# Fairness guarantee in multi-class classification

Christophe Denis, Romuald Elie, Mohamed Hebiri, François Hu

# Fairness guarantee in multi-class classification

Christophe Denis[(1)], Romuald Elie[(1)], Mohamed Hebiri[(1)], and François Hu[(2)]
[(1)]LAMA, UMR-CNRS 8050, Université Gustave Eiffel
[(2)] Département de Mathématiques et Statistique, Université de Montreal

**Abstract**

Algorithmic Fairness is an established area of machine learning, willing to reduce the influence of hidden bias in the data. Yet, despite its wide range of applications, very few works consider the multi-class classification setting from the fairness perspective. We focus on this question and extend the definition of approximate fairness in the case of *Demographic Parity* to multi-class classification. We specify the corresponding expressions of the optimal fair classifiers. This suggests a plug-in data-driven procedure, for which we establish theoretical guarantees. The enhanced estimator is proved to mimic the behavior of the optimal rule both in terms of fairness and risk. Notably, fairness guarantees are distribution-free. The approach is evaluated on both synthetic and real datasets and reveals very effective in decision making with a preset level of unfairness. In addition, our method is competitive (if not better) with the state-of-the-art in binary and multi-class tasks.

## 1 Introduction

Algorithmic fairness has become very popular during the last decade Zemel et al. [2013], Lum and Johndrow [2016], Calders et al. [2009], Zafar et al. [2017], Agarwal et al. [2019, 2018], Donini et al. [2018b], Chzhen et al. [2019], Chiappa et al. [2020], Barocas et al. [2018] as it addresses an important social concern: mitigating historical bias contained in the data. This is a crucial issue in many applications such as loan assessment or criminal sentencing among others. The main objective in algorithmic fairness consists in reducing the influence of a sensitive attribute on a prediction. Several notions of fairness have been considered in the literature for binary classification Zafar et al. [2019], Barocas et al. [2018]. All of them impose some independence condition between the sensitive feature and the prediction. This independence can be desired on some or all values of the label space, see *Equality of odds* or *Equal opportunity* [Hardt et al., 2016b]. In this paper, we focus on the well established *Demographic Parity* (DP) [Calders et al., 2009] that requires the independence between the sensitive feature and the prediction function, while not relying on labels. DP has a recognized interest in many applications, such as loan agreement without gender attributes or crime prediction without ethnicity discrimination [Hajian et al., 2011, Kamiran et al., 2013, Barocas and Selbst, 2014, Feldman et al., 2015]. Previously mentioned references consider either the regression or the binary classification frameworks, although most (modern) applications fall within the scope of multi-class classification (*e.g.* image recognition or text categorization). As an example, one might cite hiring tools based on Machine Learning (ML) models to give candidates one- to five-star ratings and favors men for software developers and other technical positions [Dastin, 2018].

The present work fills two gaps in the literature: i) it extends algorithmic fairness to the multi-class setting; ii) it properly studies the approximate fairness (also called as $\varepsilon$-fairness) from the theoretical point of view. Indeed, approximate fairness is known to be very efficient from a practical perspective Barocas et al. [2018], Zafar et al. [2019]. Nevertheless, main existing theoretical results only focus on exact fairness constraints, that is, they do not allow for deviating from a perfectly fair algorithm.

### 1.1 Main contributions

Overall, we emphasize that the present paper considers both the theoretical and the practical aspects of approximate fairness under the popular demographic parity constraint. Up to our knowledge, this is the first contribution that combines both aspects in the multi-class setting.

We establish a closed formula of the optimal predictor for both exact and approximate fairness constraints. Our proposed procedure is a post-processing algorithm which relies on solving a constrained

minimization problem. Specifically, in a first step we estimate the conditional probabilities of the output label given the sensitive attribute and the feature vectors, while a second step of the algorithm is dedicated to enforce fairness by shifting the estimated conditional probabilities in an optimal manner. We derive fairness and risk guarantees for our estimation procedure with explicit finite sample bounds. We also highlight the numerical performance of our algorithm and show that it performs as good as state-of-the-art multi-class methods for fair (or approximate fair) prediction. One of the main striking features of our procedure is that it can be applied to any off-the-shelf estimator of the conditional probabilities and it succeeds to enforce fairness at any pre-specified level.

We want to underline that the extensions, with respect to the existing literature, to multi-class and to approximate fairness are two theoretical aspects of the contribution. Both considerations involve additional technical arguments. In particular, dealing with approximate fairness is a new technical challenge. It is worth noticing that even in the binary classification setting, the control of the unfairness of the algorithms has not been analyzed theoretically.

Let us now summarize our main contributions:

- We provide an optimal solution for the multi-class problem under exact or approximate DP constraints. In particular, we derive a closed formula for the optimal (approximate) fair classifier.
- Based on this formula, we build a data-driven procedure that mimics the performance of the optimal rule both in terms of risk and fairness. Notably, our fairness guarantees are *distribution-free* and are established both in expectation and with high probability.
- We also established rates of convergence for the resulting classifier *w.r.t.* a suitable risk that combines both the error rate and the unfairness measure. A salient point of our theoretical findings is that our procedure achieves fast rates of convergence under a Margin type assumption.
- The approach is illustrated on several real and synthetic datasets with various bias levels. It provides robust and effective decision making rules with a preset level of unfairness.

## 1.2 Related works

There are mainly three ways to build fair prediction: i) *pre-processing* methods mitigate bias in the data before applying classical ML algorithms, see for instance Adebayo and Kagal [2016], Calmon et al. [2017], Zemel et al. [2013]; ii) *in-processing* methods reduce bias during training, see for instance [Agarwal et al., 2018, Donini et al., 2018a, Agarwal et al., 2019]; iii) *post-processing* methods enforce fairness after fitting, see for instance Hardt et al. [2016a], Chiappa et al. [2020], Chzhen et al. [2020b], Le Gouic et al. [2020]. The present work falls within the last category. In a related study, [Chzhen et al., 2019] exhibits fair binary classifiers under *Equal Opportunity* constraints. In contrast, we focus on the multi-class setting, while imposing *DP* constraints and we also treat the case of approximate fairness.

Up to our knowledge, only few works consider fairness in the multi-class setting. In Ye and Xie [2020], the authors enforce fairness by sub-sample selection and is in-processing. In contrast, we keep the whole sample and enforce fairness in a post-processing manner. Besides, from a high-level perspective, the procedure described in [Ye and Xie, 2020] imposes fairness on each component of the score function. It is clear that such methodology can be generalized to any convex empirical risk minimization (ERM) problem such as SVM or quadratic risk. But, since the decision rule in the multi-class setting relies on the maximizer over scores, we rather directly impose fairness on the maximizer itself.

The multi-class framework is also considered in [Zhang et al., 2018, Tavker et al., 2020, Alghamdi et al., 2022]. However, the authors in [Zhang et al., 2018, Tavker et al., 2020] do not provide an explicit formulation of the optimal fair rule and their theoretical fairness guarantee is not distribution free. In addition, they only consider numerical experiments for binary classification. Finally, the recent work in [Alghamdi et al., 2022] consider projecting an unfair classifier into a set of fair classifiers. However, as illustrated in Section 4.3, their method seems to fail in *exact* fairness. Our method provides valuable benefits on all these aspects.

## 1.3 Outline of the paper

In Section 2, we define the Demographic Parity constraint and the notion of exact/approximate fair classifier in the multi-class classification setup. An explicit expression of the optimal fair classifier is also

provided in Section 2. The corresponding data-driven procedure together with its statistical guarantees on risk and fairness are presented in Section 3. The numerical performance of the procedure is illustrated on both synthetic and real datasets in Section 4. The paper concludes with a discussion and perspective Section 5. For ease of readability, proofs and technical arguments are postponed to the Appendix of the paper.

# 2 Multi-class classification with demographic parity

Let $(X, S, Y)$ be a random tuple with distribution $\mathbb{P}$, where $X \in \mathcal{X} \subset \mathbb{R}^d$, $S \in \mathcal{S} := \{-1, 1\}$ and $Y \in [K] := \{1, \dots, K\}$ with $K$ a fixed number of classes. The distribution of the sensitive feature $S$ is denoted by $(\pi_s)_{s\in\mathcal{S}}$, and we assume that $\min_{s\in\mathcal{S}} \pi_s > 0$, meaning that we have access to both sensitive groups with non zero probability. A classification rule $g$ is a function mapping $\mathcal{X} \times \{-1, 1\}$ onto $[K]$, and its performance is evaluated through the misclassification risk

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

For $k \in [K]$, we denote $p_k(X, S) := \mathbb{P}\left(Y = k|X, S\right)$ the conditional probabilities. Recall that a Bayes classifier minimizing the misclassification risk $\mathcal{R}(\cdot)$ over the set $\mathcal{G}$ of all classifiers and is then given by

$$g^*(x, s) \in \arg\max_k p_k(x, s) \ , \quad \text{for all } (x, s) \in \mathcal{X} \times \mathcal{S} \ .$$

We introduce in Section 2.1 the Demographic parity constraint as well as the definition of an approximate fair classifier. The characterization of the optimal fair classifier and its main properties are provided in Section 2.2.

## 2.1 Demographic parity

We consider DP constraint [Calders et al., 2009] that asks for independence of the prediction function from the sensitive feature $S$. This definition naturally extends the DP constraint considered in binary classification [Agarwal et al., 2019, Chiappa et al., 2020, Gordaliza et al., 2019, Jiang et al., 2019, Oneto et al., 2019].

Approximate fairness, also referred to as $\varepsilon$-fairness, is highly popular from a practical perspective, in particular when a strict fairness constraint strongly deflates the accuracy of the method. In this context, the user is allowed to adjust the fairness constraint if relevant or needed. Of course, such modularity has a cost: the solution is less fair than the exact fair one. Moreover, the chosen unfairness level has no convincing interpretation. Without clear justification, some empirical rules exist such as the forth-firth that tolerates an unfairness of 0.2 [Holzer and Holzer, 2000, Collins, 2007, Feldman et al., 2015]. In this section, we consider *approximate* fairness setting without discussing the issue of properly selecting of the unfairness level $\varepsilon$.

We define the notion of $\varepsilon$-fairness in the particular case of Demographic Parity.

**Definition 2.1** ($\varepsilon$-fairness *w.r.t.* DP)**.** *The unfairness of a classifier $g \in \mathcal{G}$ is quantified by*

$$\mathcal{U}(g) := \max_{k\in[K]} \left|\mathbb{P}\left(g(X, S) = k|S = 1\right) - \mathbb{P}\left(g(X, S) = k|S = -1\right)\right| \ .$$

*A classifier $g$ is $\varepsilon$-fair if and only if $\mathcal{U}(g) \leq \varepsilon$. In particular, $\varepsilon = 0$ means that $g$ is exactly fair.*

Alternative measures of unfairness could be considered. The maximum can for instance be replaced by a summation over $k \in [K]$. While both measures have their advantages, picking the maximum simplifies fairness evaluation in empirical studies.

## 2.2 Optimal fair classifier

Our goal is to derive an explicit formulation of the optimal $\varepsilon$-fair classifiers *w.r.t.* the misclassification risk, denoted by $g^*_{\varepsilon-\text{fair}}$, which solves $\min_{g\in\mathcal{G}_{\varepsilon-\text{fair}}} \mathcal{R}(g)$ where $\mathcal{G}_{\varepsilon-\text{fair}}$ is the set of all $\varepsilon$-fair prediction functions. Its computation requires to properly balance misclassification risk together with fairness

criterion. The first step is to write the Lagrangian of the above problem: for $\lambda^{(1)} = (\lambda_1^{(1)}, \dots, \lambda_K^{(1)}) \in \mathbb{R}_+^K$ and $\lambda^{(2)} = (\lambda_1^{(2)}, \dots, \lambda_K^{(2)}) \in \mathbb{R}_+^K$, we define the *$\varepsilon$-fair-risk* as

$$\begin{aligned}\mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}(g) := \mathcal{R}(g) &+ \sum_{k=1}^{K} \lambda_k^{(1)} [\mathbb{P}\left(g(X,S) = k | S = 1\right) - \mathbb{P}\left(g(X,S) = k | S = -1\right) - \varepsilon] \\ &+ \sum_{k=1}^{K} \lambda_k^{(2)} [\mathbb{P}\left(g(X,S) = k | S = -1\right) - \mathbb{P}\left(g(X,S) = k | S = 1\right) - \varepsilon] \ .\end{aligned}$$

In order to characterize the optimal fair classifier, we also require the following technical condition.

