$$\left|\widehat{\mathbb{P}}_{\boldsymbol{X}|S=s}\left(\left|\frac{p_s}{2\bar{\alpha}}(1-2\widehat{\eta}(\boldsymbol{X},s)-\widehat{\boldsymbol{\gamma}}_s) + \frac{\widehat{\boldsymbol{\gamma}}_s}{2\alpha_s}\right| \geq \frac{p_s}{2\bar{\alpha}}(1-\widehat{\boldsymbol{\gamma}}_s) + \widehat{\boldsymbol{\lambda}}_s + \frac{\boldsymbol{\gamma}_s}{2\alpha_s}\right) - \alpha_s\right| \leq \frac{2}{n_s}, \forall s \in [K] \ ,$$

and the assumption on n_s guarantee that the empirical probability is strictly between 0 and 1.

□

9.9.6 Reduction to linear programming

In this section we show that the minimization problem in Eq. (9.3) can be reduced to a problem of linear programming. Recall that our goal is to solve

$$\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})} \left\{ \langle \boldsymbol{\lambda}, \boldsymbol{\alpha} \rangle + \sum_{s=1}^{K} \widehat{\mathbb{E}}_{\boldsymbol{X}|S=s}(\widehat{G}(\boldsymbol{X},s,\boldsymbol{\lambda},\boldsymbol{\gamma}))_+ \right\} \ , \tag{9.21}$$

where

$$\widehat{G}(\boldsymbol{x},s,\boldsymbol{\lambda},\boldsymbol{\gamma}) = \left|\frac{p_s}{2\bar{\alpha}}(1-2\widehat{\eta}(\boldsymbol{x},s)-\langle\boldsymbol{\gamma},\mathbf{1}\rangle) + \frac{\boldsymbol{\gamma}_s}{2\alpha_s}\right| - \frac{p_s}{2\bar{\alpha}}(1-\langle\boldsymbol{\gamma},\mathbf{1}\rangle) - \boldsymbol{\lambda}_s - \frac{\boldsymbol{\gamma}_s}{2\alpha_s} \ .$$

Similarly to the support vector machines, the reduction is achieved via the slack variables ζ_i, $i = 1,\dots,n$. With these slack variables the above problem can be expressed as

$$\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\zeta})} \langle \boldsymbol{\lambda}, \boldsymbol{\alpha} \rangle + \sum_{s=1}^{K} \sum_{i \in \mathcal{I}_s} \frac{\zeta_i}{n_s}$$

$$\text{s.t.} \begin{cases} \zeta_i \geq 0 & \forall i \in [n] \\ 0 \leq \zeta_i + \boldsymbol{\lambda}_s + \frac{p_s}{\bar{\alpha}}\widehat{\eta}(\boldsymbol{x}_i,s) & \forall i \in \mathcal{I}_s \forall s \in [K] \\ 0 \leq \zeta_i + \left\langle \boldsymbol{\gamma}, \frac{1}{\alpha_s}\boldsymbol{e}_s - \frac{p_s}{\bar{\alpha}}\mathbf{1}\right\rangle + \boldsymbol{\lambda}_s + \frac{p_s}{\bar{\alpha}}(1-\widehat{\eta}(\boldsymbol{x}_i,s)) & \forall i \in \mathcal{I}_s \forall s \in [K] \end{cases} \tag{LP-Primal}$$

To prove this result it is sufficient to observe that for all $x \in \mathbb{R}$ it holds that

$$(x)_+ = \min_{\zeta \geq x, \zeta \geq 0} \zeta \ .$$

Introduce the following notation

$$\boldsymbol{c} = \Bigg(\underbrace{1/n_1, \dots, 1/n_1}_{\mathcal{I}_1}, \dots \underbrace{1/n_s, \dots, 1/n_s}_{\mathcal{I}_s}, \dots, \underbrace{1/n_K, \dots, 1/n_K}_{\mathcal{I}_K}, \alpha_1, \dots, \alpha_K, 0 \dots, 0 \Bigg)$$

$$\boldsymbol{y} = (\boldsymbol{\zeta}^\top, \boldsymbol{\lambda}^\top, \boldsymbol{\gamma}^\top)$$

$$\boldsymbol{b} = \frac{1}{\bar{\alpha}} \Bigg((p_1 \widehat{\eta}(\boldsymbol{x}_i, s))_{i \in \mathcal{I}_1}, \dots, (p_K \widehat{\eta}(\boldsymbol{x}_i, s))_{i \in \mathcal{I}_K}, \\ (p_1 (1 - \widehat{\eta}(\boldsymbol{x}_i, s)))_{i \in \mathcal{I}_1}, \dots, (p_K (1 - \widehat{\eta}(\boldsymbol{x}_i, s)))_{i \in \mathcal{I}_K} \Bigg)$$

$$\mathbf{A} = \left[\begin{array}{cccc|c|c} -\mathbf{I}_{n_1 \times n_1} & \mathbf{0}_{n_1 \times n_2} & \dots & \mathbf{0}_{n_1 \times n_K} & -\mathbf{E}^1_{n_1 \times K} & \mathbf{0}_{n_2 \times K} \\ \mathbf{0}_{n_2 \times n_1} & -\mathbf{I}_{n_2 \times n_2} & \dots & \mathbf{0}_{n_2 \times n_K} & -\mathbf{E}^2_{n_2 \times K} & \mathbf{0}_{n_1 \times K} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0}_{n_K \times n_1} & \mathbf{0}_{n_K \times n_2} & \dots & -\mathbf{I}_{n_K \times n_K} & -\mathbf{E}^K_{n_K \times K} & \mathbf{0}_{n_K \times K} \\ \hline -\mathbf{I}_{n_1 \times n_1} & \mathbf{0}_{n_1 \times n_2} & \dots & \mathbf{0}_{n_1 \times n_K} & -\mathbf{E}^1_{n_1 \times K} & \frac{p_1}{\bar{\alpha}} \mathbf{1}_{n_1 \times K} - \frac{1}{\alpha_1} \mathbf{E}^1_{n_1 \times K} \\ \mathbf{0}_{n_2 \times n_1} & -\mathbf{I}_{n_2 \times n_2} & \dots & \mathbf{0}_{n_2 \times n_K} & -\mathbf{E}^2_{n_2 \times K} & \frac{p_2}{\bar{\alpha}} \mathbf{1}_{n_2 \times K} - \frac{1}{\alpha_2} \mathbf{E}^2_{n_2 \times K} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0}_{n_K \times n_1} & \mathbf{0}_{n_K \times n_2} & \dots & -\mathbf{I}_{n_K \times n_K} & -\mathbf{E}^K_{n_K \times K} & \frac{p_K}{\bar{\alpha}} \mathbf{1}_{n_K \times K} - \frac{1}{\alpha_K} \mathbf{E}^K_{n_K \times K} \end{array}\right]$$

where $\mathbf{E}^s_{n \times m}$ is a $n \times m$ matrix composed of zeros and ones, whose s^{th} column is equal to $\mathbf{1}$ and all other elements are zero, $\mathbf{1}_{n \times m}$ is a matrix of ones of size $n \times m$. Using the above notation, the problem in (**LP-Primal**) can be written as

$$\min_{\boldsymbol{y} \in \mathbb{R}^{n+2K}} \langle \boldsymbol{c}, \boldsymbol{y} \rangle \\ \text{s.t.} \begin{cases} \mathbf{A}\boldsymbol{y} \leq \boldsymbol{b} \\ y_i \geq 0 \qquad i \in [n] \end{cases} \tag{**LP-Primal-compacted**}$$

While the dimension of matrix $\mathbf{A}$ is $2n \times (n + 2K)$, this matrix has at most $4n + nK$ non-zero elements. This fact can be exploited if $n \gg K$, that it, the amount of *unlabeled* data is large compared to the amount of groups.

