Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto,
Massimiliano Pontil

Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification

Evgenii Chzhen[1,2], Christophe Denis[1], Mohamed Hebiri[1],
Luca Oneto[3], and Massimiliano Pontil[4,5]

[1]Université Paris-Est, [2]Université Paris-Sud, [3]Università di Pisa,
[4]Istituto Italiano di Tecnologia, [5]University College London

February 3, 2020

Abstract

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

1 Introduction

As machine learning becomes more and more spread in our society, the potential risk of using algorithms that behave unfairly is rising. As a result there is growing interest to design learning methods that meet "fairness" requirements, see [5, 9, 10, 17, 19, 22–24, 28, 31, 33, 47, 48, 50, 52] and references therein. A central goal is to make sure that sensitive information does not "unfairly" influence the outcomes of learning methods. For instance, if we wish to predict whether a university student applicant should be offered a scholarship based on curriculum, we would like our model to not unfairly use additional sensitive information such as gender or race.

Several measures of fairness of a classifier have been studied in the literature [49], ranging from Demographic Parity [8], Equal Odds and Equal Opportunity [22], Disparate Treatment, Impact, and Mistreatment [48], among others. In this paper, we study the problem of learning a binary classifier which satisfies the Equal Opportunity fairness constraint. It requires that the true positive rate of the classifier is the same across the sensitive groups. This notion has been used extensively in the literature either as a postprocessing step [22] on a learned classifier or directly during training, see for example [17] and references therein.

We address the important problem of devising statistically consistent and computationally efficient learning procedures that meet the fairness constraint. Specifically, we make four contributions. First, we derive in Proposition 2.3 the expression for the optimal equal opportunity classifier,

derived via thresholding of the Bayes regressor. Second, inspired by the above result we proposed a semi-supervised plug-in type method, which first estimates the regression function on labeled data and then estimates the unknown threshold using unlabeled data. Consequently, we establish in Theorem 4.5 that the proposed procedure is consistent, that is, it asymptotically satisfies the equal opportunity constraint and its risk converges to the risk of the optimal equal opportunity classifier. Finally, we present numerical experiments which indicate that our method is often superior or competitive with the state-of-the-art on benchmark datasets.

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

The paper is organized in the following manner. In Section 2, we introduce the problem and derive a form of the optimal equal opportunity classifier. Section 3 is devoted to the description of our method. In Section 4 we introduce assumptions used throughout this work and establish that the proposed learning algorithm is consistent. Finally, Section 5 presents numerical experiments with our method.

1.1 Related work

In this section we review previous contributions on the subject. Works on algorithmic fairness can be divided in three families. Our algorithm falls within the first family, which modifies a pre-trained classifier in order to increase its fairness properties while maintaining as much as possible the classification performance, see [6, 20, 22, 38] and references therein. Importantly, for our approach the post-processing step requires only unlabeled data, which is often easier to collect than its labeled counterpart. Methods in the second family enforce fairness directly during the training step, e.g. [2, 12, 17, 37]. The third family of methods implements fairness by modifying the data representation and then employs standard machine learning methods, see e.g. [1, 9, 17, 25–27, 50] as representative examples.

To the best of our knowledge the formula for the optimal fair classifier presented here is novel. In [22] the authors note that the optimal equalized odds or equal opportunity classifier can be derived from the Bayes optimal regressor, however, no explicit expression for this threshold is provided. The idea of recalibrating the Bayes classifier is also discussed in a number of papers, see for example [35, 38] and references therein. More importantly, the problem of deriving *efficient* and *consistent* estimators under fairness constraints has received limited attention in the literature. In [17], the authors present consistency results under restrictive assumptions on the model class. Furthermore, they only consider convex approximations of the risk and fairness constraint and it is not clear how to relate their results to the original problem with the miss-classification risk. In [2], the authors reduce the problem of fair classification to a sequence of cost-sensitive problems by leveraging the saddle point formulation. They show that their algorithm is consistent in both risk and fairness constraints. However, similarly to [17], the authors of [2] assume that the family of possible classifiers admits a bounded Rademacher complexity.

Plug-in methods in classification problems are well established and are well studied from statistical perspective, see [4, 16, 46] and references therein; in particular, it is known that one can build a plug-in type classifier which is optimal in minimax sense [4, 46]. Until very recently, theoretical studies on such methods were reduced to an efficient estimation of the regression function. Indeed, in

standard settings of classification the threshold is always known beforehand, thus, all the information about the optimal classifier is wrapped into the distribution of the label conditionally on the feature.

More recently, classification problems with a distribution dependent threshold have emerged. Prominent examples include classification with non-decomposable measures [30, 45, 51], classification with reject option [15, 32], and confidence set setup of multi-class classification [11, 14, 40], among others. A typical estimation algorithm in these scenarios is based on the plug-in strategy, which uses extra data to estimate the unknown threshold. Interestingly, in some setups a practitioner does not need to have access to two labeled samples and optimal estimation can be efficiently performed in semi-supervised manner [11, 14].

2 Optimal Equal Opportunity classifier

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

$$\mathcal{R}(g) := \mathbb{P}\left(g(X, S) \neq Y\right) \quad . \tag{1}$$

A *fair* optimal classifier is formally defined as

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

There are various definitions of fairness available in the literature, each having its critics and its supporter. In this work, we employ the following definition introduced in [22]. We refer the reader to this work as well as [2, 17, 35] for a discussion, motivation of this definition, and a comparison to other fairness definitions.

Definition 2.1 (Equal Opportunity [22])**.** *A classifier $(x, s) \mapsto g(x, s) \in \{0,1\}$ is called fair if*

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

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

Note, that the definition of fairness depends on the underlying distribution $\mathbb{P}$ and hence the whole class $\mathcal{F}(\mathbb{P})$ of the fair classifiers should be estimated. Further, notice that the class $\mathcal{F}(\mathbb{P})$ is non-empty as it always contains a classifier $g(x, s) \equiv 0$.

Using this notion of fairness we define an optimal equal opportunity classifier as a solution of the optimization problem

$$\min_{g \in \mathcal{G}} \left\{\mathcal{R}(g) \,:\, \mathbb{P}\left(g(X, S) = 1 \,|\, Y = 1, S = 1\right) = \mathbb{P}\left(g(X, S) = 1 \,|\, Y = 1, S = 0\right)\right\} \quad . \tag{2}$$

We now introduce an assumption on the regression function that plays an important role in establishing the form of the optimal fair classifier.

Assumption 2.2. *For each $s \in \{0,1\}$ we require the mapping $t \mapsto \mathbb{P}(\eta(X,S) \le t \,|\, S=s)$ to be continuous on $(0,1)$, where for all $(x,s) \in \mathbb{R}^d \times \{0,1\}$, we let the regression function*

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

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

The first part of Assumption 2.2 is achieved by many distributions and has been introduced in various contexts, see e.g. [11, 15, 32, 40, 45] and references therein. It says that, for every $s \in \{0,1\}$ the random variable $\eta(X,s)$ does not have atoms, that is, the event $\{\eta(X,s)=t\}$ has probability zero. The second part of the assumption states that the regression function $\eta(X,s)$ must surpass the level $1/2$ on a set of non-zero measure. Informally, returning to scholarship example mentioned in the introduction, this assumption means that there are individuals from *both* groups who are more likely to be offered a scholarship based on their curriculum.

In the following result we establish that the optimal equal opportunity classifier is obtained by recalibrating the Bayes classifier.

Proposition 2.3 (Optimal Rule)**.** *Under Assumption 2.2 an optimal classifier g^* can be obtained for all $(x,s) \in \mathbb{R}^d \times \{0,1\}$ as*

$$g^*(x,1) = \mathbf{1}_{\left\{1 \le \eta(x,1)\left(2 - \frac{\theta^*}{\mathbb{P}(Y=1,S=1)}\right)\right\}}, \quad g^*(x,0) = \mathbf{1}_{\left\{1 \le \eta(x,0)\left(2 + \frac{\theta^*}{\mathbb{P}(Y=1,S=0)}\right)\right\}} \tag{3}$$

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

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

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

Proof sketch. The proof relies on weak duality. The first step of the proof is to write the minimization problem for g^* using a "min-max" problem formulation. We consider the corresponding dual "max-min" problem and show that it can be analytically solved. Then, the continuity part of Assumption 2.2 allows to demonstrate that the solution of the "max-min" problem gives a solution of the "min-max" problem. The second part of Assumption 2.2 is used to prove that $|\theta^*| \le 2$. □

Before proceeding further, let us define a notion of unfairness, which plays a key role in our statistical analysis; it is sometimes referred to as difference of equal opportunity (DEO) in the literature [see e.g. 17].

Definition 2.4 (Unfairness)**.** *For any classifier g we define its unfairness as*

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

A principal goal of this paper is to construct a classification algorithm $\hat{g}$ which satisfies

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

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

3 Proposed procedure

In this section, we present the proposed plug-in algorithm and begin to study its theoretical properties.

We assume that we have at our disposal two datasets, labeled $\mathcal{D}_n$ and unlabeled $\mathcal{D}_N$ defined as

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

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

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

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

Remark 3.1. *In theory, the empirical distributions might be not well defined, since they are* only *valid if the unlabeled dataset $\mathcal{D}_N$ is composed of features from* both *groups. We show how to bypass this problem theoretically in supplementary material. Nevertheless, this remark has little to no impact in practice and in most situations these quantities are well defined.*

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

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

Using the above estimators a straightforward procedure to mimic the optimal classifier g^* provided by Proposition 2.3 is to employ a *plug-in rule* $\hat{g}$, obtained by replacing all the unknown quantities by either their empirical versions or their estimates. Specifically, we let $\hat{g}$ at $(x, s) \in \mathbb{R}^d \times \{0, 1\}$ as

$$\hat{g}(x,1) = \mathbf{1}_{\left\{1 \le \hat{\eta}(x,1)\left(2 - \frac{\hat{\theta}}{\hat{\mathbb{P}}(Y=1,S=1)}\right)\right\}}, \quad \hat{g}(x,0) = \mathbf{1}_{\left\{1 \le \hat{\eta}(x,0)\left(2 + \frac{\hat{\theta}}{\hat{\mathbb{P}}(Y=1,S=0)}\right)\right\}} \ . \tag{4}$$

It remains to define the value of $\hat{\theta}$, clearly it is desirable to mimic the condition that is satisfied by θ^* in Proposition 2.3. To this end, we make use of the unlabeled data $\mathcal{D}_N$ and of the estimator $\hat{\eta}$ previously built from the labeled dataset $\mathcal{D}_n$. Consequently, we define a data-driven version of unfairness $\Delta(g, \mathbb{P})$, which allows to construct an approximation $\hat{\theta}$ of the true value θ^*.