**Assumption 2.2** (Continuity assumption)**.** *The mapping $t \mapsto \mathbb{P}\left(p_k(X,S) - p_j(X,S) \leq t | S = s\right)$ is assumed continuous, for any $k, j \in [K]$ and $s \in \mathcal{S}$.*

Assumption 2.2 implies that the distribution of the differences $p_k(X,S) - p_j(X,S)$ has no atoms. It is required to derive a closed expression for $g^*_{\varepsilon-\text{fair}}$ and insures an accurate calibration of the fairness at the prescribed level. Notice that in the binary case ($K = 2$), it boils down to the continuity of $t \mapsto \mathbb{P}\left(p_k(X,S) \leq t | S = s\right)$ considered in [Chzhen et al., 2019]. However when $K \geq 3$, these two conditions differ and we stress that Assumption 2.2 is a well tailored condition for the multi-class problem. We are now in position to provide a characterization of optimal $\varepsilon$-fair classifier.

**Theorem 2.3.** *Let $H : \mathbb{R}_+^{2K} \to \mathbb{R}$ be the function*

$$H(\lambda^{(1)}, \lambda^{(2)}) = \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[ \max_k \left( \pi_s p_k(X,s) - s(\lambda_k^{(1)} - \lambda_k^{(2)}) \right) \right] + \varepsilon \sum_{k=1}^{K} (\lambda_k^{(1)} + \lambda_k^{(2)}) \ .$$

*Let Assumption 2.2 be satisfied and define $(\lambda^{*(1)}, \lambda^{*(2)}) \in \arg\min_{(\lambda^{(1)},\lambda^{(2)}) \in \mathbb{R}_+^{2K}} H(\lambda^{(1)}, \lambda^{(2)})$. Then, $g^*_{\varepsilon-\text{fair}} \in \arg\min_{g \in \mathcal{G}_{\varepsilon-\text{fair}}} \mathcal{R}(g)$ if and only if $g^*_{\varepsilon-\text{fair}} \in \arg\min_{g \in \mathcal{G}} \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g)$.*
*In addition, for all $(x,s) \in \mathcal{X} \times \mathcal{S}$, we can rewrite the optimal classifier as*

$$g^*_{\varepsilon-\text{fair}}(x,s) = \arg\max_{k \in [K]} \left( \pi_s p_k(x,s) - s(\lambda_k^{*(1)} - \lambda_k^{*(2)}) \right).$$

Theorem 2.3 entails a closed form expression of optimal fair classifiers, which is the bedrock of our procedure: any optimal fair classifier is simply maximizing scores, that are obtained by shifting the original conditional probabilities in a proper manner. The above result also points out that the optimum of the risk $\mathcal{R}$ over the class of fair classifiers also minimizes the fair-risk $\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}$. Hence, by construction, $\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}$ is a risk measure that efficiently balances both classification accuracy and unfairness. An important consequence of the proof of Theorem 2.3 is the following proposition that more precisely characterizes the Lagrange multipliers $(\lambda^{*(1)}, \lambda^{*(2)})$, and the level of unfairness of the $\varepsilon$-fair predictor.

**Proposition 2.4.** *Let $\varepsilon \geq 0$. For each $k \in [K]$, we have that $\lambda_k^{*(1)} \lambda_k^{*(2)} = 0$ and $\lambda_k^{*(1)} + \lambda_k^{*(2)} \geq 0$. Besides, if for some $k$*

*i) $\lambda_k^{*(1)} > 0$, then $\mathbb{P}_{X|S=1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) - \mathbb{P}_{X|S=-1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) = \varepsilon$,*

*ii) $\lambda_k^{*(2)} > 0$, then $\mathbb{P}_{X|S=1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) - \mathbb{P}_{X|S=-1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) = -\varepsilon$.*

From the above result, we easily deduce the following corollary.

**Corollary 2.5.** *Let $\varepsilon \geq 0$. It holds that*

*i) either the Bayes classifiers satisfies $\mathcal{U}(g^*) \leq \varepsilon$ and then $g^* = g^*_{\varepsilon-\text{fair}}$. In this case $\lambda^{*(1)} = \lambda^{*(2)} = 0$;*

*ii) or the $\varepsilon$-fair classifier satisfies $\mathcal{U}(g^*_{\varepsilon-\text{fair}}) = \varepsilon$.*

A straightforward consequence of the above Proposition 2.4 and Corollary 2.5 is that

$$0 \leq \mathcal{R}(g^*_{\varepsilon-\text{fair}}) = \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g^*_{\varepsilon-\text{fair}}) \leq \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g) \leq R(g) + C\left(\mathcal{U}(g) - \varepsilon\right),$$

for all $g \in \mathcal{G}$ and for some constant $C > 0$ that depends on $K$. In the case of exact fairness (*e.g.* $\varepsilon = 0$) the following remark gives a specific characterization of the exact fair classifier.

**Remark 2.6** (Exact fairness)**.** *All previous results simplify in the exact fairness case setting where $\varepsilon = 0$. Considering the reparametrization $\beta_k^* := \lambda_k^{*(1)} - \lambda_k^{*(2)} \in \mathbb{R}$, we deduce the optimal fair classifier in this case*

$$g_{\text{fair}}^*(x,s) \in \arg\max_k \left(\pi_s p_k(x,s) - s\beta_k^*\right), \ \ (x,s) \in \mathcal{X} \times \mathcal{S},$$

*where*

$$\beta^* \in \arg\min_{\beta \in \mathbb{R}^K} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_k \left(\pi_s p_k(X,s) - s\beta_k\right)\right] \ .$$

*In view of Corollary 2.5, we have $\mathcal{U}(g_{\text{fair}}^*) = 0$.*

**Binary classification** Finally, we conclude this section with a particular focus on the binary classification setting where specific characterization of the optimal fair predictor can be obtained.

**Corollary 2.7.** *Let $\varepsilon \geq 0$. In the binary setting ($K = 2$ with label space $\mathcal{Y} = \{0,1\}$), the fairness constraint reduces to a single condition and the optimal fair classifier simplifies as*

$$g_{\text{fair}}^*(x,s) = \mathbb{1}_{\{p_1(x,s) \geq \frac{1}{2} + \frac{s\beta^*}{2\pi_s}\}}, \quad (x,s) \in \mathcal{X} \times \mathcal{S} \ ,$$

*where, with the notation $F_s(t) = \mathbb{P}\left(p_1(X,S) \leq t \mid S = s\right)$ we have*

*i) $\beta^* = 0$ if $\left|F_1\left(\frac{1}{2}\right) - F_{-1}\left(\frac{1}{2}\right)\right| \leq \varepsilon$;*

*ii) $\beta^*$ is solution in $\beta$ of $\left|F_1\left(\frac{\beta+\pi_1}{2\pi_1}\right) = F_{-1}\left(\frac{-\beta+\pi_{-1}}{2\pi_{-1}}\right)\right| = \varepsilon$ otherwise.*

The proof of this result follows directly from Theorem 2.3 by considering classifiers $g$ that satisfy the fairness constraint $|\mathbb{P}\left(g(X,S) = 1 \mid S = -1\right) - \mathbb{P}\left(g(X,S) = 1 \mid S = 1\right)| \leq \varepsilon$. This constraint ensures that the condition is also satisfied for $g(X,S) = 0$ since $g$ is a binary function.

The above result highlights several important facts about the characterization of the optimal fair classifier in the binary setting. First, the optimal rule is deduced just by thresholding the conditional probability $p_1$. The thresholding is not at the classical level $1/2$ (*e.g.* without fairness constraint) but at a shifting of this value by $\frac{s\beta^*}{2\pi_s}$ to enforce fairness. Second, observe that the rule only depends on $p_1$ (and not $p_0$) for the same reason as in classical binary classification, that is $p_0 = 1 - p_1$. This yields to a reduction of the number of Lagrange parameters into a single one $\beta^*$. Notice that the case $\beta^* = 0$ means that the Bayes rule is already fair and then coincides with the $\varepsilon$-fair optimal predictor. In contrast, if $\beta^* \neq 0$, the optimal $\varepsilon$-fair rule differs from the Bayes rule and the modification of the rule is deduced by shifting the conditional probability.

# 3 Data-driven procedure

This section is devoted to the definition and the theoretical study of our empirical procedure that relies on the *plug-in* principle. The construction of our estimator is formally presented in Section 3.1 while its statistical properties are provided in Section 3.2.

## 3.1 Plug-in estimator

The enhanced estimation procedure is in two steps. According to the definition of the optimal $\varepsilon$-fair predictor given in Theorem 2.3, we first build estimators of the conditional probabilities $(p_k)_k$ and then proceed with the estimation of the parameters $\lambda^*$ and $(\pi_s)_{s \in \mathcal{S}}$. Notably, our data-driven procedure is semi-supervised as it relies on two independent datasets, one labeled and another unlabeled.

The first *labeled* dataset $\mathcal{D}_n = (X_i, S_i, Y_i)_{i=1,\dots,n}$ contains *i.i.d.* samples from the distribution $\mathbb{P}$. It allows to train estimators $(\hat{p}_k)_k$ of the conditional probabilities $(p_k)_k$ by the means of any machine learning supervised algorithm, *e.g.,* Random Forest, SVM. At this level, it is important to stress a key feature of the algorithm. Ones the empirical conditional probabilities $\hat{p}_k$ are trained, the theoretical analysis of the risk and the unfairness of the plug-in rule requires continuity conditions on the random variables $\hat{p}_k(X,S)$ (conditional on the learning sample, see Assumption 2.2). Notably, this is automatically satisfied whenever perturbing $(\hat{p}_k)_k$ with a continuous random noise (with a small magnitude to avoid deflating

the statistical properties of the estimate). We insure such a property simply by randomization. Indeed, let $u$ be a non negative real number. For each $k \in [K]$, we introduce

$$\bar{p}_k(X, S, \zeta_k) := \hat{p}_k(X, S) + \zeta_k,$$

with $(\zeta_k)_{k\in[K]}$ being *i.i.d.* according to a uniform distribution on $[0, u]$. This perturbation improves the fairness calibration in both theory and practice due to the fact that atoms for the random variables $\hat{p}_k(X, S) - \hat{p}_j(X, S)$ are avoided in this case.

The second *unlabeled* dataset $\mathcal{D}'_N$ contains $N$ *i.i.d.* copies of $(X, S)$. It is used to calibrate fairness. For $s \in \mathcal{S}$, the number of observations corresponding to $S = s$ is denoted by $N_s$, so that $N_{-1} + N_1 = N$. On the one hand, the feature vectors in $\mathcal{D}'_N$ are denoted by $X^s_1, \ldots, X^s_{N_s}$ and are *i.i.d.* data from the distribution $\mathbb{P}_{X^s}$ of $X|S = s$. On the other hand, the sensitive features from $\mathcal{D}'_N$ are denotes by $(S_1, \ldots, S_N)$. The latter are *i.i.d.* and are used to compute empirical frequencies $(\hat{\pi}_s)_{s\in\mathcal{S}}$ as estimates of $(\pi_s)_{s\in\mathcal{S}}$ (recall that $\pi_s = \mathbb{P}(S = s)$). Now notice that the estimation of parameters $(\lambda^{*(1)}, \lambda^{*(2)})$ only involves marginal distributions of $\mathbb{P}_{X|S=s}$ and $\mathbb{P}_S$. Therefore, this estimation part relies on the estimators $\hat{\pi}_s$, on the feature vectors $(X^s_1, \ldots, X^s_{N_s})$, and on independent copies $(\zeta^s_{k,i})_{k\in[K],i\in[Ns]}$ of a Uniform distribution on $[0, u]$ (for $s \in \mathcal{S}$). In particular, we define $(\hat{\lambda}^{(1)}, \hat{\lambda}^{(2)})$ as a minimizer over $\mathbb{R}^{2K}_+$ of $\hat{H}(\lambda^{(1)}, \lambda^{(2)})$ that is defined by (see the population counterpart given in Theorem 2.3)

$$\hat{H}(\lambda^{(1)}, \lambda^{(2)}) \;:=\; \sum_{s\in\mathcal{S}} \frac{1}{N_s} \sum_{i=1}^{N_s} \left[ \max_k \left( \hat{\pi}_s \bar{p}_k(X^s_i, s, \zeta^s_{k,i}) - s(\lambda^{(1)}_k - \lambda^{(2)}_k) \right) \right] \;+\; \varepsilon \sum_{k=1}^{K} (\lambda^{(1)}_k \;+\; \lambda^{(2)}_k) \;. \quad (1)$$

Finally, our *randomized* fair algorithm $\hat{g}$ is defined as

$$\hat{g}_\varepsilon(x, s) = \arg\max_{k\in[K]} \left( \hat{\pi}_s \bar{p}_k(x, s, \zeta_k) - s(\hat{\lambda}^{(1)}_k - \hat{\lambda}^{(2)}_k) \right) \;, \qquad (x, s) \in \mathcal{X} \times \mathcal{S} \;, \quad (2)$$

Note that the construction of the plug-in rule $\hat{g}$ relies on $(x, s)$ but also on the perturbations $\zeta_k$ and $\zeta^s_{k,i}$ for $k \in [K]$, $i \in [N_s]$, and $s \in \mathcal{S}$, that are easily collected as *i.i.d.* uniform random variables.