Bibliography

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach (2018). "A Reductions Approach to Fair Classification". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 60–69. URL: http://proceedings.mlr.press/v80/agarwal18a.html.

Agarwal, Alekh, Miroslav Dudík, and Zhiwei Steven Wu (2019). "Fair regression: Quantitative definitions and reduction-based algorithms". In: *arXiv preprint arXiv:1905.12843*.

Agueh, Martial and Guillaume Carlier (2011). "Barycenters in the Wasserstein space". In: *SIAM Journal on Mathematical Analysis* 43.2, pp. 904–924.

Ali, Syed Mumtaz and Samuel D Silvey (1966). "A general class of coefficients of divergence of one distribution from another". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 28.1, pp. 131–142.

Altschuler, Jason, Victor-Emmanuel Brunel, and Alan Malek (Feb. 2018). "Best Arm Identification for Contaminated Bandits". In: *arXiv e-prints*, arXiv:1802.09514, arXiv:1802.09514. arXiv: 1802.09514.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (May 23, 2016). "Machine Bias". In: *ProPublica*. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 05/23/2016).

Antoniou, Antreas, Amos Storkey, and Harrison Edwards (2017). "Data augmentation generative adversarial networks". In: *arXiv preprint arXiv:1711.04340*.

Arcones, Miguel A (1994). "Some strong limit theorems for M-estimators". In: *Stochastic Processes and Their Applications* 53.2, pp. 241–268.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein gan". In: *arXiv preprint arXiv:1701.07875*.

Audibert, Jean-Yves and Sébastien Bubeck (2010). "Best arm identification in multi-armed bandits". In.

Audibert, Jean-Yves and Olivier Catoni (2011). "Robust linear least squares regression". In: *The Annals of Statistics* 39.5, pp. 2766–2794.

Audibert, Jean-Yves and Alexandre B Tsybakov (2007). "Fast learning rates for plug-in classifiers". In: *The Annals of statistics* 35.2, pp. 608–633.

Baharlouei, Sina, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn (2019). "Rényi Fair Inference". In: *arXiv preprint arXiv:1906.12005*.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2017). "Fairness in machine learning". In: *NIPS Tutorial* 1.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2019). *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org.

Barocas, Solon and Andrew D Selbst (2016). "Big data's disparate impact". In: *Calif. L. Rev.* 104, p. 671.

Barrio, Eustasio del, Paul Deheuvels, and Sara van de Geer (2007). *Lectures on empirical processes*. EMS Series of Lectures in Mathematics. Theory and statistical applications, With a preface by Juan A. Cuesta Albertos and Carlos Matrán. European Mathematical Society (EMS), Zürich, pp. x+254.

Barrio, Eustasio del, Paula Gordaliza, and Jean-Michel Loubes (2020). "Review of Mathematical frameworks for Fairness in Machine Learning". In: *arXiv preprint arXiv:2005.13755*.

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

Bassetti, Federico, Antonella Bodini, and Eugenio Regazzini (2006). "On minimum Kantorovich distance estimators". In: *Statistics & probability letters* 76.12, pp. 1298–1302.

Bateni, Amir-Hossein and Arnak S. Dalalyan (2020). "Confidence regions and minimax rates in outlier-robust estimation on the probability simplex". In: *Electron. J. Statist.* 14.2, pp. 2653–2677. DOI: 10.1214/20-EJS1731. URL: https://doi.org/10.1214/20-EJS1731.

Bellec, Pierre C (2017). "Optimal exponential bounds for aggregation of density estimators". In: *Bernoulli* 23.1, pp. 219–248.

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

Bertsimas, Dimitris, Vivek F Farias, and Nikolaos Trichakis (2012). "On the efficiency-fairness trade-off". In: *Management Science* 58.12, pp. 2234–2250.

Biau, Gérard, Benoît Cadre, Maxime Sangnier, and Ugo Tanielian (2018). "Some theoretical properties of GANs". In: *arXiv preprint arXiv:1803.07819*.

Biau, Gérard, Maxime Sangnier, and Ugo Tanielian (2020). "Some Theoretical Insights into Wasserstein GANs". In: *arXiv preprint arXiv:2006.02682*.

Biau, Gérard and Erwan Scornet (2016). "A random forest guided tour". In: *Test* 25.2, pp. 197–227.

Bird, Sarah, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker (n.d.). *Fairlearn: A toolkit for assessing and improving fairness in AI*. Tech. rep.

Birman, Mikhail Shlemovich and Mikhail Zakharovich Solomyak (1967). "Piecewise-polynomial approximations of functions of the classes W_p^α". In: *Matematicheskii Sbornik* 115.3, pp. 331–355.

Bishop, Christopher M (2006). *Pattern recognition and machine learning*. springer.

Bobkov, S. and M. Ledoux (2019). *One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances*. Memoirs of the American Mathematical Society. American Mathematical Society. ISBN: 9781470436506.

Boucheron, Stéphane, Gábor Lugosi, and Pascal Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, p. 172.

Bousquet, Olivier and Nikita Zhivotovskiy (2019). "Fast classification rates without standard margin assumptions". In: *arXiv preprint arXiv:1910.12756*.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Briol, Francois-Xavier, Alessandro Barp, Andrew B Duncan, and Mark Girolami (2019). "Statistical Inference for Generative Models with Maximum Mean Discrepancy". In: *arXiv preprint arXiv:1906.05944*.

Brock, Andrew, Jeff Donahue, and Karen Simonyan (2018). "Large scale gan training for high fidelity natural image synthesis". In: *arXiv preprint arXiv:1809.11096*.

Bühlmann, Peter (2013). "Causal statistical inference in high dimensions". In: *Mathematical Methods of Operations Research* 77.3, pp. 357–370.

Bühlmann, Peter and Sara Van de Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Buolamwini, Joy and Timnit Gebru (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR, pp. 77–91.

Calders, T., F. Kamiran, and M. Pechenizkiy (2009). "Building classifiers with independency constraints". In: *IEEE international conference on Data mining*.