Definition 3.2 (Empirical unfairness)**.** *For any classifier g, an estimator $\hat{\eta}$ based on $\mathcal{D}_n$, and unlabeled sample $\mathcal{D}_N$ the empirical unfairness is defined as*

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

Notice that the empirical unfairness $\hat{\Delta}(g, \mathbb{P})$ is data-driven, that is, it does not involve unknown quantities. One might wonder why it is an empirical version of the quantity $\Delta(g, \mathbb{P})$ in Definition 2.4

and what is the reason to introduce it. The definition reveals itself when we rewrite the population of unfairness $\Delta(g, \mathbb{P})$ using[1] the identity

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

Using the above expression we can rewrite

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

Hence, the passage from the population unfairness to its empirical version in Definition 3.2 formally reduces to substituting "hats" to all the unknown quantities.

Using Definition 3.2, a logical estimator $\hat{\theta}$ of θ^* can be obtained as

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

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

$$\hat{g}_\theta(x,1) = \mathbf{1}_{\left\{1 \le \hat{\eta}(x,1)\left(2 - \frac{\theta}{\hat{\mathbb{P}}(Y=1,S=1)}\right)\right\}}, \quad \hat{g}_\theta(x,0) = \mathbf{1}_{\left\{1 \le \hat{\eta}(x,0)\left(2 + \frac{\theta}{\hat{\mathbb{P}}(Y=1,S=0)}\right)\right\}} \ . \tag{5}$$

In this case, the algorithm $\hat{g}$ that we propose is such that $\hat{g} \equiv \hat{g}_{\hat{\theta}}$. It is crucial to mention that since the quantity $\hat{\Delta}(\hat{g}_\theta, \mathbb{P})$ is empirical, then there might be no θ which delivers zero for the empirical unfairness. This is exactly the reason we perform a minimization of this quantity.

Remark 3.3. *Even though we believe that the introduction of the unlabeled sample is one of the strong points of our approach, this sample may not be available on some benchmark datasets. In this case, we can simply randomly split the data into two parts disregarding labels in one of them, or alternatively we can use the same sample twice. The second path is not directly justified by our theoretical results, yet, let us suggest the following intuitive explanation for this approach. On the first and the second steps, our procedure approximates two* independent *parts of the distribution $\mathbb{P}$ of the random tuple (X, S, Y). Indeed, following the factorization $\mathbb{P} = \mathbb{P}_{Y|X,S} \otimes \mathbb{P}_{(X,S)}$, the first step of our procedure approximates $\mathbb{P}_{Y|X,S}$, whereas the second step is aimed at $\mathbb{P}_{(X,S)}$ which is independent from $\mathbb{P}_{Y|X,S}$. In our experiments, reported in Section 5, we exploited the same set of data for both $\mathcal{D}_n$ and $\mathcal{D}_N$, since no unlabelled sample were available and splitting the dataset would have reduced the quality of the trained model because the datasets have a small sample size.*

4 Consistency

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

Assumption 4.1. *The estimator $\hat{\eta}$ which is constructed on $\mathcal{D}_n$ satisfies for all $s \in \{0, 1\}$*

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

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

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

Remark 4.2. *There are two parts in Assumption 4.1, the first one requires a consistent estimator in* ℓ_1 *norm. This first assumption is rather weak, since there are many different available consistent estimators for the regression function in the literature, including the Maximum likelihood estimator [45] for Gaussian Generative Model, local polynomial estimator [4] for* β*-Hölder smooth regression function* $\eta(\cdot, s)$*, regularized logistic regression [42] for Generalized Linear Model,* k*-Nearest Neighbors estimator [16] for Lipschitz regression function* $\eta(\cdot, s)$*, and random forest type estimators in various settings [3, 7, 21, 41].*
The second part of Assumption 4.1 means that $\mathbb{E}_{X|S=s}[\hat{\eta}(X,s)]$ *is lower bounded by a positive term vanishing as* N, n *grow to infinity. This condition can be introduced artificially to any predefined estimator. Indeed, assume that we have a consistent estimator* $\tilde{\eta}$ *and let* $\hat{\eta}(x,s) = \max\{\tilde{\eta}(x,s), c_{n,N}\}$*, then the second item of the assumption is satisfied in even a stronger form. Moreover, this estimator* $\hat{\eta}$ *remains consistent, since using the triangle inequality and the fact that* $|\hat{\eta}(x,s) - \tilde{\eta}(x,s)| \leq c_{n,N}$ *for all* $x \in \mathbb{R}^d$*, we have*

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

Additionally, we impose one more condition on the estimator $\hat{\eta}$ that was already successfully used in the context of confidence set classification [11, 15].

Assumption 4.3. *The estimator* $\hat{\eta}$ *is such that for all* $s \in \{0,1\}$ *the mapping*

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

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

In our settings this assumption allows us to show that the value of $\hat{\Delta}(\hat{g}, \mathbb{P})$ cannot be large, that is, the empirical unfairness of the proposed procedure is small or zero. As we shall see, a control on the empirical unfairness $\hat{\Delta}(\hat{g}, \mathbb{P})$ in Definition 3.2 is crucial in proving that the proposed procedure $\hat{g}$ achieves both asymptotic fairness and risk consistency.

Remark 4.4. *Assumption 4.3 is equivalent to say that there are no atoms in the estimated regression function. It can be fulfilled by a simple modification of any preliminary estimator, by adding a* small *deterministic "noise", the amplitude of which must be decreasing with* n, N *in order to preserve statistical consistency.*

Our remarks suggest that both Assumptions 4.1 and 4.3 can be easily satisfied in a variety of practical settings and the most demanding part of these assumptions is the consistency of $\hat{\eta}$.

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

Theorem 4.5 (Asymptotic properties)**.** *Under Assumptions 2.2, 4.1, and 4.3 the proposed algorithm satisfies*

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

Proof sketch. In order to establish statistical consistency of the proposed procedure, we follow the strategy of [11, 15], that is, we first introduce an intermediate pseudo-estimator $\tilde{g}$ as follows

$$\tilde{g}(x,1) = \mathbf{1}_{\left\{1 \leq \hat{\eta}(x,1)\left(2 - \frac{\tilde{\theta}}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]\mathbb{P}(S=1)}\right)\right\}}, \ \tilde{g}(x,0) = \mathbf{1}_{\left\{1 \leq \hat{\eta}(x,0)\left(2 + \frac{\tilde{\theta}}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]\mathbb{P}(S=0)}\right)\right\}}, \tag{6}$$

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

$$\frac{\mathbb{E}_{X|S=1}\left[\hat{\eta}(X,1)\tilde{g}(X,1)\right]}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]} = \frac{\mathbb{E}_{X|S=0}\left[\hat{\eta}(X,0)\tilde{g}(X,0)\right]}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]} \ . \tag{7}$$

Note that by Assumption 4.3 such a value $\tilde{\theta}$ always exists. Intuitively, the classifier $\tilde{g}$ "*knows*" the marginal distribution of (X, S), that is, it knows both $\mathbb{P}_{X|S}$ and $\mathbb{P}_S$. It is seen as an idealized version of $\hat{g}$, where the uncertainty is only induced by the lack of knowledge of the regression function η.

We express the excess risk as a sum of two terms, $\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\tilde{g})] - \mathcal{R}(g^*) + \mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}[\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})]$. We show that the first can be bounded by the ℓ_1 distance between $\hat{\eta}$ and η, and thanks to the consistency of $\hat{\eta}$ it does converge to zero. The handling of the second term is move involved, but we are able to show that it reduces to a study of suprema of empirical processes conditionally on the labeled sample $\mathcal{D}_n$.

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

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

At last, the continuity Assumption 4.3 alongside with means of theory of empirical processes allow to demonstrate that the term $\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}[\hat{\Delta}(\hat{g},\mathbb{P})]$ converges to zero when N growth. □

Remark 4.6. *Let us mention that it is possible to present our result in a finite sample regime, since our proof of consistency is based on non-asymptotic theory of empirical processes. However, the actual rate of convergence depends on the rate of ℓ_1-norm estimation of the regression function η, which can vary significantly from one setup to another. That is why we decided to present our result in the asymptotic sense.*

5 Experimental results

In this section, we present numerical experiments with the proposed method. The source code we used to perform the experiments can be found at `https://github.com/lucaoneto/NIPS2019_Fairness`.

We follow the protocol outlined in [17]. We consider the following datasets: Arrhythmia, COMPAS, Adult, German, and Drug[2] and compare the following algorithms: Linear Support Vector Machines (Lin.SVM), Support Vector Machines with the Gaussian kernel (SVM), Linear Logistic Regression (Lin.LR), Logistic Regression with the Gaussian kernel (LR), Hardt method [22] to all approaches (Hardt), Zafar method [48] implemented with the code provided by the authors for the linear case[3], the Linear (Lin.Donini) and the Non Linear methods (Donini) proposed in [17] and freely available[4], and also Random Forests (RF). Then, since Lin.SVM, SVM, Lin.LR, LR, and RF have also the possibility to output a probability together with the classification, we applied our method in all these cases.

In all experiments, we collect statistics concerning the classification accuracy (ACC), namely probability to correctly classify a sample, and the Difference of Equal Opportunity (DEO) in Definition 2.1. For Arrhythmia, COMPAS, German and Drug datasets we split the data in two parts (70% train and 30% test), this procedure is repeated 30 times, and we reported the average performance on the test set alongside its standard deviation. For the Adult dataset, we used the

[2]For more information about these datasets please refer to [17].