**Remark 3.1.** *Classical datasets often contain only labeled samples. Then, our approach requires to split the data into two independent samples $\mathcal{D}_n$ and $\mathcal{D}'_N$, by removing labels in the latter. As illustrated in Section 4.2, this splitting step is important to get the right level of fairness.*

## 3.2 Statistical guarantees

We are now in position to derive fairness and risk guarantees of our plug-in procedure. We need the following additional notation: $\pi_{\min} := \min_{s\in\mathcal{S}} \pi_s$ and $N_{\min} = \min(N_1; N_{-1})$.

### 3.2.1 Universal fairness guarantee

We first focus on fairness assessment and prove that the plug-in estimator $\hat{g}$ is asymptotically $\varepsilon$-fair, that is, it satisfies the requirement of Definition 2.1. This control on the fairness will be established both in expectation and with high probability. In addition, we prove that the convergence rate of the unfairness to zero is parametric with the number of unlabeled data $N$. Notably, the fairness guarantee is distribution-free and holds for any estimators of the conditional probabilities.

**Theorem 3.2.** *Let $\varepsilon \geq 0$. There exists a constant $C > 0$ depending only on $K$ and $\pi_{\min}$ such that, for any estimators $\hat{p}_k$ of the conditional probabilities, we have*

$$\mathbb{E}\left[\mathcal{U}(\hat{g}_\varepsilon)\right] \leq \varepsilon + \frac{C}{\sqrt{N}} \;.$$

This first finite-sample bound on the fairness illustrates a key feature of our post-processing approach. It makes (asymptotically) $\varepsilon$-fair any off-the-shelf (unconstrained/unfair) estimators of the conditional probabilities. This post-processing step is especially appealing when the cost of re-training an existing learning algorithm is high. While the former result provides a control of the unfairness on our algorithm in expectation, it is also appealing to have a thinner analysis of the unfairness through a high probability control.

**Theorem 3.3.** *Let $0 < \delta < 1$ and define $C_\delta = 4K\sqrt{2\log(\frac{4K}{\delta})}$. Assume that $\varepsilon > \frac{\sqrt{2}C_\delta}{\sqrt{\pi_{\min}N}}$ and that $N \geq 2\frac{\log(1/\delta)}{\pi_{\min}^2}$. Then there exists an event $\mathcal{A}(\delta)$ that holds with probability $1-(K+2)\delta$ on which we have*

$$\frac{C_\delta}{\sqrt{N_{\min}}} < \varepsilon, \quad \text{and } \forall k \in [K],\ \hat{\lambda}_k^{(1)}\hat{\lambda}_k^{(2)} = 0.$$

*Besides on $\mathcal{A}(\delta)$, the following holds*

*1) either $|\mathcal{U}(\hat{g}_\varepsilon) - \varepsilon| \leq \frac{C_\delta}{\sqrt{N_{\min}}}$;*

*2) or $\mathcal{U}(\hat{g}_\varepsilon) < \varepsilon - \frac{C_\delta}{\sqrt{N_{\min}}}$, and then we have $\hat{g} = \hat{g}_\varepsilon$ (for each $k \in [K]$, $\hat{\lambda}_k^{(1)} = \hat{\lambda}_k^{(2)} = 0$).*

This result has several levels of understanding. It highlights that the bound on the unfairness established in Theorem 3.2 is also valid with high probability, that is, there exists some constant $C > 0$ such that $\mathcal{U}(\hat{g}_\varepsilon) \leq \varepsilon + \frac{C}{\sqrt{N_{\min}}}$ with high probability. However, this result covers two significantly different situations for $\hat{g}_\varepsilon$: the first case is when the unfairness of $\hat{g}_\varepsilon$ is small *w.r.t.* to $\varepsilon$. This means that the unconstrained classifier $\hat{g}$ is already $\varepsilon$-fair and the action of the fairness constraint on our prediction function is null. In this case, we have $\hat{g}_\varepsilon = \hat{g}$. The second case, which is also the most expected one, is when at least one coordinate of the Lagrangian is non zero (*e.g.* either $\hat{\lambda}_k^{(1)}$ or $\hat{\lambda}_k^{(2)}$ is non zero for some $k$). Here, imposing the fairness constraint is relevant and the unfairness of $\hat{g}_\varepsilon$ falls within a small interval around $\varepsilon$.

From another perspective, all these conclusions are valid under some conditions on the desired level of unfairness $\varepsilon$ and the sample size $N$. It is assumed that $N$ is large enough to make the fairness constraint meaningful. However, it could be interesting to consider the case where $\varepsilon$ is smaller than the rate $\frac{1}{\sqrt{\pi_{\min}N}}$. (Observe that $\pi_{\min}N$ is the expectation of $N_{\min}$.) In this case, our statements shows that all values of $\varepsilon \in [0, \frac{1}{\sqrt{\pi_{\min}N}}]$ lead, from the theoretical perspective, to the same bound on the unfairness of the resulting classifier.

### 3.2.2 Consistency result

In this part, we provide a control on the misclassification risk of $\hat{g}_\varepsilon$. Let us define the $\ell_1$-norm in $\mathbb{R}^K$ between the estimator $\hat{\mathbf{p}} := (\hat{p}_1, \ldots, \hat{p}_K)$ and the vector of the conditional probabilities $\mathbf{p} := (p_1, \ldots, p_K)$ by $\|\hat{\mathbf{p}}(X,S) - \mathbf{p}(X,S)\|_1 = \sum_{k\in[K]} |\hat{p}_k(X,S) - p_k(X,S)|$. We then derive the following bound.

**Theorem 3.4.** *Let Assumption 2.2 be satisfied. Assume that $\dfrac{N}{\log(N)} \geq 2\pi_{\min}^{-2}$, then it holds that*

$$\mathbb{E}[\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(\hat{g}_\varepsilon)] - \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g^*_{\varepsilon-\text{fair}}) \leq C\left(\mathbb{E}\left[\|\hat{\mathbf{p}}(X,S) - \mathbf{p}(X,S)\|_1\right] + \sum_{s\in\mathcal{S}} \mathbb{E}\left[|\hat{\pi}_s - \pi_s|\right] + \frac{\log(N)}{\sqrt{N}} + u\right),$$

*where $C > 0$ depends on $K$ and $\pi_{\min}$.*

This result highlights that the excess fair-risk of $\hat{g}$ depends on i) the $L_1$-risk of $\hat{\mathbf{p}}$ for estimating the conditional probabilities; ii) the efficiency of the estimators $(\hat{\pi}_s)_{s\in\mathcal{S}}$; iii) a bound on the unfairness of the classifier; and iv) the upper-bound $u$ on the regularizing perturbations. In view of Theorem 3.3, $\hat{g}_\varepsilon$ is consistent *w.r.t.* the misclassification risk as soon as the estimator $\hat{\mathbf{p}}$ is consistent in $L_1$-norm. In particular, we can establish the following result.

**Corollary 3.5.** *Let $\varepsilon \geq 0$, if $\mathbb{E}\left[\|\hat{\mathbf{p}}(X,S) - \mathbf{p}(X,S)\|_1\right] \to 0$ and $u = u_n \to 0$ when $n \to \infty$, we have*

$$\left|\mathbb{E}\left[\mathcal{R}(\hat{g}_\varepsilon)\right] - \mathcal{R}(g^*_{\varepsilon-\text{fair}})\right| \to 0, \qquad \text{as} \quad n, N \to \infty\ .$$

Theorem 3.2 and Corrolary 3.5 directly imply that $\hat{g}_\varepsilon$ performs asymptotically as well as $g^*_{\varepsilon-\text{fair}}$ both in terms of fairness and accuracy provided that the estimators of $p_k$ are consistent *w.r.t.* the $L_1$ risk.

### 3.2.3 Rates of convergence

This section is dedicated to the study of rates of convergence *w.r.t* the excess fair-risk. To this end, we require additional assumptions on the regression functions $p_k$.

**Assumption 3.6.** *(Smoothness assumption) For all $k \in [K]$, the regression function $p_k$ is Lipschitz.*

The bound on the excess-risk provided in Theorem 3.4 depends on $\mathbb{E}\left[\|\hat{\mathbf{p}}(X,S) - \mathbf{p}(X,S)\|_1\right]$. Imposing additional regularity constraint on $\mathbf{p}$, this term can further be controlled. For instance, if we assume that for each $k \in [K]$, the regression functions $p_k$ are Lipschitz then well established nonparametric estimators of $p_k$, such as local polynomials or kernel based methods, lead to

$$\mathbb{E}\left[\|\hat{\mathbf{p}}(X,S) - \mathbf{p}(X,S)\|_1\right] \leq Cn^{-1/(2+d)} \ .$$

In this case a straightforward consequence of Theorem 3.4 is that for $u \leq n^{-1/(2+d)}$

$$\mathbb{E}[\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(\hat{g}_\varepsilon)] - \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g^*_{\varepsilon-\text{fair}}) \leq C\left(n^{-1/(2+d)} \bigvee N^{-1/2}\right) \ .$$

In particular, if $N$ is sufficiently large, that is $N^{-1/2} = \mathcal{O}\left(n^{-1/(2+d)}\right)$, the obtained rates is of the same order as the minimax rates in classification setting without fairness constraint Audibert and Tsybakov [2007]. Interestingly, it is possible to obtain faster rates under a stronger assumption than Assumption 2.2.

**Assumption 3.7.** *(Density assumption) For any $k, j \in [K]$ and $s \in \mathcal{S}$, we assume that conditional on $S = s$, the random variable $p_k(X,S) - p_j(X,S)$ admits a bounded density.*

Note that under Assumption 3.7, the Tsybakov's margin condition is satisfied with parameter $\alpha = 1$. Taking advantage of the margin condition, we can establish the following result.

**Theorem 3.8.** *For $\varepsilon > 0$ and for a sample size $N$ such that $\dfrac{N}{\log(N)} \geq 2\pi_{\min}^{-2}$, the following holds*

$$\mathbb{E}[\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(\hat{g}_\varepsilon)] - \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g^*_{\varepsilon-\text{fair}}) \leq C\left(\mathbb{E}\left[\|\hat{\mathbf{p}}(X,S) - \mathbf{p}(X,S)\|_\infty^2\right] + \frac{\log^2(N)}{N} + u^2\right) \ ,$$

*where $C > 0$ depends on $K$ and $\pi_{\min}$.*

The major consequence of the above result is that fast rates of convergence (faster than $n^{-1/2}$) can be obtained for the excess fair-risk. Specifically, if under Assumption 3.6, the estimator satisfies

$$\mathbb{E}\left[\|\hat{\mathbf{p}}(X,S) - \mathbf{p}(X,S)\|_\infty\right] \leq C\log(n)n^{-1/(2+d)} \ , \tag{3}$$

(which is again the case for popular methods) under Assumption 3.7, it holds that

$$\mathbb{E}[\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(\hat{g}_\varepsilon)] - \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g^*_{\varepsilon-\text{fair}}) \leq C\log^2(n)\left(n^{-2/(2+d)} \bigvee N^{-1}\right) \ .$$

Interestingly, if the size of the unlabeled sample $N$ is sufficiently large ($N \geq \log(n)^{-2}n^{2/(2+d)}$), then up to a logarithmic factor the established rates of convergence is of the same order as the minimax *fast* rates of convergence for plug-in classifiers (see Audibert and Tsybakov [2007]) in supervised classification without fairness constraint. Hence, we manage to show that fast rates can be also achieved in the algorithmic fairness framework under Margin type assumption. Note that the condition required in Equation (3) is, for instance fulfilled by local polynomial estimator or $k$NN classifiers under Assumption 3.6. Finally, we also want to point out that we restrict our analysis to the case where the regression functions $p_k$ are Lipschitz to ease the presentation. However, we can extend our results to the case where the regression functions are in a Hölder class.

# 4 Numerical Evaluation

We now evaluate our method numerically[1]. Section 4.2 illustrates the efficiency of the $\varepsilon$-fairness algorithm on synthetic data, while experiments on real datasets are provided in Section 4.3. Up to our knowledge, imposing the fairness constraint in multi-class classification in a model-agnostic post-processing approach is only addressed in [Alghamdi et al., 2022]. Therefore we will mainly compare our method to [Alghamdi et al., 2022] for multi-class tasks and to the state-of-the-art in-processing approach [Agarwal et al., 2019] that is designed for binary tasks.

[1]The source of our method can be found at `https://github.com/curiousML/epsilon-fairness`.

## 4.1 Implementation of the algorithm

Let us focus on the implementation of the algorithm producing an $\varepsilon$-fairness classifier. Although the exact fairness setting allows for improvements using accelerated gradient descent, we do not focus on this point and simply identify the exact fair algorithm to the approximate fair one with $\varepsilon = 0$.

The proposed approximate fair algorithm is defined in Eq. (2) and requires to solve an optimization problem in Eq. (1). The implementation–pseudo-code is provided in Algorithm 1.