Calders, T., A. Karim, F. Kamiran, W. Ali, and X. Zhang (2013). "Controlling attribute effect in linear regression". In: *IEEE International Conference on Data Mining*.

Calmon, F., D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney (2017). "Optimized Pre-Processing for Discrimination Prevention". In: *Neural Information Processing Systems*.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization. Ecole d'été de probabilités de Saint-Flour XXXI-2001*. Collection : Lecture notes in mathematics n°1851. Springer, pp. viii–272. URL: https://hal.archives-ouvertes.fr/hal-00104952.

Cayton, Lawrence (2005). "Algorithms for manifold learning". In: *Univ. of California at San Diego Tech. Rep* 12.1-17, p. 1.

Chen, Mengjie, Chao Gao, and Zhao Ren (2016). "A general decision theory for Huber's ϵ-contamination model". In: *Electronic Journal of Statistics* 10.2, pp. 3752–3774.

Chen, Mengjie, Chao Gao, and Zhao Ren (2018). "Robust covariance and scatter matrix estimation under Huber's contamination model". In: *The Annals of Statistics* 46.5, pp. 1932–1960.

Chen, Minshuo, Wenjing Liao, Hongyuan Zha, and Tuo Zhao (2020). "Statistical Guarantees of Generative Adversarial Networks for Distribution Estimation". In: *arXiv preprint arXiv:2002.03938*.

Chernozhukov, Victor (2005). "Extremal quantile regression". In: *Ann. Statist.* 33.2, pp. 806–839.

Chiappa, Silvia, Ray Jiang, Tom Stepleton, Aldo Pacchiano, Heinrich Jiang, and John Aslanides (2020). "A General Approach to Fairness with Optimal Transport". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pp. 3633–3640. URL: https://aaai.org/ojs/index.php/AAAI/article/view/5771.

Chierichetti, F., R. Kumar, S. Lattanzi, and S. Vassilvitskii (2017). "Fair Clustering Through Fairlets". In: *Neural Information Processing Systems*.

Chouldechova, Alexandra (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big data* 5.2, pp. 153–163.

Chow, C. (1957). "An optimum character recognition system using decision functions". In: *IRE Transactions on Electronic Computers* 4, pp. 247–254.

Chow, C. (1970). "On optimum error and reject trade-off". In: *IEEE Trans. Inform. Theory* 16, pp. 41–46.

Chzhen, Evgenii, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil (2019). "Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification". In: *NeurIPS 2019-33th Annual Conference on Neural Information Processing Systems*.

Chzhen, Evgenii, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil (2020a). "Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees". In: *arXiv preprint arXiv*.

Chzhen, Evgenii, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil (2020b). "Fair regression via plug-in estimator and recalibration with statistical guarantees". In.

Chzhen, Evgenii, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil (2020c). "Fair regression with wasserstein barycenters". In: *Advances in Neural Information Processing Systems* 33.

Chzhen, Evgenii and Nicolas Schreuder (2020a). "A minimax framework for quantifying risk-fairness trade-off in regression". In: *arXiv preprint arXiv:2007.14265*.

Chzhen, Evgenii and Nicolas Schreuder (2020b). "An example of prediction which complies with Demographic Parity and equalizes group-wise risks in the context of regression". In: *arXiv preprint arXiv:2011.07158*.

Collins, John R. (1977). "Upper bounds on asymptotic variances of M-estimators of location". In: *Ann. Statist.* 5.4, pp. 646–657.

Costa, Jose A and Alfred O Hero (2004a). "Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets". In: *2004 12th European Signal Processing Conference*. IEEE, pp. 369–372.

Costa, Jose A. and Alfred O. Hero (2004b). "Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets". In: *2004 12th European Signal Processing Conference*, pp. 369–372.

Csiszár, Imre (1964). "Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten". In: *Magyer Tud. Akad. Mat. Kutato Int. Koezl.* 8, pp. 85–108.

Csiszár, Imre (1967). "Information-type measures of difference of probability distributions and indirect observation". In: *studia scientiarum Mathematicarum Hungarica* 2, pp. 229–318.

Cuturi, Marco (2013). "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems*, pp. 2292–2300.

Denis, Christophe and Mohamed Hebiri (2020). "Consistency of plug-in confidence sets for classification in semi-supervised learning". In: *Journal of Nonparametric Statistics* 32.1, pp. 42–72.

Devroye, Luc (1986). *Non-Uniform Random Variate Generation*. Springer. ISBN: 978-1-4613-8645-2. DOI: 10.1007/978-1-4613-8643-8. URL: https://doi.org/10.1007/978-1-4613-8643-8.

Devroye, Luc, László Györfi, and Gábor Lugosi (2013). *A probabilistic theory of pattern recognition*. Vol. 31. Springer Science & Business Media.

Dieterich, William, Christina Mendoza, and Tim Brennan (2016). "COMPAS risk scales: Demonstrating accuracy equity and predictive parity". In: *Northpoint Inc* 7.7.4, p. 1.

Diggle, Peter J. and Richard J. Gratton (1984). "Monte Carlo Methods of Inference for Implicit Statistical Models". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.2, pp. 193–212. DOI: https://doi.org/10.1111/j.2517-6161.1984.tb01290.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1984.tb01290.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1984.tb01290.x.

Donini, M., L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil (2018). "Empirical risk minimization under fairness constraints". In: *Neural Information Processing Systems*.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml.

Dudley, R. M. (1968). "The speed of mean Glivenko-Cantelli convergence". In: *Ann. Math. Statist.* 40, pp. 40–50.

Dudley, Richard Mansfield (1969). "The speed of mean Glivenko-Cantelli convergence". In: *The Annals of Mathematical Statistics* 40.1, pp. 40–50.

Dwork, C., N. Immorlica, A. T. Kalai, and M. D. M. Leiserson (2018). "Decoupled Classifiers for Group-Fair and Efficient Machine Learning". In: *Conference on Fairness, Accountability and Transparency*.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel (2012). "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226.

Dwork, Cynthia, Christina Ilvento, and Meena Jagadeesan (2020). "Individual fairness in pipelines". In: *arXiv preprint arXiv:2004.05167*.

Facco, Elena, Maria d'Errico, Alex Rodriguez, and Alessandro Laio (2018). "Estimating the intrinsic dimension of datasets by a minimal neighborhood information". In: *CoRR* abs/1803.06992. arXiv: 1803.06992. URL: http://arxiv.org/abs/1803.06992.

Farrell, Roger H (1964). "Asymptotic behavior of expected sample size in certain one sided tests". In: *The Annals of Mathematical Statistics*, pp. 36–72.

Fefferman, Charles, Sanjoy Mitter, and Hariharan Narayanan (2016). "Testing the manifold hypothesis". In: *Journal of the American Mathematical Society* 29.4, pp. 983–1049.