[3]Python code for [48]: `https://github.com/mbilalzafar/fair-classification`

[4]Python code for [17]: `https://github.com/jmikko/fair_ERM`

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

Table 1: Results (average ± standard deviation, when a fixed test set is not provided) for all the datasets, concerning ACC and DEO.

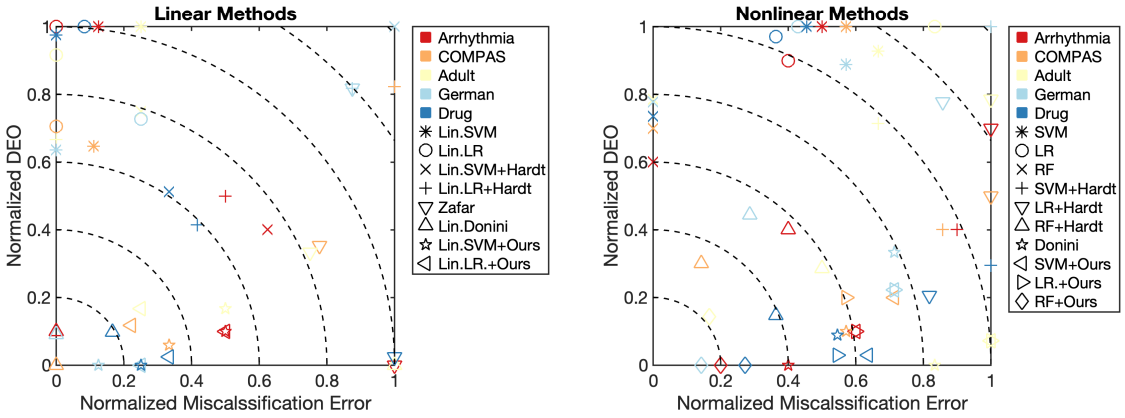

Figure 1: Results of Table 1 of linear (left) and nonlinear (right) methods when the error and the DEO are normalized in $[0, 1]$ column-wise. Different colors and symbols refer to different datasets and method respectively. The closer a point is to the origin, the better the result is.

RF+Ours	COMPAS		Adult	
	ACC	DEO	ACC	DEO
$\mathcal{D}_n$=1/10	0.68 ± 0.03	0.07 ± 0.02	0.79 ± 0.02	0.06 ± 0.02
$\mathcal{D}_n$=1/10, $\mathcal{D}_N$=1/10	0.68 ± 0.03	0.07 ± 0.02	0.79 ± 0.02	0.06 ± 0.02
$\mathcal{D}_n$=1/10, $\mathcal{D}_N$=2/10	0.68 ± 0.03	0.07 ± 0.02	0.79 ± 0.02	0.06 ± 0.02
$\mathcal{D}_n$=1/10, $\mathcal{D}_N$=4/10	0.70 ± 0.02	0.06 ± 0.02	0.79 ± 0.02	0.05 ± 0.01
$\mathcal{D}_n$=1/10, $\mathcal{D}_N$=8/10	0.71 ± 0.02	0.05 ± 0.01	0.80 ± 0.02	0.04 ± 0.01

Table 2: Impact of the size of the unlabeled dataset on ACC and DEO. The size of the labeled sample $\mathcal{D}_n$ is fixed to 1/10 of the original dataset. The unlabeled $\mathcal{D}_N$ is initially empty (as in previous experiments of Table 1), and then increases from 1/10 to 8/10 of the original dataset.

provided split of train and test sets. Unless otherwise stated, we employ two steps in the 10-fold CV procedure proposed in [17] to select the best hyperparameters with the training set[5]. In the first step, the value of the hyperparameters with the highest accuracy is identified. In the second step, we shortlist all the hyperparameters with accuracy close to the best one (in our case, above 90% of the best accuracy). Finally, from this list, we select the hyperparameters with the lowest DEO.

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

From Table 1 and Figure 1 it is possible to observe that the proposed method outperforms all methods except the one of [17] for which we obtain comparable performance. Nevertheless, note that our method is more general than the one of [17], since it can be applied to any algorithms which return a probability estimator (better if consistent since this will allow us to have a full consistent approach also from the fairness point of view). In fact, on these datasets, RF, which cannot be made trivially fair with the approach proposed in [17], outperforms all the available methods.

Note that the results reported in Table 1 differ from the one reported in [17] since the proposed method requires the knowledge of the sensitive variable at classification time, so Table 1 reports just this case. That is, the functional form of the model explicitly depends on the sensitive variable $s \in \{0, 1\}$. Many authors, point out that this may not be permitted in several practical scenarios (see e.g. [19, 39] and reference therein). Yet, removing the sensitive variable from the functional form of the model does not ensure that the sensitive variable is not considered by the model itself. We refer to [36] for the in-depth discussion on this issue. Further, the method in [22] explicitly requires the knowledge of the sensitive variable for their thresholding procedure. In Appendix E we show how to modify our method in order to derive a fair optimal classifier without the sensitive variable s in the functional form of the model. Moreover, we propose a modification of our approach which does not use s at decision time and perform additional numerical comparison in this context. We arrive to similar conclusions about the performance of our method as in this section. Yet, the consistency results are not available for this methods and are left for future investigation.

In Table 2 we demonstrate the impact of the unlabeled data size on the performance of the proposed algorithm. Since the above benchmark datasets are not provide with additional unlabeled data, we deploy the following data generation procedure: we randomly select 1/10 observations in each dataset and assign it to the labeled sample $\mathcal{D}_n$; consequently, the size of the unlabeled sample $\mathcal{D}_N$ increases from 0 to 8/10 samples that were not assigned to the labeled sample $\mathcal{D}_n$. This data generation procedure is applied to COMPAS and Adult datasets. Finally, we apply our method using the random forest algorithm using the cross-validation scheme employed in the previous experiments. The above above pipeline is repeated 30 times and the variance of the results is reported in Table 2. We can see that both DEO and ACC are improving with N, highlighting the importance of the unlabeled data. We believe that the improvement could have been more significant if the unlabeled data were provided initially.

[5]The regularization parameter (for all method) and the RBF kernel with 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 search in $\{d, \lceil d^{15/16} \rceil, \lceil d^{7/8} \rceil, \lceil d^{3/4} \rceil, \lceil d^{1/2} \rceil, \lceil d^{1/4} \rceil, \lceil d^{1/8} \rceil, \lceil d^{1/16} \rceil, 1\}$ where d is the number of features in the dataset.

6 Conclusion

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

Acknowledgments

This work was supported in part by SAP SE, by the Amazon AWS Machine Learning Research Award and by the Labex Bézout of Université Paris-Est.

References

[1] J. Adebayo and L. Kagal. Iterative orthogonal feature projection for diagnosing bias in black-box models. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

[2] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

[3] S. Arlot and R. Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.

[4] J. Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

[5] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018.

[6] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[7] L. Breiman. Consistency for a simple model of random forests. Technical report, Statistics Department University Of California At Berkeley, 2004.

[8] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *IEEE international conference on Data mining*, 2009.

[9] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Neural Information Processing Systems*, 2017.

[10] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Neural Information Processing Systems*, 2017.

[11] E. Chzhen, C. Denis, and M. Hebiri. Minimax semi-supervised confidence sets for multi-class classification. *arXiv preprint arXiv:1904.12527*, 2019.

[12] A. Cotter, M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, and S. You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. *arXiv preprint arXiv:1807.00028*, 2018.

[13] F. Cribari-Neto, N. Garcia, and K. Vasconcellos. A note on inverse moments of binomial variates. *Brazilian Review of Econometrics*, 20(2):269–277, 2000.

[14] C. Denis and M. Hebiri. Confidence sets with expected sizes for multiclass classification. *Journal of Machine Learning Research*, 18(1):3571–3598, 2017.

[15] C. Denis and M. Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *Journal of Nonparametric Statistics*, 0(0):1–31, 2019.

[16] L. Devroye. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory*, 24(2):142–151, 1978.

[17] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Neural Information Processing Systems*, 2018.

[18] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

[19] C. Dwork, N. Immorlica, A. T. Kalai, and M. D. M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, 2018.

[20] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining*, 2015.

[21] R. Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562, 2012.

[22] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Neural Information Processing Systems*, 2016.

[23] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fair learning in markovian environments. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

[24] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Neural Information Processing Systems*, 2016.

[25] F. Kamiran and T. Calders. Classifying without discriminating. In *International Conference on Computer, Control and Communication*, 2009.

[26] F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Machine Learning Conference*, 2010.

[27] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[28] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Neural Information Processing Systems*, 2017.

[29] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

[30] O. Koyejo, N. Natarajan, P. Ravikumar, and I. Dhillon. Consistent multilabel classification. In *Neural Information Processing Systems*, 2015.

[31] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Neural Information Processing Systems*, 2017.

[32] J. Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.

[33] K. Lum and J. Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.

[34] P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[35] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, 2018.

[36] L. Oneto, M. Donini, A. Elders, and M. Pontil. Taking advantage of multitask learning for fair classification. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

[37] L. Oneto, M. Donini, and M. Pontil. General fair empirical risk minimization. *arXiv preprint arXiv:1901.10080*, 2019.

[38] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Weinberger. On fairness and calibration. In *Neural Information Processing Systems*, 2017.

[39] J. E. Roemer and A. Trannoy. Equality of opportunity. In *Handbook of income distribution*, 2015.

[40] M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, pages 1–12, 2018.

[41] E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *Ann. Statist.*, 43(4):1716–1741, 08 2015.

[42] S. Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[43] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, 2015.

[44] J. Wellner. Empirical processes: Theory and applications. Technical report, Delft University of Technology, 2005.

[45] B. Yan, S. Koyejo, K. Zhong, and P. Ravikumar. Binary classification with karmic, threshold-quasi-concave metrics. In *International Conference on Machine Learning*, 2018.

[46] Y. Yang. Minimax nonparametric classification: Rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.

[47] S. Yao and B. Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Neural Information Processing Systems*, 2017.

[48] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, 2017.

[49] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

[50] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, 2013.