**Algorithm 1** $\varepsilon$-fairness calibration

**Input:** Approximate fairness parameter $\varepsilon$, new data point $(x, s)$, base estimators $(\bar{p}_k)_k$, unlabeled sample $\mathcal{D}'_N$, $(\zeta_k)_k$ and *i.i.d* uniform perturbations $(\zeta^s_{k,i})_{k,i,s}$ in $[0, 10^{-5}]$.

**Step 0**. Split $\mathcal{D}'_N$ and construct the samples $(S_1, \ldots, S_N)$ and $\{X^s_1, \ldots, X^s_{N_s}\}$, for $s \in \mathcal{S}$;

**Step 1.** Compute the empirical frequencies $(\hat{\pi}_s)_s$ based on $(S_1, \ldots, S_N)$;

**Step 2.** Compute $\hat{\lambda}^{(1)} = (\hat{\lambda}^{(1)}_1, \ldots, \hat{\lambda}^{(1)}_K)$ and $\hat{\lambda}^{(2)} = (\hat{\lambda}^{(2)}_1, \ldots, \hat{\lambda}^{(2)}_K)$ as a solution of Eq. (1);
`Sequential quadratic programming of Section 4.1 can be used for this step.`

**Step 3.** Compute $\hat{g}$ thanks to Eq. (2);

**Output:** $\varepsilon$-fair classification $\hat{g}(x, s)$ at point $(x, s)$.

First of all, base estimators $(\bar{p}_k)_k$ are needed as inputs of the algorithm. We emphasize that we can fit any off-the-shelf estimators on the labeled dataset $\mathcal{D}_n$. In particular, one can use efficient ML algorithms that are already pre-trained and that are eventually expensive to re-train. This is one of the main advantages of post-training approaches over in-processing ones. In addition, randomization in the definition of $\bar{p}_k$ provides good theoretical properties for fairness calibration (*c.f.* Section 3.2).

Once $(\bar{p}_k)_k$ are computed, the fair classifier $\hat{g}$ relies on the estimators $\hat{\lambda}^{(1)}$ and $\hat{\lambda}^{(2)}$ computed in **Step 2.** of the algorithm. This requires solving the minimization problem in Equation (1). The corresponding objective function is convex but non-smooth due to the evaluation of the *max* function. We regularize the objective function by replacing the hard-max by a soft-max. Namely, for $\beta$ a positive real number designating the temperature parameter and $a = (a_1, \ldots, a_K)^\top \in \mathbb{R}^K$, we set

$$\mathrm{softmax}(a) := \sum_{k=1}^{K} \sigma_\beta(a)_k \cdot a_k, \quad \text{where} \quad \sigma_\beta(a)_k := \frac{\exp(a_k/\beta)}{\sum_{k=1}^{K} \exp(a_k/\beta)} \ .$$

Whenever $\beta \to 0$, the soft-max reduces to the max function. Problem (1) with the soft-max relaxation is smooth enough to be solved by a constrained optimization method, such as sequential quadratic programming [Fu et al., 2019, Nie, 2007]. Empirical study shows that $\beta = 0.005$ enables a good accuracy of the algorithm, without deviating too much from the original solution.

Instead of regularizing the objective function, one can alternatively use sampling methods such as cross-entropy optimization [Rubinstein, 1999] on the original objective function. Despite their precision, the downside of this method is the induced computational complexity, that grows much faster with the dimension than the complexity induced by smoothing techniques. Hence, the regularization approach has been preferred in the following numerical study.

## 4.2 Evaluation on synthetic data

Before illustrating our method on real datasets, we choose to evaluate our methodology on synthetic data, in order to better understand its performance.

### 4.2.1 Synthetic data

Let us define the synthetic data $(X, S, Y)$. For all $k \in [K]$ we set $\mathbb{P}(Y = k) = 1/K$. Conditional on $Y = k$, features $X \in \mathbf{R}^d$ follow a Gaussian mixture of $m$ components:

$$(X|Y = k) \sim \frac{1}{m} \sum_{i=1}^{m} \mathcal{N}_d(c^k + \mu^k_i, I_d)$$

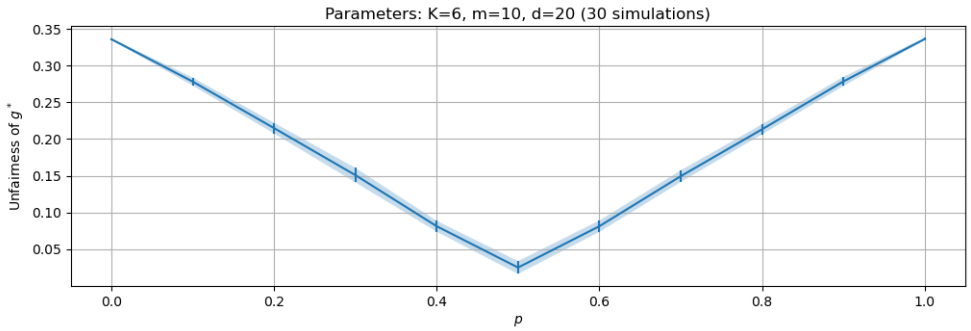


Figure 1: Unfairness of the Bayes classifier $g^*$ w.r.t. parameter $p$. We report the means and standard deviations over 30 simulations.

with $c^k \sim \mathcal{U}_d(-1,1)$, and $\mu_1^k, \ldots, \mu_m^k \sim \mathcal{N}_d(0, I_d)$; while the sensitive feature $S \in \{-1,+1\}$ follows a Bernoulli *contamination* with parameter $p$ or $1-p$ depending on $k$:

$$(S|Y=k) \sim 2 \cdot \mathcal{B}(p) - 1 \quad \text{if } k \leq \lfloor K/2 \rfloor \qquad \text{and} \qquad (S|Y=k) \sim 2 \cdot \mathcal{B}(1-p) - 1 \quad \text{if } k > \lfloor K/2 \rfloor \, .$$

From this model, we can deduce an expression of the Bayes classifier $g^*$. Indeed for each $k \in [K]$, since conditional on $Y=k$, the random variables $X$ and $S$ are independent and $\mathbb{P}(Y=k) = 1/K$, we have from the Bayes formula

$$p_k(x,s) = \frac{f_{X|Y=k}(x)\mathbb{P}(S=s|Y=k)}{\sum_{j=1}^K f_{X|Y=j}(x)\mathbb{P}(S=s|Y=j)} \ ,$$

where $f_{X|Y=k}$ is the density of $X$ conditional on $Y=k$. In view of the expression of the conditional probabilities $p_k$, the Bayes classifier $g^*$ can be expressed as

$$g^*(x,s) \in \arg\max_{k \in [K]} f_{X|Y=k}(x)\mathbb{P}(S=s|Y=k) \ .$$

We exploit the above formula to evaluate the unfairness of $g^*$ *w.r.t.* the parameter $p$. Figure 1 displays the obtained results. Interestingly, we see that parameter $p$ measures the historical bias in the dataset. Hence, this synthetic data structure enables to challenge different aspects of the algorithm. In particular, the data becomes fair when $p = 0.5$ and completely unfair when $p \in \{0,1\}$ (see also Figure 9 in Appendix D for an illustration). As default parameters, we set $K=6$, $p=0.75$, $m=10$, and $d=20$.

#### 4.2.2 Simulation scheme

We compare our method to the unfair approach. We set $u = 10^{-5}$ and estimate the conditional probabilities $p_k$ by Random Forest (RF) with default parameters in `scikit-learn`. We generate $n = 5000$ synthetic examples and split the data into three sets (60% training, 20% hold-out and 20% unlabeled).

The performance of a classifier $g$ is evaluated by its empirical accuracy $\mathrm{Acc}(g)$ on the hold-out set $\mathcal{T}$

$$Acc(g) = \frac{1}{|\mathcal{T}|} \sum_{(X,S,Y) \in \mathcal{T}} \mathbb{1}_{\{g(X,S)=Y\}} \ .$$

The unfairness of $g$ is measured on the hold-out set by the empirical counterpart $\hat{\mathcal{U}}(g)$ of the unfairness given in Definition 2.1, that is,

$$\hat{\mathcal{U}}(g) = \max_{k \in [K]} \left| \hat{\nu}_{g|-1}(k) - \hat{\nu}_{g|1}(k) \right| \ ,$$

where $\hat{\nu}_{g|s}(k) = \dfrac{1}{|\mathcal{T}^s|} \displaystyle\sum_{(X,S,Y) \in \mathcal{T}^s} \mathbb{1}_{\{g(X,S)=k\}}$ is the empirical distribution of $g(X,S)|S=s$ on the conditional hold-out test $\mathcal{T}^s = \{(X,S,Y) \in \mathcal{T} \mid S=s\}$.

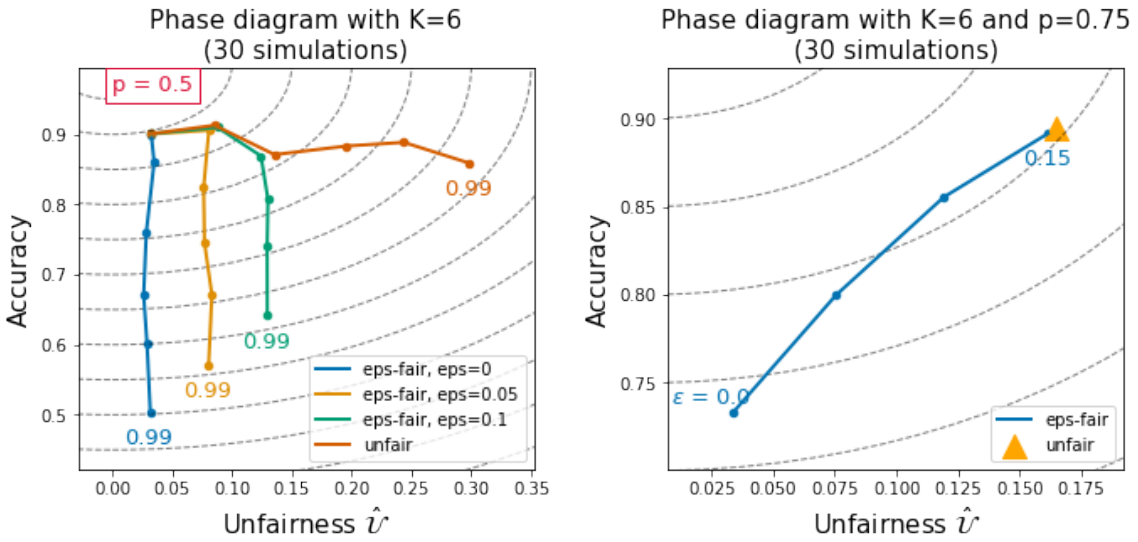


Figure 2: (Accuracy, Unfairness) phase diagrams *w.r.t.* *Left* the level of bias $p$ between 0.5 and 0.99; *Right* the accuracy-fairness trade-off parameter $\varepsilon$. Top-left corner gives the best trade-off.

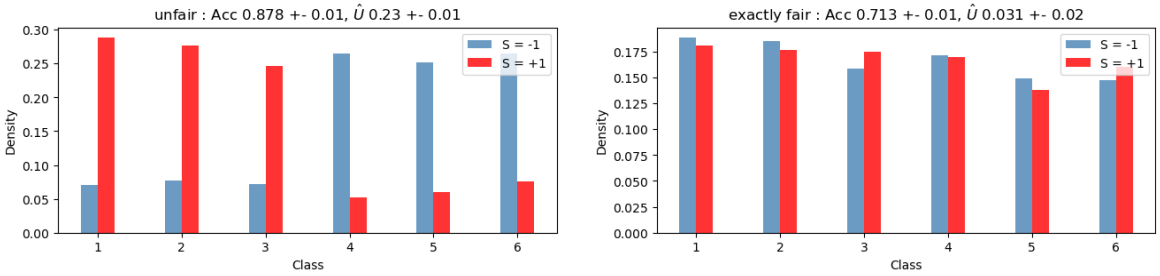


Figure 3: Empirical distribution of $\hat{g}$ on 30 simulations. *Left*: unfair classifier; *Right*: exactly-fair classifier.

### 4.2.3 Fairness versus Accuracy

Figure 2-Left illustrates how fairness and accuracy vary across different levels of unfairness, quantified by $p \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$, in both the unfair and fair random forests with $\varepsilon \in \{0, 0.05, 0.1\}$. Figure 2-Right presents the fairness and accuracy of our $\varepsilon$-fairness method for $\varepsilon \in \{0, 0.05, 0.1, 0.15\}$. Note that the performance evolves as expected: enforcing fairness degrades the accuracy and the trade-off accuracy-fairness is controlled by the parameter $\varepsilon$. From Figure 2-Right, for exact fairness ($\varepsilon = 0$), the gain in fairness is particularly salient and effective. By contrast, whenever $\varepsilon = 0.15$, the fair classifier becomes similar to the unfair method, confirming the result in Section 4.2.1 that the original unfairness of the problem is around $\varepsilon = 0.15$. From Figure 2-Left, we additionally notice that: 1) the fairness efficiency of the algorithm is particularly significant for datasets with large historical bias ($p = 0.9$ or 0.99); 2) our method succeeds at reaching the required unfairness level up to small approximation terms (vertical curves as soon as the unfairness bound $\varepsilon$ is reached); 3) as claimed in Theorem 3.3, when the unconstrained classifier is already $\varepsilon$-fair, the action of the fairness constraint on $\hat{g}_{\varepsilon-\text{fair}}$ is null and we have $\hat{g}_{\varepsilon-\text{fair}} = \hat{g}$ (horizontal parts of the curves). We also illustrates in Figure 3 that the distribution of $\hat{g}_{\text{fair}}$ is independent from $S$.