Fisher, Ronald A (1936). "The use of multiple measurements in taxonomic problems". In: *Annals of eugenics* 7.2, pp. 179–188.

Fitzsimons, J., A. Al Ali, M. Osborne, and S. Roberts (2018). "Equality Constrained Decision Trees: For the Algorithmic Enforcement of Group Fairness". In: *arXiv preprint arXiv:1810.05041*.

Fitzsimons, Jack, AbdulRahman Al Ali, Michael Osborne, and Stephen Roberts (2019). "A general framework for fair regression". In: *Entropy* 21.8, p. 741.

Fraenkel, LE (1978). "Formulae for high derivatives of composite functions". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 83. 2. Cambridge University Press, pp. 159–165.

Fréchet, Maurice (1957). "Sur la distance de deux lois de probabilité". In: *Comtes Rendus Hebdomadaires des Seances de l'Academie des Sciences* 244.6, pp. 689–692.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Gabillon, Victor, Mohammad Ghavamzadeh, and Alessandro Lazaric (2012). "Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence". In: *Advances in NeurIPS 25*, pp. 3221–3229.

Gajane, Pratik and Mykola Pechenizkiy (2017). "On formalizing fairness in prediction with machine learning". In: *arXiv preprint arXiv:1710.03184*.

Gangbo, Wilfrid and Andrzej Święch (1998). "Optimal maps for the multidimensional Monge-Kantorovich problem". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 51.1, pp. 23–45.

Garivier, Aurélien and Florencia Leonardi (2011). "Context tree selection: A unifying view". In: *Stochastic Processes and their Applications* 121.11, pp. 2488–2506.

Genevay, Aude, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré (2018). "Sample complexity of Sinkhorn divergences". In: *arXiv preprint arXiv:1810.02733*.

Genevay, Aude, Gabriel Peyré, and Marco Cuturi (2017). "Learning generative models with sinkhorn divergences". In: *arXiv preprint arXiv:1706.00292*.

Genevay, Aude, Gabriel Peyré, and Marco Cuturi (2018). "Learning generative models with Sinkhorn divergences". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1608–1617.

Genovese, Christopher R, Marco Perone Pacifico, Verdinelli Isabella, and Larry Wasserman (2012). "Minimax manifold estimation". In.

Gilbert, E. (1952). "A comparison of signalling alphabets". In: *The Bell system technical journal* 31.3, pp. 504–522.

Giné, Evarist and Richard Nickl (2016). *Mathematical foundations of infinite-dimensional statistical models*. Vol. 40. Cambridge University Press.

Goldt, Sebastian, Marc Mezard, Florent Krzakala, and Lenka Zdeborová (2020). "Modelling the influence of data structure on learning in neural networks: the hidden manifold model". In.

Goodfellow, Ian (2016). "Nips 2016 tutorial: Generative adversarial networks". In: *arXiv preprint arXiv:1701.00160*.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016). *Deep learning*. Vol. 1. 2. MIT press Cambridge.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). "Generative adversarial nets". In: *Advances in neural information processing systems*, pp. 2672–2680.

Gordaliza, P., E. Del Barrio, G. Fabrice, and J. M. Loubes (2019). "Obtaining fairness using optimal transport theory". In: *International Conference on Machine Learning*.

Grandvalet, Yves, Alain Rakotomamonjy, Joseph Keshet, and Stéphane Canu (2008). "Support vector machines with a reject option". In: *Advances in neural information processing systems* 21, pp. 537–544.

Györfi, L., Z. Györfi, and I. Vajda (Jan. 1979). "Bayesian decision with rejection". In: *Problems of Control and Information Theory* 8.

Györfi, László, Michael Kohler, Adam Krzyzak, and Harro Walk (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Haas, Christian (2019). "The Price of Fairness-A Framework to Explore Trade-Offs in Algorithmic Fairness". In: *arXiv preprint arXiv*.

Haberman, Shelby J (1989). "Concavity and estimation". In: *The Annals of Statistics*, pp. 1631–1661.

Hardt, M., E. Price, and N. Srebro (2016). "Equality of opportunity in supervised learning". In: *Neural Information Processing Systems*.

He, Xuming and Gang Wang (1995). "Law of the iterated logarithm and invariance principle for M-estimators". In: *Proceedings of the American Mathematical Society* 123.2, pp. 563–573.

Herbei, R. and M. Wegkamp (2006). "Classification with reject option". In: *Canad. J. Statist.* 34.4, pp. 709–721.

Hoeffding, Wassily (1994). "Probability inequalities for sums of bounded random variables". In: *The Collected Works of Wassily Hoeffding*. Springer, pp. 409–426.

Hoerl, Arthur E. and Robert W. Kennard (2000). "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 42.1, pp. 80–86.

Howard, Steven R, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon (2018). "Uniform, nonparametric, non-asymptotic confidence sequences". In: *arXiv preprint arXiv:1810.08240*.

Hsu, Daniel, Sham M Kakade, and Tong Zhang (2012). "Random design analysis of ridge regression". In: *Conference on learning theory*, pp. 9–1.

Hsu, Daniel and Sivan Sabato (2016). "Loss Minimization and Parameter Estimation with Heavy Tails". In: *Journal of Machine Learning Research* 17.18, pp. 1–40.

Huber, Peter J (1964). "Robust estimation of a location parameter". In: *The annals of mathematical statistics* 35.1, pp. 73–101.

Huber, Peter J. and Elvezio M. Ronchetti (2009). *Robust statistics*. Second. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ.

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros (2017). "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134.

Jacod, Jean and Philip Protter (2012). *Probability essentials*. Springer Science & Business Media.

Jamieson, Kevin, Matthew Malloy, Robert Nowak, and Sébastien Bubeck (2014). "lil'UCB: An optimal exploration algorithm for multi-armed bandits". In: *Conference on Learning Theory*, pp. 423–439.

Jamieson, Kevin G. and Robert D. Nowak (2014). "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting". In: *48th Annual Conference on Information Sciences and Systems, CISS*, pp. 1–6.

Jiang, R., A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa (2019). "Wasserstein fair classification". In: *arXiv preprint arXiv:1907.12059*.

Jones, Erik, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang (2020). *Selective Classification Can Magnify Disparities Across Groups*. arXiv: 2010.14134 [cs.LG].

Jung, Christopher, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu (2019). "Eliciting and enforcing subjective individual fairness". In: *arXiv preprint arXiv:1905.10660*.

Karmarkar, Narendra (1984). "A new polynomial-time algorithm for linear programming". In: *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pp. 302–311.

Karras, Tero, Samuli Laine, and Timo Aila (2019). "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4401–4410.

Kaufmann, Emilie, Olivier Cappé, and Aurélien Garivier (2016). "On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models". In: *Journal of Machine Learning Research* 17, 1:1–1:42.

Kaufmann, Emilie and Wouter M. Koolen (2018). "Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals". In: *CoRR* abs/1811.11419.