[51] M. J. Zhao, N. Edakunni, A. Pocock, and G. Brown. Beyond fano's inequality: bounds on the optimal f-score, ber, and cost-sensitive risk and their implications. *Journal of Machine Learning Research*, 14:1033–1090, 2013.

[52] I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

Supplementary material for "Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification"

A Optimal classifier

Proof of Proposition 2.3. Let us study the following minimization problem

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

Using the weak duality we can write

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

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

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

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

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

moreover, we can write for the rest

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

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

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

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

$$g^*_\lambda(x,1) = \mathbf{1}_{\left\{\eta(X,1)\left(\frac{\lambda}{\mathbb{P}(Y=1 \mid S=1)} - 2\mathbb{P}(S=1)\right) + \mathbb{P}(S=1) \leq 0\right\}} = \mathbf{1}_{\left\{1 - \eta(X,1)\left(2 - \frac{\lambda}{\mathbb{P}(Y=1,S=1)}\right) \leq 0\right\}}$$
$$g^*_\lambda(x,0) = \mathbf{1}_{\left\{\eta(X,0)\left(-\frac{\lambda}{\mathbb{P}(Y=1 \mid S=0)} - 2\mathbb{P}(S=0)\right) + \mathbb{P}(S=0) \leq 0\right\}} = \mathbf{1}_{\left\{1 - \eta(X,0)\left(2 + \frac{\lambda}{\mathbb{P}(Y=1,S=0)}\right) \leq 0\right\}} \ .$$

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

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

It is important to observe that the mappings

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

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

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

Clearly, under Assumption 2.2 this subgradient is reduced to the gradient almost surely, thus we have the following condition on the optimal value of λ^*

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

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

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

This implies that the classifier $g^*_{\lambda^*}$ is fair, that is, it satisfies Definition 2.1. Finally, it remains to show that $g^*_{\lambda^*}$ is actually an optimal classifier, indeed, since $g^*_{\lambda^*}$ is fair we can write on the one hand

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

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

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

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

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

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

where the last inequality is due to Assumption 2.2. We arrive to contradiction, therefore $(2-\theta^*/\mathbb{P}(Y=1,S=1)) > 0$. Similarly, we show that $(2+\theta^*/\mathbb{P}(Y=1,S=0)) > 0$. Combination of both inequalities and the fact that for all $s \in \{0,1\}$ we have $\mathbb{P}(Y=1,S=s) \leq 1$ implies that $|\theta^*| \leq 2$. □

B Auxiliary results

Before proceeding to the proof of our main result in Theorem 4.5, let us first introduce several auxiliary results. We suggest the reader to first understand these results omitting its proofs before proceeding further. We will use $C > 0$ as a generic constant which actually could be different from line to line, yet, this constant is always independent from n, N.

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

The next lemma can be found in [13].

Lemma B.2. *Let* Z *be a binomial random variable with parameters* N, p*, then for every* $\alpha \in \mathbb{R}$

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

Lemma B.3. *For any classifier* g *we have*

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

Proof. We can write

$$\begin{aligned}\mathcal{R}(g) &:= \mathbb{P}(Y \neq g(X,S)) = \mathbb{E}[Y(1-g(X,S))] + \mathbb{E}[(1-Y)g(X,S)] \\ &= \mathbb{E}_{(X,S)}\eta(X,S)(1-g(X,S)) + \mathbb{E}_{(X,S)}(1-\eta(X,S))g(X,S) \\ &= \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\eta(X,S)-1)g(X,S)] \ .\end{aligned}$$

□

In what follows we shall often use the relations:

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

which holds for all $s \in \{0,1\}$.

C Proof of Theorem 4.5

Below we gather extra tools which are directly related to the proof of our main result, proof are provided in Appendix D. First lemma gives an upper on the quantity of unfairness $\Delta(g,\mathbb{P})$ in terms of its empirical version in Definition 3.2.

Lemma C.1. *Let g be* any *classifier (data depended or not) and $\hat{\eta}$ be an estimator of the regression function η constructed on $\mathcal{D}_n$. Then, almost surely we have*

$$\begin{aligned}\Delta(g,\mathbb{P}) \leq \hat{\Delta}(g,\mathbb{P}) &+ \underbrace{2\frac{\mathbb{E}_{X|S=1}\left|\eta(X,1)-\hat{\eta}(X,1)\right|}{\mathbb{P}\left(Y=1 \,|\, S=1\right)}}_{\textit{how good is } \hat{\eta}} + \underbrace{2\frac{\mathbb{E}_{X|S=0}\left|\eta(X,0)-\hat{\eta}(X,0)\right|}{\mathbb{P}\left(Y=1 \,|\, S=0\right)}}_{\textit{how good is } \hat{\eta}}\\ &+ \underbrace{\frac{\left|(\mathbb{E}_{X|S=1}-\hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)g(X,1)\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)}}_{\textit{empirical process}} + \underbrace{\frac{\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)g(X,0)\right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)}}_{\textit{empirical process}}\\ &+ \frac{\left|(\hat{\mathbb{E}}_{X|S=1}-\mathbb{E}_{X|S=1})\hat{\eta}(X,1)\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} + \frac{\left|(\hat{\mathbb{E}}_{X|S=0}-\mathbb{E}_{X|S=0})\hat{\eta}(X,0)\right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)}\ .\end{aligned}$$

Let us elaborate on the above result. The second and the third terms are responsible for the estimation of η and can be controlled in various parametric on nonparametric models. The third and the fourth terms can be handled with the theory of empirical processes in the considered classifier g is data dependent. The last two terms can be handled conditionally on the first labeled samples by the use of the multiplicative Chernoff's bound or (if we do not mind losing a constant factor of 2) can be handled by the empirical process used to bound the third and the fourth terms.

The next lemma gives an upper bound on the empirical processes of Lemma C.1.

Lemma C.2. *There exists a constant $C>0$ that depends only on $\mathbb{P}(S=0)$ and $\mathbb{P}(S=1)$ such that almost surely for all $s \in \{0,1\}$ we have*

$$\mathbb{E}_{\mathcal{D}_N}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=s}-\hat{\mathbb{E}}_{X|S=s})\hat{\eta}(X,s)\mathbf{1}_{\{t\leq\hat{\eta}(X,s)\}}\right| \leq C\sqrt{\frac{1}{N}}\ .$$

The next result is obvious, yet, is used several times in our proof, that is why we present it separately.

Lemma C.3. *For any functions $h_1, h_0 : \mathbb{R}^d \to [0,1]$, any $\theta \in \mathbb{R}$, any $a_1, a_0, b_1, b_0 \in (0,1)$ we have*

$$\begin{aligned}
&\mathbb{E}_{X|S=1}\left[\frac{\theta h_1(X)}{a_1}\mathbf{1}_{\left\{b_1(2h_1(X)-1)-\frac{\theta h_1(X)}{a_1}\geq 0\right\}}\right] \\
&\qquad = \mathbb{E}_{X|S=1}\left[(2h_1(X)-1)\mathbf{1}_{\left\{b_1(2h_1(X)-1)-\frac{\theta h_1(X)}{a_1}\geq 0\right\}}\right] b_1 \\
&\qquad - \mathbb{E}_{X|S=1}\left(b_1(2h_1(X)-1) - \frac{\theta h_1(X)}{a_1}\right)_+ , \\
&\mathbb{E}_{X|S=0}\left[\frac{\theta h_0(X)}{a_0}\mathbf{1}_{\left\{b_0(2h_0(X)-1)+\frac{\theta h_0(X)}{a_0}\geq 0\right\}}\right] \\
&\qquad = -\mathbb{E}_{X|S=0}\left[(2h_0(X)-1)\mathbf{1}_{\left\{b_0(2h_0(X)-1)+\frac{\theta h_0(X)}{a_0}\geq 0\right\}}\right] b_0 \\
&\qquad + \mathbb{E}_{X|S=0}\left(b_0(2h_0(X)-1) + \frac{\theta h_0(X)}{a_0}\right)_+ ,
\end{aligned}$$

moreover, the expectation $\mathbb{E}_{X|S=s}$ can be replaced by $\hat{\mathbb{E}}_{X|S=s}$ for all $s \in \{0,1\}$.

C.1 Proof of asymptotic fairness (Part I of Theorem 4.5)

Proof. The first step is to show that under Assumption 4.3 the term $\hat{\Delta}(\hat{g}, \mathbb{P})$ cannot be too big. Indeed, notice that for every $\theta \in [-2, 2]$, thanks to the triangle inequality we can write almost surely

$$\begin{aligned}
\hat{\Delta}(\hat{g}_\theta, \mathbb{P}) \leq &\left|\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)}\right| \\
&+ \left|\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)}\right| \\
&+ \left|\frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} - \frac{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)}\right| .
\end{aligned} \tag{8}$$

Our goal is to take care of each of the three terms appearing on the right hand side of the inequality. The technique used for the second and the third term is identical, whereas the first term is a bit more involved. Let us start with the second term on the right hand side of Eq. (8). For this term we

can write almost surely

$$
\begin{aligned}
&\left|\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)}\right| \\
&\leq \left|\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)}\right| \\
&+ \left|\frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)}\right| \\
&= \frac{\left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\hat{g}_\theta(X,1)\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \\
&+ \underbrace{\frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)}}_{\leq 1} \frac{\left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\mathbf{1}_{\{0\leq\hat{\eta}(X,1)\}}\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \\
&\leq 2\frac{\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\mathbf{1}_{\{t\leq\hat{\eta}(X,1)\}}\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \ ,
\end{aligned}
$$

where the last inequality follows from the fact that $\hat{g}_\theta$ is a thresholding rule. Similarly, we show that the third term in Eq. (8) admits the following bound almost surely

$$
\begin{aligned}
&\left|\frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} - \frac{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)}\right| \\
&\leq 2\frac{\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0} - \hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbf{1}_{\{t\leq\hat{\eta}(X,0)\}}\right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \ .
\end{aligned}
$$

Therefore, we arrive at the following bound on $\hat{\Delta}(\hat{g}_\theta, \mathbb{P})$ which holds almost surely

$$
\begin{aligned}
\hat{\Delta}(\hat{g}_\theta, \mathbb{P}) \leq &\left|\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_\theta(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\hat{g}_\theta(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)}\right| \\
&+ 2\frac{\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\mathbf{1}_{\{t\leq\hat{\eta}(X,1)\}}\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \\
&+ 2\frac{\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0} - \hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbf{1}_{\{t\leq\hat{\eta}(X,0)\}}\right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \ .
\end{aligned} \tag{9}
$$

This is one of the moments when we make use of Assumption 4.3. Thanks to the continuity we can be sure that for every possible unlabeled sample $\mathcal{D}_N$, there exists $\theta'(\mathcal{D}_N)$ such that

$$
\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)\hat{g}_{\theta'(\mathcal{D}_N)}(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} = \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\hat{g}_{\theta'(\mathcal{D}_N)}(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \ .
$$

Indeed, for every possible unlabeled sample $\mathcal{D}_N$ on the left hand side we have a continuous decreasing of θ function and on the right hand side we have a continuous increasing function of θ. Therefore, such a value $\theta'(\mathcal{D}_N)$ exists.