**Splitting the sample** When an unlabeled dataset is not available, the samples $\mathcal{D}_n$ and $\mathcal{D}'_N$ follow from splitting the initial dataset, see Remark 3.1. Our theoretical study relies strongly on the independence between both datasets $\mathcal{D}_n$ and $\mathcal{D}'_N$. Figure 4 numerically illustrates the importance of such condition for the fairness but also the accuracy of our proposed method. Indeed, whenever the splitting is not

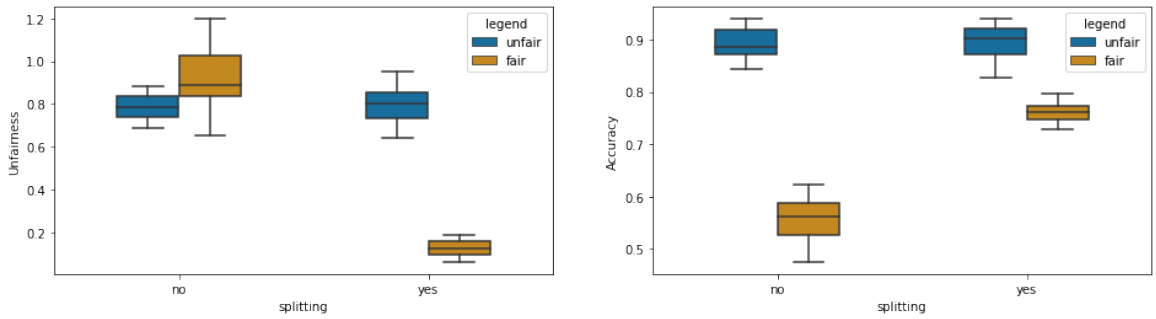


Figure 4: Empirical impact of data splitting on unfairness (Left – the lower the better) and accuracy (Right: accuracy – the higher the better). Boxplots are generated over 30 repetitions with $p = 0.75$. The non-splitting procedure involves two sets (80% training and 20% hold-out): in this particular case we use the training set (instead of the unlabeled) to compute empirical frequencies $(\hat{\pi}_s)_{s\in\mathcal{S}}$.

performed (left parts of plots), the fairness performance of the fair algorithm may even be worse than the unfair method. This emphasize that splitting is crucial and enables to avoid over-fitting on the training set.

## 4.3 Application to real datasets

In this section, we illustrate the performance of our methodology on real data and compare it with a benchmark of three State of the Art algorithms [Zhang et al., 2018, Agarwal et al., 2019, Alghamdi et al., 2022].

### 4.3.1 Datasets

The performance of the method is evaluated on two real datasets : DRUG and CRIME. Hereafter, we provide a short description of these datasets.

**Drug Consumption (DRUG)** This dataset Fehrman et al. [2017] contains demographic information such as age, gender, and education level, as well as measures of personality traits thought to influence drug use for 1885 respondents. The task is to predict cannabis use, where the 7 levels of drug use have been simplified into $K = 4$ categories (never used, not used in the past year, used in the past year, and used in the past day) for multi-class outcomes or $K = 2$ categories (used or not used in the past year) for binary outcomes. The binary sensitive feature is education level (college degree or not).

**Communities&Crime (CRIME)** This dataset contains socio-economic, law enforcement, and crime data about communities in the US with 1994 examples. The task is to predict the number of violent crimes per $10^5$ population which, we divide into $K = 5$ (multi-class outcomes) or $K = 2$ (binary outcomes) balanced classes based on equidistant quantiles. Following Calders et al. [2013], the sensitive feature is a binary variable that corresponds to the ethnicity.

### 4.3.2 Methodology

We illustrate our $\varepsilon$-fair method[2] with linear and nonlinear multi-class classification methods. For linear models, we consider one-versus-all logistic regression (reglog); for nonlinear models, Random Forest (RF) and LightGBM (GBM). For reglog, we use the default parameters in scikit-learn. For RF and GBM, we use a 3-fold cross-validation random search to select the best hyperparameters with the training set:
• For RF, we set the number of trees in $\{10, 11, \ldots, 200\}$, the maximum depth of each tree in $\{2, 3, \ldots, 16\}$, the minimum number of samples required to split an internal node in $\{2, 3, \ldots, 10\}$, and the minimum number of samples required to be at a leaf node in $\{1, \ldots, 8\}$;

[2]See `https://github.com/curiousML/epsilon-fairness`.

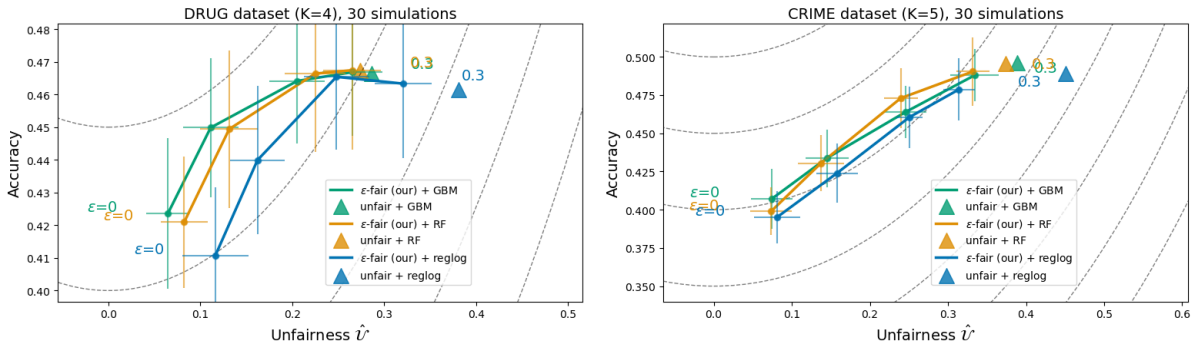


Figure 5: (Accuracy, Unfairness) phase diagrams that shows the evolution, w.r.t. the accuracy-fairness trade-off parameter $\varepsilon \in [0, 0.1, 0.2, 0.3]$. We report the means and standard deviations over the 30 repetitions. Top-left corner gives the best trade-off.

• For GBM, we set the $L1$ and $L2$ regularization term on weights both in $\{0, 0.1, 1, 2, 5, 10, 20, 50\}$, the number of boosted trees in $\{10, 11, \ldots, 200\}$, the maximum tree leaves in $\{6, 7, \ldots, 50\}$, the maximum depth of each tree in $\{2, 3, \ldots, 16\}$, and the minimum number of samples required in a child node for a split to occur in the tree in $\{10, 11, \ldots, 100\}$.

Note that the numerical experiments presented in Figure 5 confirm our findings on synthetic data. Our method have good performance in term of unfairness while the accuracy slightly increases when the level $\varepsilon$ of desired fairness increases. Besides, the performance of the $\varepsilon$-fair classifier becomes closer to the base (unfair) when the fairness constraint is released.

### 4.3.3 Benchmarks

We aim at highlighting the numerical efficiency of our method in terms of accuracy-fairness trade-off curves. For this purpose, we compare our $\varepsilon$-fairness method to the following benchmarks :

**Fair-learn** For binary classification tasks, the current state-of-the-art is established by the in-processing approach [Agarwal et al., 2019][3]. The authors present a reduction-based algorithm, which is an extention of the Fair-Lasso. The Fair-Lasso algorithm is a variant of the traditional Lasso algorithm that incorporates fairness constraints, aiming at finding a fair solution while maintaining good predictive performance. We use the following trade-off tolerances $[0.0001, 0.5, 1, 2.5, 5, 10]$.

**Fair-adversarial** The paper Zhang et al. [2018][4] presents an in-processing method for reducing bias using adversarial training: a primary model, which is trained to perform a specific task, and a bias correction model, which is trained to reduce the bias in the primary model's predictions. Note that we cannot universally apply this method on any pre-trained classifier. We use a Neural Network (NN) as the base classifier and set the following parameters: `num_epochs = 200`, `batch_size = 128`, `classifier_num_hidden_units = 50` (see the python package `AIF360`). We use the following trade-off tolerances $[0.01, 0.1, 0.5, 0.9, 1]$.

**Fair-projection** For multi-class classification tasks, we compare our result to the recent post-processing approach Alghamdi et al. [2022][5]. The authors propose a method based on information projection by reweighting the outputs of a pre-trained classifier to satisfy specific group-fairness requirements. The trade-off tolerances are $[0, 0.1, 0.2, 0.5, 0.9]$.

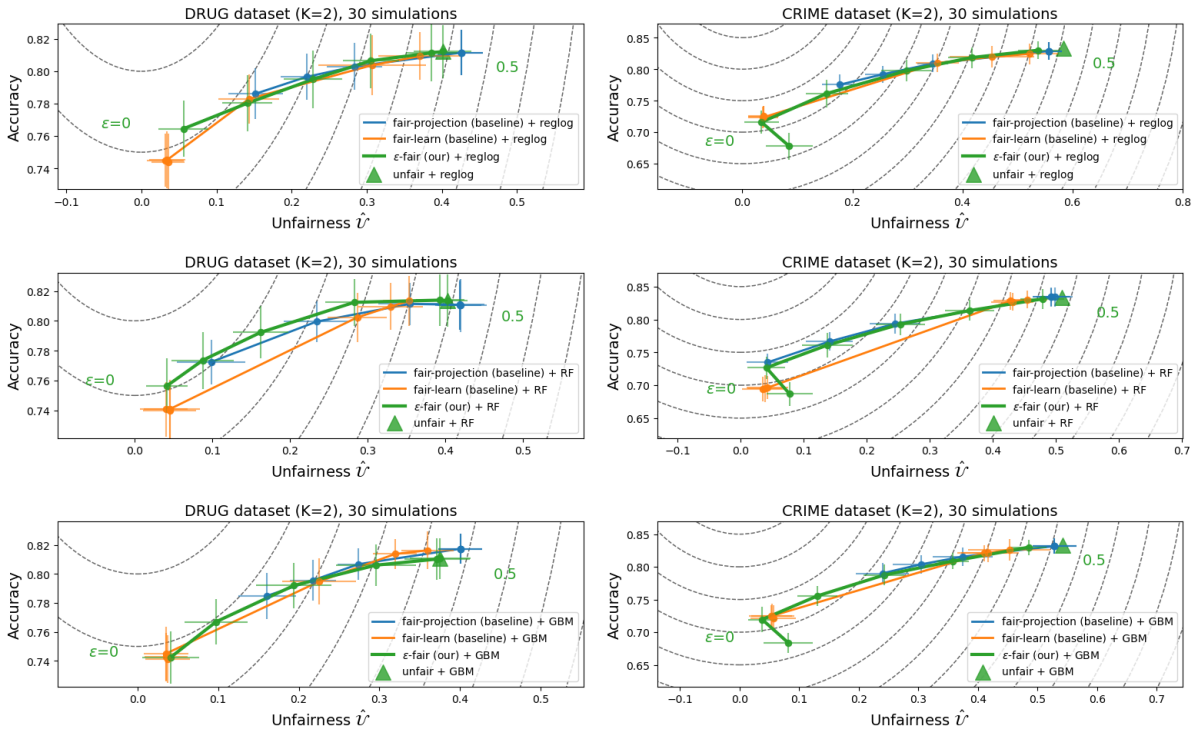


Figure 6: (Accuracy, Unfairness) phase diagrams that shows the evolution, *w.r.t.* the accuracy-fairness trade-off tolerances. We report the means and standard deviations over 30 repetitions. Top-left corner gives the best trade-off.

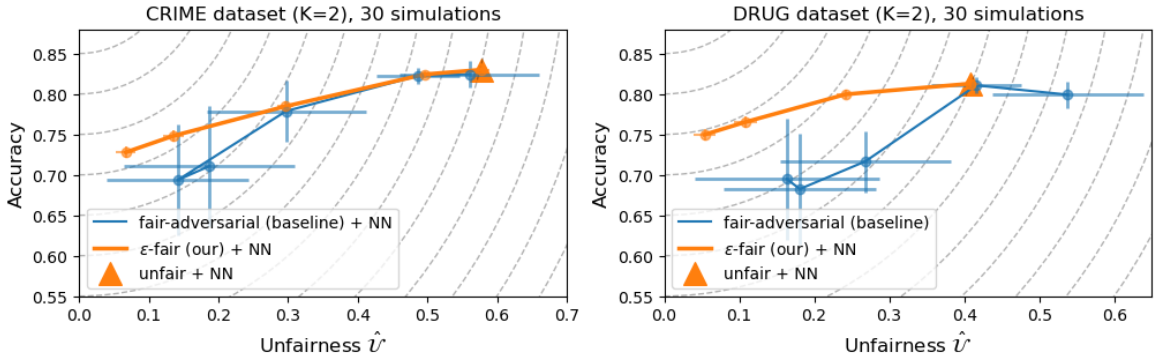


Figure 7: (Accuracy, Unfairness) phase diagrams that shows the evolution, *w.r.t.* the accuracy-fairness trade-off tolerances. For $\varepsilon$-fair classifier we vary $\varepsilon \in \{0.01, 0.1, 0.3, 0.5, 0.9\}$. We report the means and standard deviations over 30 repetitions. Top-left corner gives the best trade-off.