Kerkyacharian, Gerard, Alexandre B Tsybakov, Vladimir Temlyakov, Dominique Picard, and Vladimir Koltchinskii (2014). "Optimal exponential bounds on the accuracy of classification". In: *Constructive Approximation* 39.3, pp. 421–444.

Khachiyan, Leonid Genrikhovich (1979). "A polynomial algorithm in linear programming". In: *Doklady Akademii Nauk*. Vol. 244. 5. Russian Academy of Sciences, pp. 1093–1096.

Khintchine, Aleksandr (1924). "Über einen Satz der Wahrscheinlichkeitsrechnung". ger. In: *Fundamenta Mathematicae* 6.1, pp. 9–20.

Kilbertus, N., M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf (2017). "Avoiding Discrimination through Causal Reasoning". In: *Neural Information Processing Systems*.

Kloeckner, Benoit (2010). "A geometric study of Wasserstein spaces: Euclidean spaces". In: *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze* 9.2, pp. 297–323.

Klopp, Olga, Alexandre B. Tsybakov, and Nicolas Verzelen (2017). "Oracle inequalities for network models and sparse graphon estimation". In: *The Annals of Statistics* 45.1, pp. 316–354. DOI: 10.1214/16-AOS1454.

Klopp, Olga and Nicolas Verzelen (2019). "Optimal graphon estimation in cut distance". In: *Probability Theory and Related Fields* 174.3, pp. 1033–1090.

Kodali, Naveen, Jacob Abernethy, James Hays, and Zsolt Kira (2017). "On convergence and stability of GANs". In: *arXiv preprint arXiv:1705.07215*.

Köeppen, M., K. Yoshida, and K. Ohnishi (2014). "Evolving Fair Linear Regression for the Representation of Human-Drawn Regression Lines". In: *2014 International Conference on Intelligent Networking and Collaborative Systems*, pp. 296–303.

Kohavi, Ron (1996). "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." In: *Kdd*. Vol. 96, pp. 202–207.

Kolmogoroff, A. (1929). "Über das Gesetz des iterierten Logarithmus". In: *Mathematische Annalen* 101, pp. 126–135.

Koltchinskii, Vladimir (2011a). *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Vol. 2033. Lecture Notes in Mathematics. Heidelberg: Springer, pp. x+254.

Koltchinskii, Vladimir (2011b). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*. Vol. 2033. Springer Science & Business Media.

Komiyama, J. and H. Shimao (2017). "Two-stage Algorithm for Fairness-aware Machine Learning". In: *arXiv preprint arXiv:1710.04924*.

Komiyama, J., A. Takeda, J. Honda, and H. Shimao (2018). "Nonconvex Optimization for Regression with Fairness Constraints". In: *International Conference on Machine Learning*.

Korostelev, Aleksandr Petrovich and Alexandre B Tsybakov (2012). *Minimax theory of image reconstruction*. Vol. 82. Springer Science & Business Media.

Kourou, Konstantina, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis (2015). "Machine learning applications in cancer prognosis and prediction". In: *Computational and structural biotechnology journal* 13, pp. 8–17.

Kusner, M. J., J. Loftus, C. Russell, and R. Silva (2017). "Counterfactual fairness". In: *Neural Information Processing Systems*.

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*. Cambridge University Press. DOI: 10.1017/9781108571401.

Laurent, Beatrice and Pascal Massart (2000). "Adaptive estimation of a quadratic functional by model selection". In: *Annals of Statistics*, pp. 1302–1338.

Le Gouic, Thibaut and Jean-Michel Loubes (2017). "Existence and consistency of Wasserstein barycenters". In: *Probability Theory and Related Fields* 168.3-4, pp. 901–917.

Le Gouic, Thibaut, Jean-Michel Loubes, and Philippe Rigollet (2020). "Projection to fairness in statistical learning". In: *arXiv preprint arXiv:2005.11720*.

Lecué, Guillaume and Philippe Rigollet (Feb. 2014). "Optimal learning with Q-aggregation". In: *Ann. Statist.* 42.1, pp. 211–224.

LeCun, Yann (1998). "The MNIST database of handwritten digits". In: *http://yann. lecun. com/exdb/mnist/*.

LeCun, Yann, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel (1990). "Handwritten digit recognition with a back-propagation network". In: *Advances in neural information processing systems*, pp. 396–404.

Ledig, Christian, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, and Zehan Wang (2017). "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690.

Lee, Yin Tat and Aaron Sidford (2015). "Efficient inverse maintenance and faster algorithms for linear programming". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE, pp. 230–249.

Lei, J. (2014a). "Classification with confidence". In: *Biometrika* 101.4, pp. 755–769.

Lei, Jing (2014b). "Classification with confidence". In: *Biometrika* 101.4, pp. 755–769.

Lei, Jing (2020). "Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces". In: *Bernoulli* 26.1, pp. 767–798.

Levina, Elizaveta and Peter J Bickel (2005). "Maximum likelihood estimation of intrinsic dimension". In: *Advances in neural information processing systems*, pp. 777–784.

Liang, Tengyuan (2018). "On how well generative adversarial networks learn densities: Nonparametric and parametric results". In: *arXiv preprint arXiv:1811.03179*.

Liang, Tengyuan (2019). "Estimating Certain Integral Probability Metric (IPM) Is as Hard as Estimating under the IPM". In: *arXiv preprint arXiv:1911.00730*.

Liang, Tengyuan and James Stokes (2018). "Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks". In: *arXiv preprint arXiv:1802.06132*.

Lipton, Zachary, Julian McAuley, and Alexandra Chouldechova (2018). "Does mitigating ML's impact disparity require treatment disparity?" In: *Advances in Neural Information Processing Systems*, pp. 8125–8135.

Liu, Shuang, Olivier Bousquet, and Kamalika Chaudhuri (2017). "Approximation and convergence properties of generative adversarial learning". In: *arXiv preprint arXiv:1705.08991*.

Loftus, Joshua R, Chris Russell, Matt J Kusner, and Ricardo Silva (2018). "Causal reasoning for algorithmic fairness". In: *arXiv preprint arXiv:1805.05859*.

Loh, Po-Ling (2015). "Statistical consistency and asymptotic normality for high-dimensional robust M-estimators". In: *CoRR* abs/1501.00312. arXiv: 1501.00312. URL: http://arxiv.org/abs/1501.00312.

Luise, Giulia, Massimiliano Pontil, and Carlo Ciliberto (2020). "Generalization Properties of Optimal Transport GANs with Latent Distribution Learning". In: *arXiv preprint arXiv:2007.14641*.

Lum, K. and J. Johndrow (2016). "A statistical framework for fair predictive algorithms". In: *arXiv preprint arXiv:1610.08077*.

Madras, David, Elliot Creager, Toniann Pitassi, and Richard Zemel (2018). "Learning Adversarially Fair and Transferable Representations". In: *International Conference on Machine Learning*, pp. 3384–3393.