Taking infimum over $\theta \in [-2,2]$ on both sides of Equation (9) we obtain

$$\hat{\Delta}(\hat{g},\mathbb{P}) = \hat{\Delta}(\hat{g}_{\hat{\theta}},\mathbb{P}) \leq 2\frac{\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=1}-\hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\mathbf{1}_{\{t\leq\hat{\eta}(X,1)\}}\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} + 2\frac{\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbf{1}_{\{t\leq\hat{\eta}(X,0)\}}\right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \ . \tag{10}$$

Using Lemma C.1 and applying it to $\hat{g}$ we immediately obtain almost surely

$$\begin{aligned}\Delta(\hat{g},\mathbb{P}) \leq & 4\frac{\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=1}-\hat{\mathbb{E}}_{X|S=1})\hat{\eta}(X,1)\mathbf{1}_{\{t\leq\hat{\eta}(X,1)\}}\right|}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \\ & + 4\frac{\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbf{1}_{\{t\leq\hat{\eta}(X,0)\}}\right|}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} \\ & + 2\frac{\mathbb{E}_{X|S=1}\left|\eta(X,1)-\hat{\eta}(X,1)\right|}{\mathbb{P}\left(Y=1\,|\,S=1\right)} + 2\frac{\mathbb{E}_{X|S=0}\left|\eta(X,0)-\hat{\eta}(X,0)\right|}{\mathbb{P}\left(Y=1\,|\,S=0\right)} \ .\end{aligned}$$

Clearly, if $\hat{\eta}$ is a consistent estimator of η then the last two terms on the right hand side are converging to zero in expectation as $n \to \infty$. Therefore, it remain to provide an upper bound for the two empirical processes. Recall, that our goal is to obtain consistency in expectation, thus we take expectation *w.r.t.* $\mathcal{D}_n, \mathcal{D}_N$ from both sides of the inequality. Thanks to Lemma C.2 we have for each $s \in \{0,1\}$

$$\mathbb{E}_{\mathcal{D}_N}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=s}-\hat{\mathbb{E}}_{X|S=s})\hat{\eta}(X,s)\mathbf{1}_{\{t\leq\hat{\eta}(X,s)\}}\right| \leq C\sqrt{\frac{1}{N}} \ .$$

The arguments above imply that there exists an absolute constant $C > 0$ such that

$$\begin{aligned}\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}[\Delta(\hat{g},\mathbb{P})] \leq & 2\frac{\mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{X|S=1}\left|\eta(X,1)-\hat{\eta}(X,1)\right|}{\mathbb{P}\left(Y=1\,|\,S=1\right)} + 2\frac{\mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{X|S=0}\left|\eta(X,0)-\hat{\eta}(X,0)\right|}{\mathbb{P}\left(Y=1\,|\,S=0\right)} \\ & + C\sqrt{\frac{1}{N}}\mathbb{E}_{\mathcal{D}_n}\frac{1}{\min\{\mathbb{E}_{X|S=1}\hat{\eta}(X,1),\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\}} \ .\end{aligned}$$

Using the second item of Assumption 4.1, which states that $\min\{\mathbb{E}_{X|S=1}\hat{\eta}(X,1),\mathbb{E}_{X|S=0}\hat{\eta}(X,0)\} \geq c_{n,N}$ almost surely we conclude. □

C.2 Proof of asymptotic optimality (Part II of Theorem 4.5)

In order to show that the risk of the proposed algorithm converges to the risk of the optimal classifier, we follow the strategy of [11], that is, we first introduce an intermediate pseudo-estimator $\tilde{g}$ as follows

$$\tilde{g}(x,1) = \mathbf{1}_{\left\{\mathbb{P}(S=1)\leq\hat{\eta}(x,1)\left(2\mathbb{P}(S=1)-\frac{\tilde{\theta}}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)\right\}} \ , \tag{11}$$

$$\tilde{g}(x,0) = \mathbf{1}_{\left\{\mathbb{P}(S=0)\leq\hat{\eta}(x,0)\left(2\mathbb{P}(S=0)+\frac{\tilde{\theta}}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)\right\}} \ , \tag{12}$$

where $\tilde{\theta}$ is a solution of

$$\frac{\mathbb{E}_{X|S=1}\left[\hat{\eta}(X,1)\tilde{g}_\theta(X,1)\right]}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]} = \frac{\mathbb{E}_{X|S=0}\left[\hat{\eta}(X,0)\tilde{g}_\theta(X,0)\right]}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]} \ , \tag{13}$$

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

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

Note that thanks to Assumption 4.3 such a value $\tilde{\theta}$ always exists.

Intuitively, the classifier $\tilde{g}$ knows the marginal distribution of (X, S), that is, it knows both $\mathbb{P}_{X|s}$ and $\mathbb{P}_S$. It is seen as an idealized version of $\hat{g}$, where the uncertainty is only induced by the lack of knowledge of the regression function η. We upper bound the excess risk in two steps. In the first step we upper bound $\mathcal{R}(\tilde{g}) - \mathcal{R}(g^*)$ and on the second we upper bound the difference $\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})$.

Theorem C.4 (Bound on the pseudo oracle)**.** *Let $\tilde{g}$ be the pseudo oracle classifier defined in Eq. 11 with $\hat{\eta}$ satisfying Assumptions 4.1 and 4.3, then*

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\tilde{g})] - \mathcal{R}(g^*) \leq 0 \ .$$

Proof of Theorem C.4. First of all, let us rewrite the equation for θ^* in the following form

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

Using Lemma C.3 with $h_s(\cdot) \equiv \eta(\cdot, s), a_s = \mathbb{E}_{X|S=1}[h_s(\cdot)], b_s = \mathbb{P}(S = s)$ for $s \in \{0, 1\}$ we get

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

Rearranging the terms we can arrive at

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

Notice that the left hand side of the above equality can be written as

$$\begin{aligned}
&\mathbb{E}_{(X,S)}[(2\eta(X,S)-1)g^*(X,S)] \\
&= \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1) - \frac{\theta^*\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+ \\
&\quad + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1) + \frac{\theta^*\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ .
\end{aligned} \tag{14}$$

Thus, combining the previous equality with the expression of the risk from Lemma B.3 we get

$$\begin{aligned}
\mathcal{R}(g^*) = \mathbb{E}_{(X,S)}[\eta(X,S)] &- \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1) - \frac{\theta^*\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+ \\
&- \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1) + \frac{\theta^*\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ .
\end{aligned} \tag{15}$$

Step-wise similar argument yields that for the pseudo-oracle $\tilde{g}$ we can write

$$\begin{aligned}
&\mathbb{E}_{(X,S)}[(2\hat{\eta}(X,S)-1)\tilde{g}(X,S)] \\
&= \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&\quad + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned} \tag{16}$$

Moreover, its risk satisfies

$$\begin{aligned}
\mathcal{R}(\tilde{g}) &= \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\eta(X,S)-1)\tilde{g}(X,S)] \\
&\leq \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\hat{\eta}(X,S)-1)\tilde{g}(X,S)] + 2\mathbb{E}_{(X,S)}\left|\hat{\eta}(X,S)-\eta(X,S)\right| .
\end{aligned} \tag{17}$$

Therefore, combining Eq. (15) with Eq. (17), we can write for the excess risk

$$\begin{aligned}
\mathcal{R}(\tilde{g}) - \mathcal{R}(g^*) \leq &2\mathbb{E}_{(X,S)}\left|\hat{\eta}(X,S)-\eta(X,S)\right| \\
&+ \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1) - \frac{\theta^*\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+ \\
&- \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&+ \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1) + \frac{\theta^*\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ \\
&- \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned}$$

Recall that θ^* is a minimizer of

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

thus we can replace θ^* by $\tilde{\theta}$ and obtain the following upper bound

$$\begin{aligned}
\mathcal{R}(\tilde{g}) - \mathcal{R}(g^*) \leq & 2\mathbb{E}_{(X,S)} \left|\hat{\eta}(X,S) - \eta(X,S)\right| \\
& + \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\eta(X,1)-1) - \frac{\tilde{\theta}\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right)_+ \\
& - \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\eta(X,0)-1) + \frac{\tilde{\theta}\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right)_+ \\
& - \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned}$$

Since, for all $x, y \in \mathbb{R}$ we have $(x)_+ - (y)_+ \leq (x-y)_+ \leq |x-y|$ we get

$$\begin{aligned}
\mathcal{R}(\tilde{g}) - \mathcal{R}(g^*) \leq & 4\mathbb{E}_{(X,S)} \left|\hat{\eta}(X,S) - \eta(X,S)\right| \\
& + \mathbb{E}_{X|S=1}|\tilde{\theta}| \left|\frac{\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]} - \frac{\eta(X,1)}{\mathbb{E}_{X|S=1}[\eta(X,1)]}\right| \\
& + \mathbb{E}_{X|S=0}|\tilde{\theta}| \left|\frac{\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]} - \frac{\eta(X,0)}{\mathbb{E}_{X|S=0}[\eta(X,0)]}\right| .
\end{aligned}$$