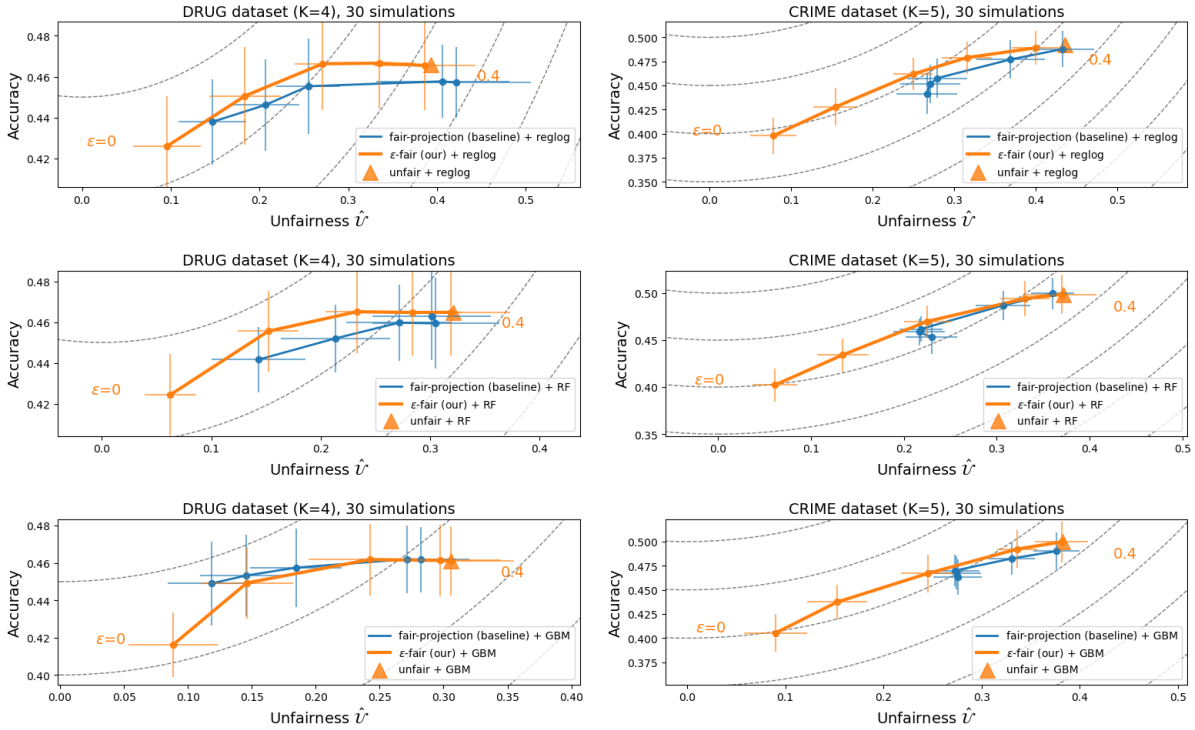


Figure 8: (Accuracy, Unfairness) phase diagrams that shows the evolution, *w.r.t.* the accuracy-fairness trade-off tolerances. We report the means and standard deviations over 30 repetitions. Top-left corner gives the best trade-off.

### 4.3.4 Results

**Performance in binary case ($K = 2$)** We analyze the efficiency of the $\varepsilon$-fairness method compared to `fair-learn`, `fair-projection` and `fair-adversarial` for binary classification. Numerical experiments on DRUG and CRIME presented in Figure 6 reveal that our method is very efficient in both accuracy and fairness and at least competitive (if not better) in several aspects :

1. **Competitive fairness.** Overall, our $\varepsilon$-fair classifier outperforms `fair-projection` classifier in terms of exact fairness ($\varepsilon = 0$) and achieves similar performance as the state-of-the-art benchmark `fair-learn`.

2. **Competitive accuracy.** Although we obtain similar accuracies using reglog and GBM, our algorithm seems more efficient than `fair-learn` using RF. Compared to `fair-projection` our algorithm is competitive in terms of accuracy for $\varepsilon \geq 0.1$ in both datasets.

From Figure 7, our $\varepsilon$-fair predictor outperforms `fair-adversarial` predictor both in terms of accuracy and fairness. Note that since `fair-learn` and `fair-adversarial` are in-processing methods their running time (using the dedicated package) is much higher than our algorithm.

**Performance in multi-class case ($K \geq 3$).** We analyze the efficiency of the $\varepsilon$-fairness method compared to the baseline `fair-projection` for multi-class classification. The numerical experiments are presented in Figure 8. In multi-class tasks, empirical results highlight the efficiency of our approach to enforce fairness when $\varepsilon$ decreases. Indeed, our methodology achieves better fairness results under the

[3]The method in [Agarwal et al., 2019] was developed for *Equality of Odds* but the code is also implemented for *Demographic Parity* see `https://github.com/fairlearn/fairlearn`.

[4]We use IBM AIF360 library `https://aif360.readthedocs.io/en/stable/modules/algorithms.html`.

[5]The code can be found at `https://github.com/HsiangHsu/Fair-Projection`.

DP constraint than `fair-projection` while maintaining competitive accuracy. Moreover, our fairness calibration is close to the pre-specified level, regardless of the base algorithm (reglog, RF or GBM).

Finally, our methodology only use a portion of the dataset to train a classifier, while reserving the remaining portion as unlabeled. Despite using relatively small datasets, consisting of about 1000 examples, our approach performed better than other benchmark methods trained on full labeled datasets.

# 5 Conclusion

In the multi-class classification framework, we provide an optimal fair classification rule under DP constraint and derive misclassification and fairness guarantees of the associated plug-in fair classifier (see Algorithm 1). We handle both exact and approximate fairness settings and show that our approach achieves distribution-free fairness and can be applied on top of any probabilistic base estimator. We also establish rates of convergence for our procedure. Up to our knowledge, the present contribution is the first statistical analysis in approximate fairness context. In particular, we consider here the multi-class setting which has rarely been studied. We finally illustrate the proficiency of our procedure on various synthetic and real datasets. Importantly, our algorithm is efficient for enforcing a pre-specified level of fairness. A natural way for further research is to extend our methodology to other notions of fairness such as *equalized odds* and also to consider settings of multi-category sensitive attributes. We believe that the present work is a relevant step to handle these two problems. On the other hand, the calibration of the level of unfairness $\varepsilon \geq 0$ is an important empirical issue. As mentioned in the introduction, there are some heuristics that provide guidelines for its calibration but one may ask for more advanced and robust approaches. In particular, a future direction of research is to describe a methodology that statistically justifies a data-driven calibration of this parameter in order to optimally compromise risk and unfairness.

# References


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.

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

A. Agarwal, M. Dudik, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, 2019.

W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P.W. Michalak, S. Asoodeh, and F. Calmon. Beyond adult and COMPAS: Fair multi-class prediction via information projection. In *In Neural Information Processing Systems*, 2022.

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

S. Barocas and A. Selbst. Big Data's Disparate Impact. *SSRN eLibrary*, 2014.

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

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

T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *IEEE International Conference on Data Mining*, 2013.

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.

S. Chiappa, R. Jiang, T. Stepleton, A. Pacchiano, H. Jiang, and J. Aslanides. A general approach to fairness with optimal transport. In *AAAI*, 2020.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems*, 2019.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression via plug-in estimator and recalibrationwith statistical guarantees. In *Advances in Neural Information Processing Systems*, 2020a.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair Regression via Plug-In Estimator and Recalibration. NeurIPS20, 2020b.

B. Collins. Tackling unconscious bias in hiring practices: The plight of the rooney rule. *NYU Law Review*, 82:870–912, 2007.

J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, pages 296–299. Auerbach Publications, 2018.

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*, 2018a.

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*, 2018b.

E. Fehrman, A. Muhammad, E. Mirkes, V. Egan, and A. Gorban. The five factor model of personality and evaluation of drug consumption risk. In *Data science*, pages 231–242. Springer, 2017.

M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Z. Fu, G. Liu, and L. Guo. Sequential quadratic programming method for nonlinear least squares estimation and its application. *Mathematical Problems in Engineering*, 2019.

P. Gordaliza, E. Del Barrio, G. Fabrice, and J. M. Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, 2019.

S. Hajian, J. Domingo-Ferrer, and A. Martínez-Ballesté. Discrimination prevention in data mining for intrusion and crime detection. In *2011 IEEE Symposium on Computational Intelligence in Cyber Security (CICS)*, pages 47–54, 2011.

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

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

H. Holzer and D. Holzer. Assessing affirmative action. *Journal of Economic Literature*, 38(3):483–568, 2000.

R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. *arXiv preprint arXiv:1907.12059*, 2019.

F. Kamiran, I. Zliobaite, and T. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl. Inf. Syst.*, 35(3):613–644, 2013.

T. Le Gouic, J.-M. Loubes, and P. Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

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

P.-y. Nie. Sequential penalty quadratic programming filter methods for nonlinear programming. *Nonlinear Analysis: Real World Applications*, 8(1):118–129, 2007.

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

R. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and computing in applied probability*, 1(2):127–190, 1999.

S.K. Tavker, H.G. Ramaswamy, and H. Narasimhan. Consistent plug-in classifiers for complex objectives and constraints. In *Advances in Neural Information Processing Systems*, volume 33, pages 20366–20377, 2020.

Q. Ye and W. Xie. Unbiased subdata selection for fair classification: A unified framework and scalable algorithms. *arXiv preprint arXiv:2012.12356*, 2020.

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.

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.

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

B. Hu Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593, 2018.

# Appendix

In this section, we gather the proofs of our results. Section A is devoted to useful technical results. In Section B we give the proof of the results related to the optimal fair predictors while Section C is dedicated to the theoretical properties of our estimation procedure. Finally, we provide additional numerical results in Section D. In all the sequel, $C$ denotes a generic constant, whose value may vary from line to line.

## A Technical results

**Lemma A.1** (Hoeffding)**.** *Let $Z \sim \mathcal{B}(N,p)$, with $p \in (0,1)$. We then have for all $t>0$ and $N > \frac{t}{p}$*

$$\mathbb{P}(Z \leq t) \leq \exp\left(-2N(p-t/N)^2\right).$$

**Lemma A.2.** *Let $Z \sim \mathcal{B}(N,p)$. We have that*

$$\mathbb{E}\left[\frac{\mathbb{1}_{\{Z\geq 1\}}}{Z}\right] \leq \frac{2}{(N+1)p}$$

**Proposition A.3.** *Let $f:\ \mathbb{R}^M \to \mathbb{R}$ be a convex continuous function, and $\mathcal{H} \subset \mathbb{R}^M$ a closed convex set. We consider the minimizer $\mathbf{x}^*$ of the function $f$ over the set $\mathcal{H}$*

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{H}} f(\mathbf{x}).$$

*Then, there exists a subgradient $\mathfrak{h}$ in the subdifferential $\partial f(\mathbf{x}^*)$ of $f$ at the point $\mathbf{x}^*$ such that*

$$\mathfrak{h}^T(\mathbf{y}-\mathbf{x}^*) \geq 0, \quad \forall \mathbf{y} \in \mathcal{H}.$$

From the above proposition, it is easy to show the following result.

**Corollary A.4.** *Let $f:\ \mathbb{R}^M \to \mathbb{R}$ be a convex continuous function. Let $\mathcal{H} = \mathbb{R}^M_+$. We consider the minimizer $\mathbf{x}^*$ of the function $f$ over the set $\mathcal{H}$. Let $\mathcal{M} := \{m \in [M],\ \mathbf{x}^*_m \neq 0\}$. Then there exists a subgradient $\mathfrak{h} \in \partial f(\mathbf{x}^*)$, such that for all $m \in [M]$ we have $\mathfrak{h}_m \geq 0$ and in particular,*

$$\forall m \in \mathcal{M}, \quad \mathfrak{h}_m = 0.$$

## B Proof of Section 2

We begin with an auxiliary lemma, which provides an alternative useful representation of $\mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}(g)$.