Madras, David, Toni Pitassi, and Richard Zemel (2018). "Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/09d37c08f7b129e96277388757530c72-Paper.pdf.

Maillard, Odalric-Ambrym (2019). "Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds". In: *Proceedings of ALT*. Vol. 98. PMLR, pp. 610–632.

Makhlouf, Karima, Sami Zhioua, and Catuscia Palamidessi (2020). "Survey on Causal-based Machine Learning Fairness Notions". In: *arXiv preprint arXiv:2010.09553*.

Maronna, Ricardo Antonio (1976). "Robust M-Estimators of Multivariate Location and Scatter". In: *The Annals of Statistics* 4.1, pp. 51–67.

Mary, Jérémie, Clément Calauzènes, and Noureddine El Karoui (2019). "Fairness-aware learning for continuous attributes and treatments". In: *International Conference on Machine Learning*, pp. 4382–4391.

Massart, Pascal (1990). "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". In: *The annals of Probability*, pp. 1269–1283.

Matousek, Jiri and Bernd Gärtner (2007). *Understanding and using linear programming*. Springer Science & Business Media.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan (2019). "A survey on bias and fairness in machine learning". In: *arXiv preprint arXiv:1908.09635*.

Menon, A. K. and R. C. Williamson (2018a). "The cost of fairness in binary classification". In: *Conference on Fairness, Accountability and Transparency*.

Menon, Aditya Krishna and Robert C Williamson (23–24 Feb 2018b). "The cost of fairness in binary classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, pp. 107–118. URL: http://proceedings.mlr.press/v81/menon18a.html.

Mohamed, Shakir and Balaji Lakshminarayanan (2016). "Learning in implicit generative models". In: *arXiv preprint arXiv:1610.03483*.

Mourtada, Jaouad (2019). "Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices". In: *arXiv preprint arXiv:1912.10754*.

Mourtada, Jaouad, Stéphane Gaïffas, and Erwan Scornet (2020). "Minimax optimal rates for Mondrian trees and forests". In: *Annals of Statistics* 48.4, pp. 2253–2276.

Mouzannar, Hussein, Mesrob I Ohannessian, and Nathan Srebro (2019). "From fair decision making to social equality". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 359–368.

Mukherjee, Debarghya, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun (13–18 Jul 2020). "Two Simple Ways to Learn Individual Fairness Metrics from Data". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 7097–7107. URL: http://proceedings.mlr.press/v119/mukherjee20a.html.

Müller, Alfred (1997). "Integral probability metrics and their generating classes of functions". In: *Advances in Applied Probability* 29.2, pp. 429–443.

Nadeem, Malik Sajjad Ahmed, Jean-Daniel Zucker, and Blaise Hanczar (2009). "Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option". In: *Machine Learning in Systems Biology*, pp. 65–81.

Nagarajan, Vaishnavh and J Zico Kolter (2017). "Gradient descent GAN optimization is locally stable". In: *Advances in neural information processing systems*, pp. 5585–5595.

Negahban, Sahand N., Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu (Nov. 2012). "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers". In: *Statist. Sci.* 27.4, pp. 538–557.

Nemirovski, A (2000). "TOPICS IN NON-PARAMETRIC STATISTICS". In: *Lecture Notes in Mathematics* 1738, pp. 86–282.

Neu, Gergely and Nikita Zhivotovskiy (2020). "Fast rates for online prediction with abstention". In: *Conference on Learning Theory*. PMLR, pp. 3030–3048.

Neveu, Jacques (1972). *Martingales à temps discret*. Vol. 1. 1. Masson Paris.

Nickl, Richard and Benedikt M Pötscher (2007). "Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov-and Sobolev-type". In: *Journal of Theoretical Probability* 20.2, pp. 177–199.

Nie, Weili and Ankit B Patel (2020). "Towards a better understanding and regularization of GAN training dynamics". In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 281–291.

Niemiro, Wojciech (1992). "Asymptotics for M-estimators defined by convex minimization". In: *The Annals of Statistics* 20.3, pp. 1514–1533.

Niles-Weed, Jonathan and Philippe Rigollet (2019). "Estimation of Wasserstein distances in the Spiked Transport Model". In: *arXiv preprint arXiv:1909.07513*.

Olfat, Matt, Stephen Sloan, Pedro Hespanhol, Matt Porter, Ram Vasudevan, and Anil Aswani (2020). "Covariance-Robust Dynamic Watermarking". In: *arXiv preprint arXiv:2003.13908*.

Oneto, L., M. Donini, and M. Pontil (2019a). "General Fair Empirical Risk Minimization". In: *arXiv preprint arXiv:1901.10080*.

Oneto, L., M. Donini, and M. Pontil (2019b). "General fair empirical risk minimization". In: *arXiv preprint arXiv:1901.10080*.

Oneto, Luca and Silvia Chiappa (2020). "Fairness in Machine Learning". In: *Recent Trends in Learning From Data*. Springer, pp. 155–196.

Oneto, Luca, Michele Donini, Andreas Maurer, and Massimiliano Pontil (2019). "Learning fair and transferable representations". In: *arXiv preprint arXiv:1906.10673*.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499*.

Osborne, Martin and Ariel Rubinstein (1994). *A Course in Game Theory*. Tech. rep. The MIT Press.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher,

M. Perrot, and E. Duchesnay (2011a). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011b). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini (2008). "Discrimination-aware data mining". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 560–568.

Pérez-Suay, Adrián, Valero Laparra, Gonzalo Mateo-Garcia, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls (2017). "Fair Kernel Learning". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part I*. Ed. by Michelangelo Ceci, Jaakko Hollmén, Ljupco Todorovski, Celine Vens, and Saso Dzeroski. Vol. 10534. Lecture Notes in Computer Science. Springer, pp. 339–355. DOI: 10.1007/978-3-319-71249-9_21. URL: https://doi.org/10.1007/978-3-319-71249-9%5C_21.

Pfau, David and Oriol Vinyals (2016). "Connecting generative adversarial networks and actor-critic methods". In: *arXiv preprint arXiv:1610.01945*.

Plečko, Drago and Nicolai Meinshausen (2019). "Fair Data Adaptation with Quantile Preservation". In: *arXiv preprint arXiv:1911.06685*.

Pless, Robert and Richard Souvenir (2009). "A survey of manifold learning for images". In: *IPSJ Transactions on Computer Vision and Applications* 1, pp. 83–94.

Poggio, Tomaso A., Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao (2017). "Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review". In: *Int. J. Autom. Comput.* 14.5, pp. 503–519. DOI: 10.1007/s11633-017-1054-2. URL: https://doi.org/10.1007/s11633-017-1054-2.

Portnoy, Stephen (1984). "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency". In: *Ann. Statist.* 12.4, pp. 1298–1309.