For the same reason why $|\theta^*| \leq 2$ we have $|\tilde{\theta}| \leq 2$, thus for all $s \in \{0,1\}$ we have

$$\begin{aligned}
& \mathbb{E}_{X|S=s}|\tilde{\theta}| \left|\frac{\hat{\eta}(X,s)}{\mathbb{E}_{X|S=s}[\hat{\eta}(X,s)]} - \frac{\eta(X,s)}{\mathbb{E}_{X|S=s}[\eta(X,s)]}\right| \\
& \leq 2\mathbb{E}_{X|S=s} \left|\frac{\hat{\eta}(X,s)}{\mathbb{E}_{X|S=s}[\hat{\eta}(X,s)]} - \frac{\eta(X,s)}{\mathbb{E}_{X|S=s}[\eta(X,s)]}\right| \\
& \leq 2\mathbb{E}_{X|S=s} \left|\frac{\hat{\eta}(X,s)}{\mathbb{E}_{X|S=s}[\eta(X,s)]} - \frac{\eta(X,s)}{\mathbb{E}_{X|S=s}[\eta(X,s)]}\right| \\
& + 2\mathbb{E}_{X|S=s} \left|\frac{\hat{\eta}(X,s)}{\mathbb{E}_{X|S=s}[\hat{\eta}(X,s)]} - \frac{\hat{\eta}(X,s)}{\mathbb{E}_{X|S=s}[\eta(X,s)]}\right| \\
& \leq 4\frac{\mathbb{E}_{X|S=s}\left|\eta(X,s) - \hat{\eta}(X,s)\right|}{\mathbb{E}_{X|S=s}[\eta(X,s)]} .
\end{aligned}$$

Thanks to Assumption 4.1, these terms converge to zero in expectation. □

Theorem C.5. *Let $\hat{g}$ be the proposed classifier with $\hat{\eta}$ satisfying Assumptions 4.1 and 4.3, then*

$$\lim_{n\to\infty} \mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}[\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})] \leq 0 .$$

Proof. Our goal is to upper bound the quantity $\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})$. We start by providing a

bound on $\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})$ which holds almost surely. Recall the equality of Equation (16)

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

Using this and the expression of the risk given in Lemma B.3 we can obtain the following lower bound on the risk of $\tilde{g}$

$$\begin{aligned}
\mathcal{R}(\tilde{g}) &= \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\eta(X,S)-1)\tilde{g}(X,S)] \\
&\geq \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\hat{\eta}(X,S)-1)\tilde{g}(X,S)] - 2\mathbb{E}_{(X,S)}\left|\hat{\eta}(X,S)-\eta(X,S)\right| \\
&= \mathbb{E}_{(X,S)}[\eta(X,S)] - 2\mathbb{E}_{(X,S)}\left|\hat{\eta}(X,S)-\eta(X,S)\right| \\
&\quad - \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
&\quad - \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned} \tag{18}$$

We have thanks to Lemma C.3 used with $h_s(\cdot) = \hat{\eta}(\cdot, s), a_s = \hat{\mathbb{E}}_{X|S=s}[h_s(X)], b_s = \hat{\mathbb{P}}(S=s)$ for all $s \in \{0,1\}$

$$\begin{aligned}
\frac{\hat{\mathbb{E}}_{X|S=1}\hat{\theta}\hat{\eta}(X,1)\hat{g}(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} &= \hat{\mathbb{E}}_{X|S=1}[(2\hat{\eta}(X,1)-1)\hat{g}(X,1)]\hat{\mathbb{P}}(S=1) \\
&\quad - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ ,
\end{aligned} \tag{19}$$

and

$$\begin{aligned}
\frac{\hat{\mathbb{E}}_{X|S=0}\hat{\theta}\hat{\eta}(X,0)\hat{g}(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)} &= -\hat{\mathbb{E}}_{X|S=0}[(2\hat{\eta}(X,0)-1)\hat{g}(X,0)]\hat{\mathbb{P}}(S=0) \\
&\quad + \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .
\end{aligned} \tag{20}$$

Recall, that thanks to Definition 3.2 of the empirical unfairness we have

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

Since, $|\hat{\theta}| \leq 2$, subtracting Eq. (20) from Eq. (19) and taking absolute value combined with the

triangle inequality we get

$$
\begin{aligned}
&\hat{\mathbb{E}}_{(X,S)}(2\hat{\eta}(X,S)-1)\hat{g}(X,S) \\
&= \hat{\mathbb{E}}_{X|S=0}[(2\hat{\eta}(X,0)-1)\hat{g}(X,0)]\hat{\mathbb{P}}(S=0) + \hat{\mathbb{E}}_{X|S=1}[(2\hat{\eta}(x,1)-1)\hat{g}(X,1)]\hat{\mathbb{P}}(S=1) \qquad (21) \\
&\geq -2\hat{\Delta}(\hat{g},\mathbb{P}) + \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\
&\quad + \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ .
\end{aligned}
$$

Note that using the bound above we can get the following upper bound on the risk of the proposed classifier

$$
\begin{aligned}
\mathcal{R}(\hat{g}) &= \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\eta(X,S)-1)\hat{g}(X,S)] \\
&\leq \mathbb{E}_{(X,S)}[\eta(X,S)] - \mathbb{E}_{(X,S)}[(2\hat{\eta}(X,S)-1)\hat{g}(X,S)] \\
&\quad + 2\mathbb{E}_{(X,S)}\left|\eta(X,S)-\hat{\eta}(X,S)\right| \qquad \text{(replaced } \eta \text{ by } \hat{\eta}\text{)} \\
&\leq \mathbb{E}_{(X,S)}[\eta(X,S)] - \hat{\mathbb{E}}_{(X,S)}[(2\hat{\eta}(X,S)-1)\hat{g}(X,S)] + 2\mathbb{E}_{(X,S)}\left|\eta(X,S)-\hat{\eta}(X,S)\right| \\
&\quad + \left|(\mathbb{E}_{(X,S)} - \hat{\mathbb{E}}_{(X,S)})[(2\hat{\eta}(X,S)-1)\hat{g}(X,S)]\right| \qquad \text{(replaced } \mathbb{E}_{(X,S)} \text{ by } \hat{\mathbb{E}}_{(X,S)}\text{)} \\
&\leq \mathbb{E}_{(X,S)}[\eta(X,S)] - \hat{\mathbb{E}}_{(X,S)}[(2\hat{\eta}(X,S)-1)\hat{g}(X,S)] + 2\mathbb{E}_{(X,S)}\left|\eta(X,S)-\hat{\eta}(X,S)\right| \\
&\quad + \sup_{t\in[0,1]}\left|(\mathbb{E}_{(X,S)} - \hat{\mathbb{E}}_{(X,S)})[(2\hat{\eta}(X,S)-1)\mathbf{1}_{\{t\leq\hat{\eta}(X,S)\}}]\right| \qquad \text{(since } \hat{g} \text{ is thresholding)} \\
&\leq \mathbb{E}_{(X,S)}[\eta(X,S)] + 2\mathbb{E}_{(X,S)}\left|\eta(X,S)-\hat{\eta}(X,S)\right| \\
&\quad + 2\hat{\Delta}(\hat{g},\mathbb{P}) + \sup_{t\in[0,1]}\left|(\mathbb{E}_{(X,S)} - \hat{\mathbb{E}}_{(X,S)})[(2\hat{\eta}(X,S)-1)\mathbf{1}_{\{t\leq\hat{\eta}(X,S)\}}]\right| \\
&\quad - \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\
&\quad - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \qquad \text{(after Eq. (21))} .
\end{aligned}
$$

Thus, combining this upper bound on $\mathcal{R}(\hat{g})$ with the lower bound on $\mathcal{R}(\tilde{g})$ given in Eq. (18) we

arrive at

$$\begin{aligned}\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g}) \leq\ & 4\mathbb{E}_{(X,S)}\left|\eta(X,S) - \hat{\eta}(X,S)\right| + 2\hat{\Delta}(\hat{g},\mathbb{P}) \\ & + \sup_{t\in[0,1]}\left|(\mathbb{E}_{(X,S)} - \hat{\mathbb{E}}_{(X,S)})[(2\hat{\eta}(X,S)-1)\mathbf{1}_{\{t\leq\hat{\eta}(X,S)\}}]\right| \\ & + \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\ & - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\ & + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\ & - \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .\end{aligned}$$

Thanks to Lemma C.2 the term $\sup_{t\in[0,1]}\left|(\mathbb{E}_{(X,S)} - \hat{\mathbb{E}}_{(X,S)})[(2\hat{\eta}(X,S)-1)\mathbf{1}_{\{t\leq\hat{\eta}(X,S)\}}]\right|$ converges to zero in expectation[6]. Equation (10) with Lemma C.2 gives the convergence to zero of $\hat{\Delta}(\hat{g},\mathbb{P})$ in expectation. Assumption 4.1 tells us that the term $\mathbb{E}_{(X,S)}\left|\eta(X,S) - \hat{\eta}(X,S)\right|$ goes to zero in expectation. Thus it only remains to bound the term

$$\begin{aligned}(*) =& \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\tilde{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\ & - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\ & + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\tilde{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\ & - \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ .\end{aligned}$$

Notice that (similarly to the case of θ^*) the condition in Eq. (13) on $\tilde{\theta}$ is the first order optimality condition for the minimum of the following function

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

thus, the objective evaluated at minimum, that is, at $\tilde{\theta}$ is less or equal than the one evaluated at $\hat{\theta}$. Which implies that in order to upper bound $(*)$ it is sufficient to provide an upper bound on

[6]Actually Lemma C.2 is stated with $\hat{\eta}(X,S)$, whereas here it is $(2\hat{\eta}(X,S)-1)$. A straightforward modification of the argument used in Lemma C.2 yields the desired result.