**Lemma B.1.** *Let $\varepsilon \geq 0$, the $\varepsilon$-fair-risk of a classifier $g$ with tuning parameters $\lambda^{(1)} = (\lambda_1^{(1)},\ldots,\lambda_K^{(1)}) \in \mathbb{R}_+^K, \lambda^{(2)} = (\lambda_1^{(2)},\ldots,\lambda_K^{(2)}) \in \mathbb{R}_+^K$ reads as:*

$$\mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}(g) = \sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\sum_{k=1}^K \left(\pi_s p_k(X,S) - s(\lambda_k^{(1)} - \lambda_k^{(2)})\right)\mathbb{1}_{\{g(X,S)\neq k\}}\right] - \varepsilon\sum_{k=1}^K(\lambda_k^{(1)}+\lambda_k^{(2)}). \quad (4)$$

*Proof of Lemma B.1.* Let $(\lambda^{(1)},\lambda^{(2)}) \in \mathbb{R}_+^{2K}$ and recall the following definition of the $\varepsilon$-fair risk

$$\mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}(g) = \mathbb{P}\left(g(X,S)\neq Y\right) - \sum_{k=1}^K\sum_{s\in\mathcal{S}} s(\lambda_k^{(1)}-\lambda_k^{(2)})\,\mathbb{E}_{X|S=s}\left[\mathbb{1}_{\{g(X,s)\neq k\}}\right] - \varepsilon\sum_{k=1}^K(\lambda_k^{(1)}+\lambda_k^{(2)}). \quad (5)$$

The result in (4) directly follows from the following decomposition

$$\begin{aligned}
\mathbb{P}\left(g(X,S)\neq Y\right) &= \sum_{k=1}^K \mathbb{E}\left[\mathbb{1}_{\{g(X,S)\neq k\}}\mathbb{1}_{\{Y=k\}}\right] \\
&= \sum_{k=1}^K\sum_{s\in\mathcal{S}} \mathbb{E}\left[\mathbb{1}_{\{g(X,S)\neq k\}}\mathbb{1}_{\{S=s\}}p_k(X,S)\right] \\
&= \sum_{k=1}^K\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\mathbb{1}_{\{g(X,s)\neq k\}}\pi_s p_k(X,s)\right] \ .
\end{aligned}$$

□

*Proof of Theorem 2.3.* The proof is divided into two parts. First, we provide the proof for $\varepsilon > 0$. Then the second part is dedicated to the proof of the result when $\varepsilon = 0$ which corresponds to the case of exact fairness.

**Proof for approximate fairness** From Lemma B.1, we deduce that $g^*_{\lambda^{(1)},\lambda^{(2)}}$ should be defined for all $(x, s) \in \mathcal{X} \times \mathcal{S}$ as

$$g^*_{\lambda^{(1)},\lambda^{(2)}}(x, s) = \arg\max_{k\in[K]} \left(\pi_s p_k(X, S) - s(\lambda_k^{(1)} - \lambda_k^{(2)})\right) \ , \tag{6}$$

since it minimizes the risk $\mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}$. Now we should maximize $\mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}(g^*_{\lambda^{(1)},\lambda^{(2)}})$ in the dual variables. Notice that the $\varepsilon$-fair risk can be written as

$$\mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}(g^*_{\lambda^{(1)},\lambda^{(2)}}) = 1 - \sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_{k\in[K]} \left(\pi_s p_k(X, S) - s(\lambda_k^{(1)} - \lambda_k^{(2)})\right)\right] - \varepsilon \sum_{k=1}^{K} (\lambda_k^{(1)} + \lambda_k^{(2)}) \ .$$

Hence, a maximizer $(\lambda^{*(1)}, \lambda^{*(2)})$ in $\mathbb{R}_+^{2K}$ of $(\lambda^{(1)}, \lambda^{(2)}) \mapsto \mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}(g^*_{\lambda^{(1)},\lambda^{(2)}})$ is solution of

$$(\lambda^{*(1)}, \lambda^{*(2)}) \in \arg\min_{(\lambda^{(1)},\lambda^{(2)})\in\mathbb{R}_+^{2K}} \underbrace{\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_{k\in[K]} \left(\pi_s p_k(X, s) - s(\lambda_k^{(1)} - \lambda_k^{(2)})\right)\right] + \varepsilon \sum_{k=1}^{K} (\lambda_k^{(1)} + \lambda_k^{(2)})}_{H(\lambda^{(1)},\lambda^{(2)})} \ .$$

The rest of the proof consists in showing that such a calibration of the tuning parameters $(\lambda^{(1)}, \lambda^{(2)})$ implies that $g^*_{\lambda^{*(1)},\lambda^{*(2)}}$ is indeed $\varepsilon$-fair. Observe that

$$H(\lambda^{(1)}, \lambda^{(2)}) \geq \varepsilon \sum_{k=1}^{K} (\lambda_k^{(1)} + \lambda_k^{(2)}) \ ,$$

and then $\lim_{\|(\lambda^{(1)},\lambda^{(2)})\|_2^2\to\infty} H(\lambda^{(1)}, \lambda^{(2)}) = +\infty$. Moreover, the mapping $H$ is continuous and convex in $(\lambda^{(1)}, \lambda^{(2)})$. Therefore the minimum $(\lambda^{*(1)}, \lambda^{*(2)})$ exists, and there exists some constant $C_\lambda > 0$ such that for all $k \in [K]$ and $j \in \{1, 2\}$ we have $|\lambda_k^{(j)}| \leq C_\lambda$.

Let us derive a subgradient $\mathfrak{h}^* = (\mathfrak{h}^{*(1)}, \mathfrak{h}^{*(2)})$ of $H$ at the optimum $(\lambda^{*(1)}, \lambda^{*(2)})$ with $\mathfrak{h}^{*(1)} = \left(\mathfrak{h}_1^{*(1)}, \ldots, \mathfrak{h}_K^{*(1)}\right)$ and $\mathfrak{h}^{*(2)} = \left(\mathfrak{h}_1^{*(2)}, \ldots, \mathfrak{h}_K^{*(2)}\right)$ being two vectors in $\mathbb{R}^K$. In order to express $\mathfrak{h}^*$ let us build the subdifferential of the function $f\left(x, (\lambda^{(1)}, \lambda^{(2)})\right) := \max_{k\in[K]} \left\{h_k^s\left(x, (\lambda^{(1)}, \lambda^{(2)})\right)\right\}$ at the point $(\lambda^{*(1)}, \lambda^{*(2)})$ with

$$h_k^s\left(x, (\lambda^{(1)}, \lambda^{(2)})\right) = \pi_s p_k(x, s) - s\left(\lambda_k^{(1)} - \lambda_k^{(2)}\right).$$

We have that

$$\begin{aligned}
&\partial f\left(x, (\lambda^{*(1)}, \lambda^{*(2)})\right) \\
&\qquad = \operatorname{conv}\left\{\nabla h_k^s\left(x, (\lambda^{*(1)}, \lambda^{*(2)})\right) \ : \ h_k^s\left(x, (\lambda^{*(1)}, \lambda^{*(2)})\right) = \max_{j\in[K]} \left\{h_j^s\left(x, (\lambda^{*(1)}, \lambda^{*(2)})\right)\right\}\right\},
\end{aligned}$$

where $\nabla h_k^s\left(x, (\lambda^{(1)}, \lambda^{(2)})\right) \in \mathbb{R}^{2K}$ is the gradient of the function $h_k^s$ *w.r.t.* $(\lambda^{(1)}, \lambda^{(2)})$. Therefore, we deduce that a subgradient $\mathfrak{h}^*$ of $H$ at $(\lambda^{*(1)}, \lambda^{*(2)})$ can be expressed for each $k \in [K]$, and $l \in \{1, 2\}$ as

$$\begin{aligned}
\mathfrak{h}_k^{*(l)} = (2l-3) \sum_{s\in\mathcal{S}} \Big\{ & s\mathbb{P}_{X|S=s}\left(\forall j \neq k \ (\pi_s p_k(X, s) - s(\lambda_k^{*(1)} - \lambda_k^{*(2)})) > (\pi_s p_j(X, s) - s(\lambda_j^{*(1)} - \lambda_j^{*(2)}))\right) \\
& + s\, u_k^s\, \mathbb{P}_{X|S=s}\Big(\forall j \neq k \ (\pi_s p_k(X, s) - s(\lambda_k^{*(1)} - \lambda_k^{*(2)})) \geq (\pi_s p_j(X, s) - s(\lambda_j^{*(1)} - \lambda_j^{*(2)})), \\
& \qquad \exists j \neq k \ (\pi_s p_k(X, s) - s(\lambda_k^{*(1)} - \lambda_k^{*(2)})) = (\pi_s p_j(X, s) - s(\lambda_j^{*(1)} - \lambda_j^{*(2)}))\Big)\Big\} + \varepsilon \ ,
\end{aligned}$$

with $u_k^s \in [0,1]$ for all $k \in [K]$ and all $s \in \mathcal{S}$. Thanks to Assumption 2.2, $p_k(X,s) - p_j(X,s)$ has no atom for all $s \in \mathcal{S}$ and then the second part of the r.h.s. of the above equation vanishes and we have

$$\mathfrak{h}_k^{*(l)} = (2l-3)\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\forall j \neq k\ (\pi_s p_k(X,s) - s(\lambda_k^{*(1)} - \lambda_k^{*(2)})) > (\pi_s p_j(X,s) - s(\lambda_j^{*(1)} - \lambda_j^{*(2)}))\right) + \varepsilon,$$

which can be written as

$$\mathfrak{h}_k^{*(l)} = (2l-3)\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) + \varepsilon.$$

Now, we apply Corollary A.4 and deduce, from the above equation, that if

- $\lambda_k^{*(1)} \neq 0$ and $\lambda_k^{*(2)} \neq 0$, we then necessary have $\mathfrak{h}_k^{*(l)} = 0$ for $l \in \{1,2\}$ and then

$$\begin{aligned}\mathbb{P}_{X|S=1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) - \mathbb{P}_{X|S=-1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) &= \varepsilon \\ \mathbb{P}_{X|S=1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) - \mathbb{P}_{X|S=-1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) &= -\varepsilon\ ,\end{aligned}$$

  which leads to a contradiction.

- $\lambda_k^{*(1)} = 0$ and $\lambda_k^{*(2)} = 0$, we get

$$\begin{aligned}\mathbb{P}_{X|S=1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) - \mathbb{P}_{X|S=-1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) &\leq \varepsilon \\ \mathbb{P}_{X|S=1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) - \mathbb{P}_{X|S=-1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) &\geq -\varepsilon\ ,\end{aligned}$$

  which gives

$$\left|\mathbb{P}_{X|S=1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) - \mathbb{P}_{X|S=-1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right)\right| \leq \varepsilon.$$

- Finally, if $\lambda_k^{*(1)}\lambda_k^{*(2)} = 0$ and $\lambda_k^{*(1)} + \lambda_k^{*(2)} > 0$, we get

$$\left|\mathbb{P}_{X|S=1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) - \mathbb{P}_{X|S=-1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right)\right| = \varepsilon.$$

Hence, we have shown that for each $k \in [K]$,

$$\left|\mathbb{P}_{X|S=1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) - \mathbb{P}_{X|S=-1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right)\right| \leq \varepsilon,$$

which means that $g^*_{\lambda^{*(1)},\lambda^{*(2)}}$ is $\varepsilon$-fair: $\mathcal{U}(g^*_{\lambda^{*(1)},\lambda^{*(2)}}) \leq \varepsilon$.

Furthermore, we also have that for each $k \in [K]$, the vector $(\lambda^{*(1)}, \lambda^{*(2)})$ meets the following constraint $\lambda_k^{*(1)}\lambda_k^{*(2)} = 0$ and $\lambda_k^{*(1)} + \lambda_k^{*(2)} \geq 0$. Since parameters $(\lambda^{*(1)}, \lambda^{*(2)})$ are bounded, we then deduce that for any classifier $g$ (see for instance (5))

$$\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g) \leq R(g) + C\left(\mathcal{U}(g) - \varepsilon\right),$$

therefore, for any $g \in \mathcal{G}_{\varepsilon-\text{fair}}$

$$\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g) \leq R(g). \tag{7}$$

Besides, considering the three above cases, we notice that

$$\left|\mathbb{P}_{X|S=1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right) - \mathbb{P}_{X|S=-1}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}(X,S) = k\right)\right| < \varepsilon \Rightarrow \lambda_k^{*(1)} = \lambda_k^{*(2)} = 0.$$

Since $g^*_{\lambda^{*(1)},\lambda^{*(2)}} \in \mathcal{G}_{\varepsilon-\text{fair}}$, the above equation and Equation (7) imply that for any $g \in \mathcal{G}_{\varepsilon-\text{fair}}$

$$R(g^*_{\lambda^{*(1)},\lambda^{*(2)}}) = \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g^*_{\lambda^{*(1)},\lambda^{*(2)}}) \leq \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g) \leq R(g),$$

which concludes the proof.