Primus, Richard A (2003). "Equal protection and disparate impact: Round three". In: *Harv. L. Rev.* 117, p. 494.

Puchkin, Nikita and Nikita Zhivotovskiy (2021). "Exponential Savings in Agnostic Active Learning through Abstention". In: *arXiv preprint arXiv:2102.00451*.

Quadrianto, Novi and Viktoriia Sharmanska (2017). "Recycling privileged learning and distribution matching for fairness". In: *Advances in Neural Information Processing Systems*, pp. 677–688.

Radford, Alec, Luke Metz, and Soumith Chintala (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434*.

Raff, E., J. Sylvester, and S. Mills (2018). "Fair forests: Regularized tree induction to minimize model bias". In: *AAAI/ACM Conference on AI, Ethics, and Society*.

Rakhlin, Alexander, Ohad Shamir, and Karthik Sridharan (2012). "Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization". In: *ICML 2012*. icml.cc / Omnipress.

Rakhlin, Alexander, Karthik Sridharan, and Alexandre B. Tsybakov (May 2017). "Empirical entropy, minimax regret and minimax risk". In: *Bernoulli* 23.2, pp. 789–824.

Richardson, Eitan and Yair Weiss (2018). "On GANs and GMMs". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grau-

man, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/0172d289da48c48de8c5ebf3de9f7ee1-Paper.pdf.

Rigollet, Phillippe and Jan-Christian Hütter (2015). "High dimensional statistics". In: *Lecture notes for course 18S997*.

Robert, Christian (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Robert, Christian P. and George Casella (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer. ISBN: 978-1-4419-1939-7. DOI: 10.1007/978-1-4757-4145-2. URL: https://doi.org/10.1007/978-1-4757-4145-2.

Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen (2016). "Improved techniques for training GANs". In: *Advances in neural information processing systems*, pp. 2234–2242.

Santambrogio, Filippo (2015). "Optimal transport for applied mathematicians". In: *Birkäuser, NY* 55.58-63, p. 94.

Sard, Arthur (1942). "The measure of the critical values of differentiable maps". In: *Bulletin of the American Mathematical Society* 48.12, pp. 883–890.

Scetbon, Meyer, Laurent Meunier, Jamal Atif, and Marco Cuturi (2020). "Equitable and Optimal Transport with Multiple Agents". In: *arXiv preprint arXiv:2006.07260*.

Schreuder, Nicolas (2020). "Bounding the expectation of the supremum of empirical processes indexed by Hölder classes". In: *Mathematical Methods of Statistics* 29, pp. 76–86.

Schreuder, Nicolas, Victor-Emmanuel Brunel, and Arnak S. Dalalyan (2020). "A nonasymptotic law of iterated logarithm for general M-estimators". In: *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 1331–1341.

Schreuder, Nicolas, Victor-Emmanuel Brunel, and Arnak S. Dalalyan (2021). "Statistical guarantees for generative models without domination". In: *Algorithmic Learning Theory*. Ed. by Vitaly Feldman, Katrina Ligett, and Sivan Sabato. Vol. 132. Proceedings of Machine Learning Research. PMLR, pp. 1051–1071.

Schreuder, Nicolas and Evgenii Chzhen (2021). "Classification with abstention but without disparities". In: *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*. Proceedings of Machine Learning Research. PMLR.

Seddik, Mohamed El Amine, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet (2020). "Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures". In: *arXiv preprint arXiv:2001.08370*.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press. ISBN: 978-1-10-705713-5. URL: http://www.cambridge.org/de/academic/subjects/computer-science/pattern-recognition-and-machine-learning/understanding-machine-learning-theory-algorithms.

Shalev-Shwartz, Shai, Nathan Srebro, and Tong Zhang (2010). "Trading Accuracy for Sparsity in Optimization Problems with Sparsity Constraints". In: *SIAM Journal on Optimization* 20.6, pp. 2807–2832.

Shin, Jaehyeok, Aaditya Ramdas, and Alessandro Rinaldo (2019). "On the bias, risk and consistency of sample means in multi-armed bandits". In: *CoRR* abs/1902.00746. URL: http://arxiv.org/abs/1902.00746.

Shiryayev, AN (1993). *Selected Works of AN Kolmogorov: Volume III: Information Theory and the Theory of Algorithms*. Vol. 27. Springer.

Singh, Shashank and Barnabás Póczos (2018). "Minimax distribution estimation in Wasserstein distance". In: *arXiv preprint arXiv:1802.08855*.

Singh, Shashank, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos (2018). "Nonparametric density estimation with adversarial losses". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran Associates Inc., pp. 10246–10257.

Srebro, Nathan and Karthik Sridharan (2010). "Note on refined Dudley integral covering number bound". In: *Unpublished results. http://ttic. uchicago. edu/karthik/dudley. pdf*.

Srebro, Nathan, Karthik Sridharan, and Ambuj Tewari (2010). "Smoothness, low noise and fast rates". In: *Advances in neural information processing systems*, pp. 2199–2207.

Sridharan, Karthik, Shai Shalev-shwartz, and Nathan Srebro (2009). "Fast Rates for Regularized Objectives". In: *Advances in NeurIPS 21*, pp. 1545–1552.

Sriperumbudur, Bharath K, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet (2012). "On the empirical estimation of integral probability metrics". In: *Electronic Journal of Statistics* 6, pp. 1550–1599.

Steinberg, Daniel, Alistair Reid, and Simon O'Callaghan (2020). "Fairness Measures for Regression via Probabilistic Classification". In: *arXiv preprint arXiv:2001.06089*.

Steinberg, Daniel, Alistair Reid, Simon O'Callaghan, Finnian Lattimore, Lachlan McCalman, and Tiberio Caetano (2020). "Fast Fair Regression via Efficient Approximations of Mutual Information". In: *arXiv preprint arXiv:2002.06200*.

Stone, Charles J (1977). "Consistent nonparametric regression". In: *The annals of statistics*, pp. 595–620.

Sugiyama, Masashi, Taiji Suzuki, and Takafumi Kanamori (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Sweetman, David (1990). *Van Gogh: His life and his art*. Crown Publishers New York.

Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. New York: Springer.

Tsybakov, Alexandre B (2003). "Optimal rates of aggregation". In: *Learning theory and kernel machines*. Springer, pp. 303–313.

Tsybakov, Alexandre B (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Uppal, A., S. Singh, and B. Poczos (2019). "Nonparametric Density Estimation: Convergence Rates for GANs under Besov IPM Losses". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pp. 9089–9100.

Vaart, A. W. van der (1998). *Asymptotic statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, pp. xvi+443.

Vaart, Aad W. van der and Jon A. Wellner (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. With applications to statistics. Springer-Verlag, New York, pp. xvi+508.

Van Der Walt, Stefan, S Chris Colbert, and Gael Varoquaux (2011). "The NumPy array: a structure for efficient numerical computation". In: *Computing in Science & Engineering* 13.2, p. 22.