$$
\begin{aligned}
(**) =& \mathbb{E}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& + \mathbb{E}_{X|S=0}\left(\mathbb{P}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\mathbb{E}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ \\
& - \hat{\mathbb{E}}_{X|S=0}\left(\hat{\mathbb{P}}(S=0)(2\hat{\eta}(X,0)-1) + \frac{\hat{\theta}\hat{\eta}(X,0)}{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}\right)_+ ,
\end{aligned}
$$

where we replaced $\tilde{\theta}$ by $\hat{\theta}$ thanks to the optimality of $\tilde{\theta}$. Let us define

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

Both bounds are following similar arguments, we demonstrate it for $(\triangle)$, clearly we have

$$
\begin{aligned}
(\triangle) \leq& \hat{\mathbb{E}}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& + \left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+\right| .
\end{aligned}
$$

For the first difference on the right hand side of this inequality we can write using the fact that $(x)_+ - (y)_+ \leq |x-y|$ for all $x, y \in \mathbb{R}$ and $|2\hat{\eta}(X,1)-1| \leq 1$ almost surely

$$
\begin{aligned}
& \hat{\mathbb{E}}_{X|S=1}\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& - \hat{\mathbb{E}}_{X|S=1}\left(\hat{\mathbb{P}}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+ \\
& \leq \left|\mathbb{P}(S=1) - \hat{\mathbb{P}}(S=1)\right| + |\hat{\theta}|\left|\frac{\hat{\mathbb{E}}_{X|S=1}[\hat{\eta}(X,1)]}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]} - 1\right|
\end{aligned}
$$

Clearly $\left|\mathbb{P}(S=1) - \hat{\mathbb{P}}(S=1)\right|$ goes to zero in expectation thanks to the law of large numbers or its finite sample variants. Besides, the term $\left|\frac{\hat{\mathbb{E}}_{X|S=0}[\hat{\eta}(X,0)]}{\mathbb{E}_{X|S=0}[\eta(X,0)]} - 1\right|$ can be seen in the following manner: let $Z \in [0,1]$ be a random variable with law $\mathbb{P}_Z$ and $Z_1, \dots, Z_M$ be its i.i.d. realization, then sequentially our question is about

$$\left|1 - \frac{\bar{Z}}{\mathbb{E}[Z]}\right| \ ,$$

with $\bar{Z} = \frac{1}{M}\sum_{i=1}^{M} Z_i$. This term converges to zero in expectation thanks to the multiplicative Chernoff inequality, which is an exponential concentration inequality that allows to obtain even a rate. Actually, even without the multiplicative Chernoff bound this term goes to zero thanks to the law of large numbers. Therefore, for convergence it remains to study the term

$$(\star) = \left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) - \frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right)_+\right| \ .$$

Notice that thanks to the second part of Assumption 4.1 and the fact that $\hat{\theta} \in [-2,2]$ we have

$$\left|\frac{\hat{\theta}\hat{\eta}(X,1)}{\mathbb{E}_{X|S=1}[\hat{\eta}(X,1)]}\right| \leq \frac{2}{c_{n,N}} \ .$$

Therefore, we can upper bound $(\star)$ as

$$(\star) \leq \sup_{t \in [-2/c_{n,N}, 2/c_{n,N}]} \left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\left(\mathbb{P}(S=1)(2\hat{\eta}(X,1)-1) + t\right)_+\right| \ ,$$

where the random quantity has been "supped-out". Introduce,

$$\begin{aligned} \mathcal{D}_{N_1} &= \{X_i \in \mathcal{D}_N \ : \ S_i = 1\} \\ \mathcal{D}_{N_0} &= \{X_i \in \mathcal{D}_N \ : \ S_i = 0\} \ , \end{aligned}$$

of size N_1 and N_0 respectively, such that $N_1 + N_0 = N$. Clearly we have $\mathcal{D}_{N_s} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{X|S=s}$ for each $s \in \{0,1\}$. Also recall that Remark B.1 implies that neither N_0 nor N_1 are equal to zero, however, both are still random. Besides, denote by $\mathcal{D}_N^S = \{S_i \ : \ (X_i, S_i) \in \mathcal{D}_N\}$ the dataset which is obtained from $\mathcal{D}_N$ by removing features. Thus,

$$\mathbb{E}_{(\mathcal{D}_N)}(\star) \leq \mathbb{E}_{\mathcal{D}_N^S}\mathbb{E}_{\mathcal{D}_{N_1}} \sup_{t \in [-2/c_{n,N}, 2/c_{n,N}]} \left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\left((2\hat{\eta}(X,1)-1)\mathbb{P}(S=1) + t\right)_+\right| \ .$$

Conditionally on $\mathcal{D}_N^S$ we can view N_0 and N_1 as fixed strictly positive integers, moreover, conditionally on $\mathcal{D}_n$ the estimator $\hat{\eta}$ is not random as it is built *only* on $\mathcal{D}_n$. Thus, we would like to control the following process

$$\mathbb{E}_{\mathcal{D}_{N_1}} \sup_{t \in [-2/c_{n,N}, 2/c_{n,N}]} \left|(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1})\left((2\hat{\eta}(X,1)-1)\mathbb{P}(S=1) + t\right)_+\right| \ ,$$

conditionally on $\mathcal{D}_N^S, \mathcal{D}_n$. First of all we rewrite this process as

$$\frac{1}{c_{n,N}}\mathbb{E}_{\mathcal{D}_{N_1}} \sup_{|t|\leq 1} \left| \left(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1}\right) \left((2\hat{\eta}(X,1)-1)\mathbb{P}(S=1)c_{n,N} + 2t\right)_+ \right| \quad .$$

Thanks to the symmetrization argument we can write

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{N_1}} \sup_{|t|\leq 1} \left| \left(\mathbb{E}_{X|S=1} - \hat{\mathbb{E}}_{X|S=1}\right) \left((2\hat{\eta}(X,1)-1)\mathbb{P}(S=1)c_{n,N} + 2t\right)_+ \right| \\
&\qquad \leq 2\mathbb{E}_{\mathcal{D}_{N_1}} \sup_{|t|\leq 1} \left| \frac{1}{N_1}\sum_{i=1}^{N_1} \varepsilon_i f_t(X_i) \right| \quad ,
\end{aligned}$$

where $f_t(\cdot) = \left((2\hat{\eta}(\cdot,1)-1)\mathbb{P}(S=1)c_{n,N} + 2t\right)_+$. Notice that for each t, t' we have for all $x \in \mathbb{R}^d$

$$|f_t(x) - f_{t'}(x)| \leq 2\left|t - t'\right| \quad ,$$

that is, the parametrization is 2-Lipschitz. Therefore, standard results in empirical processes (combine [44, Lemma 6.2] with [29, Theorem 3.2.]) tells us that there exists $C > 0$ such that

$$\mathbb{E}_{\mathcal{D}_{N_1}} \sup_{|t|\leq 1} \left| \frac{1}{N_1}\sum_{i=1}^{N_1} \varepsilon_i f_t(X_i) \right| \leq C\sqrt{\frac{1}{N_1}} \quad .$$

Thus, taking expectation *w.r.t.* $\mathcal{D}_N^s$ we get

$$\mathbb{E}_{(\mathcal{D}_N)}(\star) \leq \frac{C}{c_{n,N}}\mathbb{E}_{\mathcal{D}_N^s}\sqrt{\frac{1}{N_1}} \quad ,$$

applying Lemma B.2 we get for some $C > 0$ that depends on $\mathbb{P}(S=1)$ that

$$\mathbb{E}_{(\mathcal{D}_N)}(\star) \leq \frac{C}{c_{n,N}}\sqrt{\frac{1}{N}} \quad .$$

Thanks to Assumption 4.1 we have

$$\frac{1}{c_{n,N}\sqrt{N}} = o\left(1\right) \quad ,$$

thus, the term $\mathbb{E}_{(\mathcal{D}_N)}(\star)$ converges to zero. Repeating the same argument for $(\triangle\triangle)$ we conclude. □

D Proofs of auxiliary results

Proof of Lemma C.1. We start from the level of unfairness of g, that is, we would like to find an upper bound on

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

rewriting the expression above, our goal can be written as

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

Now, we start working with the expression above

$$\begin{aligned}
&\left|\frac{\mathbb{E}_{X|S=1}\eta(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\eta(X,1)} - \frac{\mathbb{E}_{X|S=0}\eta(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\eta(X,0)}\right| \\
&\leq \left|\frac{\mathbb{E}_{X|S=1}\eta(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\eta(X,1)} - \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)}\right| \\
&\quad + \left|\frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} - \frac{\mathbb{E}_{X|S=0}\eta(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\eta(X,0)}\right| \\
&\quad + \left|\frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)}\right| \quad .
\end{aligned}$$

The first two terms on the right hand side of the inequality can be upper-bounded in a similar way. That is why we only show the bound for the first term, that is, for $S = 1$. We have for $(*) = \left|\frac{\mathbb{E}_{X|S=1}\eta(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\eta(X,1)} - \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)}\right|$

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

thus, we have

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

Finally, it remains to upper bound

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

Recall that $\hat{\mathbb{E}}_{X|S=1}$ and $\hat{\mathbb{E}}_{X|S=0}$ stands for the expectations taken *w.r.t.* empirical measure induced by $\mathcal{D}_N$, and that $\mathcal{D}_N$ is independent from $\mathcal{D}_n$. Therefore, we can write

$$
\begin{aligned}
(**) \leq & \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} \right| \\
& + \left| \frac{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\mathbb{E}_{X|S=0}\hat{\eta}(X,0)} - \frac{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)} \right| \\
& + \left| \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)g(X,0)}{\hat{\mathbb{E}}_{X|S=0}\hat{\eta}(X,0)} \right| \ .
\end{aligned}
$$

Clearly, the last term on the right hand side of the previous inequality corresponds to our empirical criteria since everything can be easily evaluated using data. The first two terms on the right hand side of the inequality can be upper-bounded in a similar fashion, again, we only demonstrate the bound for $S = 1$. We can write

$$
\begin{aligned}
& \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} \right| \\
& \leq \left| \frac{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} \right| \\
& \quad + \left| \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\mathbb{E}_{X|S=1}\hat{\eta}(X,1)} - \frac{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)g(X,1)}{\hat{\mathbb{E}}_{X|S=1}\hat{\eta}(X,1)} \right| \ .
\end{aligned}
$$

Notice that for the first term on the right hand side of the inequality we have

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

whereas for the second term we can write

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

□

Proof of Lemma C.2. Let us first introduce two slices of $\mathcal{D}_N$ as

$$
\mathcal{D}_{N_1} = \{X_i \in \mathcal{D}_N \ : \ S_i = 1\} \ , \ \mathcal{D}_{N_0} = \{X_i \in \mathcal{D}_N \ : \ S_i = 0\}
$$

of size N_1 and N_0 respectively, such that $N_1 + N_0 = N$. Clearly we have $\mathcal{D}_{N_s} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{X|S=s}$ for each $s \in \{0, 1\}$. Besides, denote by $\mathcal{D}_N^S = \{S_i \ : \ (X_i, S_i) \in \mathcal{D}_N\}$ the which is obtained from $\mathcal{D}_N$ by removing features. Recalling Remark B.1, we have

$$
N_1 - 2 \sim \text{Bin}(N, \mathbb{P}(S = 1)), \quad N_0 - 2 \sim \text{Bin}(N, \mathbb{P}(S = 0)) \ .
$$

Clearly, since the proposed algorithm is a thresholding of $\hat{\eta}$ we have

$$\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\hat{g}(X,0)\right| \\ \leq \mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbf{1}_{\{t\leq\hat{\eta}(X,0)\}}\right| \ .$$

Further we work conditionally on $\mathcal{D}_n$. Using the classical symmetrization technique [29, Theorem 2.1.] we get

$$\mathbb{E}_{\mathcal{D}_N}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbf{1}_{\{t\leq\hat{\eta}(X,0)\}}\right| \\ = \mathbb{E}_{\mathcal{D}_N^S}\mathbb{E}_{\mathcal{D}_{N_0}}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbf{1}_{\{t\leq\hat{\eta}(X,0)\}}\right| \\ \leq 2\mathbb{E}_{\mathcal{D}_N^S}\mathbb{E}_{\mathcal{D}_{N_0}}\mathbb{E}_{\varepsilon}\sup_{t\in[0,1]}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i\hat{\eta}(X_i,0)\mathbf{1}_{\{t\leq\hat{\eta}(X_i,0)\}}\right| \ ,$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim}$ Rademacher variables. Note that the function class $x\mapsto\mathbf{1}_{\{t\leq\hat{\eta}(x,0)\}}$ has VC-dimension [43] equal to one. At this moment we will work with

$$\mathbb{E}_{\varepsilon}\sup_{t\in[0,1]}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i\hat{\eta}(X_i,0)\mathbf{1}_{\{t\leq\hat{\eta}(X_i,0)\}}\right| \ ,$$

conditionally on all the data. First of all let us introduce $\mathcal{F}=\left\{f\,:\,\exists t\in[0,1],\,f(x)=\mathbf{1}_{\{t\leq\hat{\eta}(x,0)\}}\right\}$ Thus, our process can be written as

$$\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i\varphi_i(f(X_i))\right| \ ,$$

where $\varphi_i(\cdot)=\eta(X_i,0)\times\cdot$. Clearly, we have $\varphi_i(0)=0$ and for every u,v

$$|\varphi_i(u)-\varphi_i(v)|\leq|u-v| \ .$$

That is, φ_i are contractions, and the contraction lemma [29, Theorem 2.2.] gives

$$\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i\varphi_i(f(X_i))\right|\leq\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i f(X_i)\right| \ .$$

Recall, that the class $\mathcal{F}$ is a VC-class with VC-dimension equal to one. Therefore, it is a known fact [18, 34] that there exists $C>0$ such that

$$\mathbb{E}_{\varepsilon}\sup_{f\in\mathcal{F}}\left|\frac{1}{N_0}\sum_{X_i\in\mathcal{D}_{N_0}}\varepsilon_i f(X_i)\right|\leq C\sqrt{\frac{1}{N_0}} \ ,$$

almost surely. The above implies that

$$\mathbb{E}_{\mathcal{D}_N}\sup_{t\in[0,1]}\left|(\mathbb{E}_{X|S=0}-\hat{\mathbb{E}}_{X|S=0})\hat{\eta}(X,0)\mathbf{1}_{\{t\leq\hat{\eta}(X,0)\}}\right|\leq C\mathbb{E}_{\mathcal{D}_N^S}\sqrt{\frac{1}{N_0}} \ .$$

It remains to provide an upper bound on $\mathbb{E}_{\mathcal{D}_N^S}\sqrt{\frac{1}{N_0}}$, to this end we recall that this expectation can be written as

$$\mathbb{E}\sqrt{\frac{1}{2+Z}} \ ,$$

where Z is the binomial random variable with parameters N and $\mathbb{P}(S=0)$. Thus, thanks to Lemma B.2 there exists a constant $C>0$ that depends on $\mathbb{P}(S=0)$ such that

$$\mathbb{E}\sqrt{\frac{1}{2+Z}} \leq C\sqrt{\frac{1}{N}} \ .$$

Similarly we get the bound for the case $S=1$. □

E Optimal classifier independent of sensitive feature

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

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

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

Proposition E.1 (Optimal rule)**.** *Under Assumption 2.2 an optimal classifier g^* can be obtained for all $x\in\mathbb{R}^d$ as*

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

where θ^* is such that the equality

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

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

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

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

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

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

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

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

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

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

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

E.1 Proofs

Proof of Proposition E.1. Let us study the following minimization problem

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

Using the weak duality we can write

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

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

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

Using arguments of Lemma B.3 we can write

$$\mathbb{P}(g(X) \neq Y) = \mathbb{P}(Y = 1) - \mathbb{E}_X[(2\eta(X) - 1)g(X)] \ ,$$

where $\eta(\cdot) := \mathbb{P}(Y = 1 | X = \cdot)$. Moreover, since

$$\mathbb{E}[YS] = \mathbb{E}_S[S\mathbb{E}[Y|S]] = \mathbb{E}_S[S\mathbb{E}_X[\mathbb{E}[Y|X,S]]] = \mathbb{E}_S[S\mathbb{E}_X[\eta(X,S)]] = \mathbb{P}(S=1)\mathbb{E}_X[\eta(X,1)] \ ,$$

we can write for the rest

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

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

$$\mathbb{P}(Y=1) - \mathbb{E}_X\left[g(X) \left(2\eta(X) - 1 + \lambda \left(\frac{\eta(X,0)}{\mathbb{E}_X[\eta(X,0)]} - \frac{\eta(X,1)}{\mathbb{E}_X[\eta(X,1)]} \right) \right) \right] \ .$$

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

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

Similarly to Proposition 2.3, for $\lambda = 0$ we recover the classical optimal predictor in the context of binary classification. Substituting this classifier into the objective of $(**)$ we arrive at

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

The mapping

$$\lambda \mapsto \mathbb{E}_X\left(2\eta(X)-1+\lambda\left(\frac{\eta(X,0)}{\mathbb{E}_X[\eta(X,0)]}-\frac{\eta(X,1)}{\mathbb{E}_X[\eta(X,1)]}\right)\right)_+ \ ,$$

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

$$0 \in \partial_\lambda \mathbb{E}_X\left(2\eta(X)-1+\lambda\left(\frac{\eta(X,0)}{\mathbb{E}_X[\eta(X,0)]}-\frac{\eta(X,1)}{\mathbb{E}_X[\eta(X,1)]}\right)\right)_+ \ .$$

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

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

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

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

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

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

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

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

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

E.2 Experiments without the sensitive feature

In this section we report the equivalent results to those in Table 1 and Figure 1 into Table 3 and Figure 2 when the sensitive feature is not in the functional form of the model. Note that the method of Hardt [22] is not able to deal with this setting then there are no results for this case.

From Table 3 and Figure 2 we can observe analogous results to those in Section 5. Nevertheless, note that, without the sensitive feature in the functional form of the models, the results are generally less accurate and more fair w.r.t. to the case that the sensitive feature in the functional form of the models. This results is similar to the one reported in [17].

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

Table 3: Results (average ± standard deviation, when a fixed test set is not provided) for all the datasets, concerning ACC and DEO. In this case the sensitive feature the sensitive feature is not in the functional form of the model.

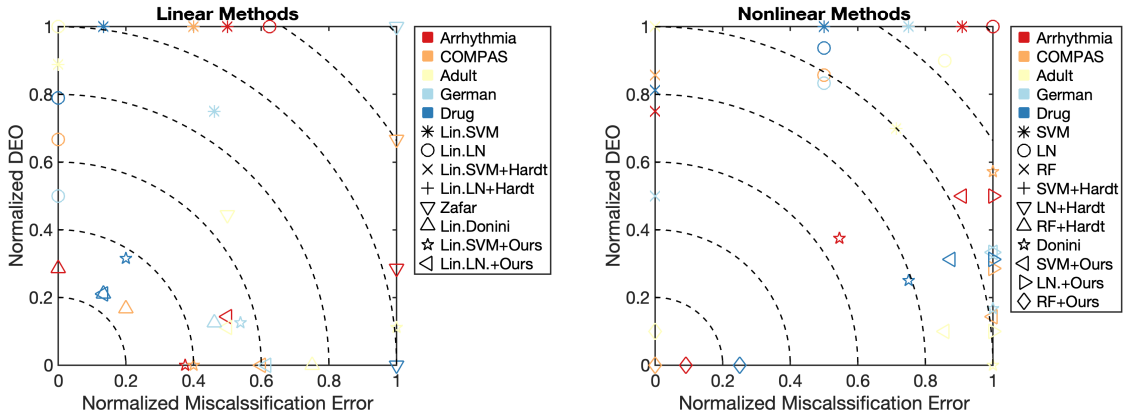

Figure 2: Results of Table 3 of linear (left) and nonlinear (right) methods when the error and the DEO are normalized in $[0, 1]$ column-wise. Different colors and symbols refer to different datasets and method respectively. The closer a point is to the origin, the better the result is. In this case the sensitive feature the sensitive feature is not in the functional form of the model.