Vapnik, V. and A. Chervonenkis (1968). "On the uniform convergence of relative frequencies of events to their probabilities". In: *Doklady Akademii Nauk SSSR* 181.4, pp. 781–787.

Varshamov, R. (1957). "Estimate of the number of signals in error correcting codes". In: *Dokl. Akad. Nauk SSSR* 117, pp. 739–741.

Vershynin, Roman (2010). "Introduction to the non-asymptotic analysis of random matrices". In: *arXiv preprint arXiv:1011.3027*.

Vershynin, Roman (2018). *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

Villani, Cédric (2008). *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stefan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, lhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antonio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods*.

Vovk, Vladimir, Alex Gammerman, and Glenn Shafer (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Wang, Jie, Peter Wonka, and Jieping Ye (2014). "Scaling SVM and Least Absolute Deviations via Exact Data Reduction". In: *ICML 2014*. Vol. 32. JMLR W.& C.P. Pp. 523–531.

Wasserman, Larry (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Weber, Max (1992). *Wissenschaft als Beruf: 1917-1919; Politik als Beruf: 1919*. Vol. 17. Mohr Siebeck.

Weed, Jonathan and Francis Bach (2019). "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance". In: *Bernoulli* 25.4A, pp. 2620–2648.

Wegkamp, M. and M. Yuan (2011). "Support vector machines with a reject option". In: *Bernoulli* 17.4, pp. 1368–1385.

Wick, Michael, Swetasudha Panda, and Jean-Baptiste Tristan (2019). "Unlocking Fairness: a Trade-off Revisited". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pp. 8783–8792. URL: http://papers.nips.cc/paper/9082-unlocking-fairness-a-trade-off-revisited.pdf.

Yuan, M. and M. Wegkamp (2010). "Classification methods with reject option based on convex risk minimization". In: *J. Mach. Learn. Res.* 11, pp. 111–130.

Zafar, M. B., I. Valera, M. Gomez Rodriguez, and K. P. Gummadi (2017). "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment". In: *International Conference on World Wide Web*.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, and C. Dwork (2013). "Learning fair representations". In: *International Conference on Machine Learning*.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

Zink, Anna and Sherri Rose (2019). "Fair regression for health care spending". In: *Biometrics* n/a.n/a.

Zliobaite, I. (2015). "On the relation between accuracy and fairness in binary classification". In: *arXiv preprint arXiv:1505.05723*.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Des compromis en apprentissage statistique: apprentissage en ligne, modèles génératifs et équité

Mots clés : Apprentissage statistique, statistique théorique, modèles génératifs, équité

Résumé : Les algorithmes d'apprentissage automatique sont reconnus pour leurs performances impressionnantes sur de nombreuses tâches que l'on croyait dédiées à l'esprit humain. Néanmoins, les algorithmes d'apprentissage automatique devenant omniprésents dans notre quotidien, il existe un besoin croissant de comprendre précisément leurs comportements et leurs limites. La théorie de l'apprentissage statistique est la branche de l'apprentissage automatique qui vise à fournir un formalisme de modélisation solide pour les problèmes d'inférence ainsi qu'une meilleure compréhension des propriétés statistiques des algorithmes d'apprentissage. La théorie de l'apprentissage statistique permet (i) de mieux comprendre les cas dans lesquels un algorithme fonctionne correctement (ii) de quantifier les compromis inhérents à l'apprentissage pour des choix algorithmiques pertinents (iii) de fournir des informations pour développer de nouveaux algorithmes. S'appuyant sur le cadre de l'apprentissage statistique, cette thèse présente des contributions liées à trois problèmes différents : l'apprentissage en ligne, la génération de données et, enfin, l'apprentissage équitable. Dans le cadre de l'apprentissage en ligne - où la taille de l'échantillon n'est pas connue à l'avance - nous fournissons des bornes de déviations uniformes en la taille de l'échantillon, dont la vitesse de convergence correspond à celle donnée par la loi du logarithme itéré pour une classe générale de M-estimateurs convexes – comprenant la moyenne, la médiane, les M-estimateurs de Huber. En ce qui concerne les modèles génératifs, nous proposons un cadre pratique pour étudier les modèles génératifs adversariaux (Goodfellow et al. 2014) d'un point de vue statistique afin d'évaluer l'impact d'une faible dimensionnalité intrinsèque des données sur l'erreur du modèle génératif. Nous établissons des limites de risque non asymptotiques pour le minimiseur du risque empirique. Enfin, notre travail sur l'apprentissage équitable consiste en une large étude de la contrainte de parité démographique, une contrainte populaire dans la littérature sur l'apprentissage équitable. Celle-ci contraint les prédicteurs à traiter des sous-groupes, définis par un attribut sensible comme le genre, pour qu'ils soient « traités de la même manière ». En particulier, nous proposons un cadre statistique minimax pour quantifier précisément le coût en risque d'introduire cette contrainte dans le cadre de la régression.

Title : A study of some trade-offs in statistical learning: online learning, generative models and fairness

Keywords : Statistical learning, theoretical statistics, online learning, generative models, fairness

Abstract : Machine learning algorithms are celebrated for their impressive performance on many tasks that we thought were dedicated to human minds, from handwritten digits recognition (LeCun et al. 1990) to cancer prognosis (Kourou et al. 2015). Nevertheless, as machine learning becomes more and more ubiquitous in our daily lives, there is a growing need for precisely understanding their behaviours and their limits. Statistical learning theory is the branch of machine learning which aims at providing a powerful modelling formalism for inference problems as well as a better understanding of the statistical properties of learning algorithms. Importantly, statistical learning theory allows one to (i) get a better understanding of the cases in which an algorithm performs well (ii) quantify trade-offs inherent to learning for better-informed algorithmic choices (iii) provide insights to develop new algorithms which will eventually outperform existing ones or tackle new tasks. Relying on the statistical learning framework, this thesis presents contributions related to three different learning problems: online learning, learning generative models and, finally, fair learning.

In the online learning setup – in which the sample size is not known in advance – we provide general anytime deviation bounds (or confidence intervals) whose width has the rate given in the Law of Iterated Logarithm for a general class of convex M-estimators – comprising the mean, the median, Huber's M-estimators.

Regarding generative models, we propose a convenient framework for studying adversarial generative models (Goodfellow et al. 2014) from a statistical perspective to assess the impact of (eventual) low intrinsic dimensionality of the data on the error of the generative model. We establish non-asymptotic risk bounds for the Empirical Risk Minimizer.

Finally, our work on fair learning consists in a broad study of the Demographic Parity constraint, a popular constraint in the fair learning literature. It essentially constrains predictors to treat groups defined by a sensitive attribute (e.g., gender or ethnicity) to be "treated the same". In particular, we propose a statistical minimax framework to precisely quantify the cost in risk of introducing this constraint in the regression setting.

Institut Polytechnique de Paris
91120 Palaiseau, France