**Proof for exact fairness** First, we apply Lemma B.1 with $\varepsilon = 0$ and then have

$$\mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}(g^*_{\lambda^{(1)},\lambda^{(2)}}) = 1 - \sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_{k\in[K]}\left(\pi_s p_k(X,S) - s(\lambda_k^{(1)} - \lambda_k^{(2)})\right)\right],$$

with

$$g^*_{\lambda^{(1)},\lambda^{(2)}}(x,s) = \arg\max_{k\in[K]}\left(\pi_s p_k(X,S) - s(\lambda_k^{(1)} - \lambda_k^{(2)})\right) \ .$$

Therefore, it is not difficult to see that using the reparametrization

$$\beta_k = \lambda_k^{(1)} - \lambda_k^{(2)}, \ \ k = 1,\dots,K, \tag{8}$$

we can write

$$\mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}(g^*_{\lambda^{(1)},\lambda^{(2)}}) = \mathcal{R}_\beta(g^*_\beta) = 1 - \sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_{k\in[K]}\left(\pi_s p_k(X,s) - s\beta_k\right)\right] \ . \tag{9}$$

Hence, a maximizer $\beta^*$ in $\mathbb{R}^K$ of $\beta \mapsto \mathcal{R}_\beta(g^*_\beta)$ takes the form

$$\beta^* \in \arg\min_{\beta\in\mathbb{R}^K} \sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_{k\in[K]}\left(\pi_s p_k(X,s) - s\beta_k\right)\right] \ .$$

The above criterion is convex in $\beta$. Therefore, first order optimality conditions for the minimization over $\beta$ of the above criterion imply that, for each $k \in [K]$,

$$\begin{aligned} 0 = &\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\forall j \neq k \ (\pi_s p_k(X,s) - s\beta_k^*) > (\pi_s p_j(X,s) - s\beta_j^*)\right) \\ &+ s u_k^s \mathbb{P}_{X|S=s}\left(\forall j \neq k \ (\pi_s p_k(X,s) - s\beta_k^*) \geq (\pi_s p_j(X,s) - s\beta_j^*), \ \exists j \neq k \ (\pi_s p_k(X,s) - s\beta_k^*) = (\pi_s p_j(X,s) - s\beta_j^*)\right) \ , \end{aligned}$$

with $u_k^s \in [0,1]$ for all $k \in [K]$ and $s \in \mathcal{S}$. As in the case where $\varepsilon > 0$, we use Assumption 2.2 on the distribution of $p_k(X,s) - p_j(X,s)$ to show that the second part of the r.h.s. vanishes. Therefore for all $k \in [K]$

$$\mathbb{P}_{X|S=1}\left(g^*_{\beta^*}(X,S) \neq k\right) = \mathbb{P}_{X|S=-1}\left(g^*_{\beta^*}(X,S) \neq k\right) \ ,$$

meaning that the classifier $g^*_{\beta^*}$ is fair. Furthermore, for any fair classifier $g \in \mathcal{G}_{\text{fair}}$, we observe that

$$\mathcal{R}(g^*_{\beta^*}) = \mathcal{R}_{\beta^*}(g^*_{\beta^*}) \leq \mathcal{R}_{\beta^*}(g) = \mathcal{R}(g),$$

so that $g^*_{\beta^*}$ is also an optimal fair classifier.

Conversely, consider any optimal fair classifier $g^*_{\text{fair}} \in \mathcal{G}_{\text{fair}}$. Combining the fairness of $g^*_{\text{fair}}$ with the optimality of $\beta^*$ over the family $(R_\beta(g^*_\beta))_{\beta\in\mathbf{R}^K}$, we deduce

$$\mathcal{R}_{\beta^*}(g^*_{\text{fair}}) = \mathcal{R}(g^*_{\text{fair}}) \leq \mathcal{R}_{\beta^*}(g^*_{\beta^*}) \leq \mathcal{R}_{\beta^*}(g), \text{for any } g \in \mathcal{G} \, .$$

Hence any optimal fair classifier is a minimizer of $\mathcal{R}_{\beta^*}$ over $\mathcal{G}$.

□

# C Proof of Section 3

We first introduce some notation. We recall that $N_{\min} = \min(N_1, N_{-1})$ and denote by $\hat{P}_{X|S=s}$ the empirical measure with respect to $(X_1^s,\dots,X_{N_s}^s)$ for $s \in \mathcal{S}$. Furthermore, throughout this section, we consider the following convention $\frac{0}{0} = 0$. Hence, if $N_s = 0$, we then have $\hat{P}_{X|S=s}(A) = 0$ for any event $A$.

We start this section with two results. Lemma-C.1 directly follows from similar arguments as in the proof of Lemma B.8 in Chzhen et al. [2020a]. Its proof is hence omitted.

**Lemma C.1.** *Conditional on the data, we have that, for each $s \in \mathcal{S}$ and $k \in [K]$,*

$$\begin{aligned}\hat{\mathbb{P}}_{X|S=s}\left(\exists j \neq k, \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) = \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)})\right) &= \frac{1}{N_s}\sum_{i=1}^{N_s} \mathbb{1}_{\left\{\exists j \neq k, \hat{h}_k^s(X_i^s, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) = \hat{h}_j^s(X_i^s, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)})\right\}} \\ &\leq \frac{K-1}{N_s} \ \ a.s.\ ,\end{aligned}$$

*where $\hat{h}_k^s : (x, \lambda^{(1)}, \lambda^{(2)}) \mapsto \hat{\pi}_s \bar{p}_k(x, s) - s\left(\lambda^{(1)} - \lambda^{(2)}\right)$.*

**Lemma C.2.** *Let us introduce for all $k \in [K]$ the random variable*

$$\hat{A}_k = \left| \sum_{s \in \mathcal{S}} s\left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s}\right)\left(\forall j \neq k\ \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)})\right)\right| \ .$$

*Then all $k \in [K]$*

1. *there exists $C_1 > 0$, that depends on $K$ such that*

$$\mathbb{E}\left[\hat{A}_k \mathbb{1}_{\{N_{\min} \geq 1\}} \mid \mathcal{D}_n, S_1, \ldots, S_N\right] \leq \frac{C_1 \mathbb{1}_{\{N_{\min} \geq 1\}}}{\sqrt{N_{\min}}} \ ;$$

2. *for all $\delta > 0$, the event $\mathcal{A}_k(\delta) = \left\{\hat{A}_k \leq K\sqrt{\frac{2\log(\frac{4K}{\delta})}{N_{\min}}}\right\} \bigcap \{N_{\min} \geq 1\}$ holds with probability greater than $1 - \delta$.*

*Proof.* 1. For this part, we work on the event $\{N_{\min} \geq 1\}$ conditionally on $\mathcal{D}_n$ and on $S_1, \ldots, S_N$. For $s \in \{-1, 1\}$, and $k \in [K]$, we have

$$\begin{aligned}&\left|\left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s}\right)\left(\forall j \neq k,\ \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)})\right)\right| = \\ &\quad\left|\left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s}\right)\left(\forall j \neq k,\ \bar{p}_k(X, s) - \bar{p}_j(X, s) > \frac{s\left((\hat{\lambda}_k^{(1)} - \hat{\lambda}_k^{(2)}) - (\hat{\lambda}_j^{(1)} - (\hat{\lambda}_j^{(2)})\right)}{\hat{\pi}_s}\right)\right| \\ &\qquad\leq \sum_{j=1}^{K} \sup_{t \in \mathbb{R}} \left|\left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s}\right)(\bar{p}_k(X, s) - \bar{p}_j(X, s) > t)\right| \ .\end{aligned}$$

Therefore, from the Dvoretzky-Kiefer-Wolfowitz Inequality, we deduce that, for each $s \in \mathcal{S}$ and $k \in [K]$

$$\mathbb{E}\left[\hat{A}_k \mathbb{1}_{\{N_{\min} \geq 1\}} \mid \mathcal{D}_n, S_1, \ldots, S_N\right] \leq \frac{C_1 \mathbb{1}_{\{N_{\min} \geq 1\}}}{\sqrt{N_{\min}}}.$$

2. From the Dvoretzky-Kiefer-Wolfowitz Inequality, conditional on $\mathcal{D}_n$ and on $(S_1, \ldots, S_N)$, we have on the event $\{N_{\min} \geq 1\}$, for each $u > 0$ and for all $j, k \in [K]$, $s \in \mathcal{S}$, and $t > 0$

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} \left|\left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s}\right)(\bar{p}_k(X, s) - \bar{p}_j(X, s) > t)\right| \geq u\right) \leq 2\exp(-2N_s u^2) \leq 2\exp(-2N_{\min} u^2).$$

Since

$$\hat{A}_k \leq \sum_{s \in \mathcal{S}} \sum_{j=1}^{K} \sup_{t \in \mathbb{R}} \left|\left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s}\right)(\bar{p}_k(X, s) - \bar{p}_j(X, s) > t)\right|,$$

we deduce for each $u > 0$ and $k \in [K]$

$$\begin{aligned}\mathbb{P}\left(\hat{A}_k \geq u\right) &\leq \sum_{s \in \mathcal{S}} \sum_{j=1}^{K} \mathbb{P}\left(\sup_{t \in \mathbb{R}} \left|\left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s}\right)(\bar{p}_k(X, s) - \bar{p}_j(X, s) > t)\right| \geq \frac{u}{2K}\right) \\ &\leq 4K\exp\left(-\frac{u^2 N_{\min}}{2K^2}\right).\end{aligned}$$

Hence, from the above inequality, we obtain that

$$\mathbb{1}_{\{N_{\min}\geq 1\}}\mathbb{P}\left(\hat{A}_k \geq K\sqrt{\frac{2\log(\frac{4K}{\delta})}{N_{\min}}}\,\middle|\,\mathcal{D}_n,(S_1,\ldots,S_N)\right) \leq \mathbb{1}_{\{N_{\min}\geq 1\}}\delta \leq \delta,$$

which yields the desired result.

□

Let us now consider the proofs of Theorem 3.2 and Theorem 3.4.

*Proof of Theorem 3.2.* As in the proof of Theorem 2.3], we consider separately the cases of approximate ($\varepsilon > 0$) and exact ($\varepsilon = 0$) fairness.

**Unfairness control in the case of approximate fairness** We first consider the case where $\varepsilon > 0$. As in Lemma C.1, we first introduce, for $s \in \mathcal{S}$ and $k \in [K]$,

$$\hat{h}_k^s : \left(x, \lambda^{(1)}, \lambda^{(2)}\right) \mapsto \hat{\pi}_s \bar{p}_k(x,s) - s\left(\lambda^{(1)} - \lambda^{(2)}\right) \ .$$

By construction, the estimator $\bar{p}_k(X,S)$ is randomized and then satisfies an analog version of Assumption 2.2. Therefore for all $s \in \mathcal{S}$ and $k \in [K]$

$$\mathbb{P}_{X|S=s}\left(\hat{g}_\varepsilon(X,S)=k\right) = \mathbb{P}_{X|S=s}\left(\forall j \neq k, \hat{h}_k^s(X,\hat{\lambda}_k^{(1)},\hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X,\hat{\lambda}_j^{(1)},\hat{\lambda}_j^{(2)})\right) \ . \tag{10}$$

Now, we consider similar arguments as in Proof of Theorem 2.3. First we observe that

$$\hat{H}\left(\lambda^{(1)},\lambda^{(2)}\right)) \geq \varepsilon \sum_{k=1}^{K}(\lambda_k^{(1)} + \lambda_k^{(2)}) \ , \tag{11}$$

where $\hat{H}$ is the empirical version of $H$ and is defined as

$$\hat{H}(\lambda^{(1)},\lambda^{(2)}) = \sum_{s\in\mathcal{S}} \hat{\mathbb{E}}_{X|S=s}\left[\max_{k\in[K]}\left(\pi_s\bar{p}_k(X,s) - s(\lambda_k^{(1)} - \lambda_k^{(2)})\right)\right] + \varepsilon\sum_{k=1}^{K}(\lambda_k^{(1)} + \lambda_k^{(2)}) \ ,$$

with $\hat{\mathbb{E}}_{X|S=s}$ being the empirical expectation over the points $X_i$ from the dataset $\mathcal{D}'_N$ such that the sensitive attribute $S_i = s$. From Equation (11), we deduce that the minimizer $(\hat{\lambda}^{(1)},\hat{\lambda}^{(2)})$ exists and is bounded by some $C'_\lambda > 0$ which depends neither on $N$ nor on $n$. Furthermore, we have that a subgradient $\hat{\mathfrak{h}}$ of $\hat{H}$ can be expressed for each $k \in [K]$ and $l \in \{1,2\}$ as follows

$$\begin{aligned}\hat{\mathfrak{h}}_k^{(l)} = (2l-3)\sum_{s\in\mathcal{S}}\Big\{ & s\hat{\mathbb{P}}_{X|S=s}\left(\forall j \neq k\ \hat{h}_k^s(X,\hat{\lambda}_k^{(1)},\hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X,\hat{\lambda}_j^{(1)},\hat{\lambda}_j^{(2)})\right) \\ & + s\,u_k^s\,\hat{\mathbb{P}}_{X|S=s}\Big(\forall j \neq k\ \hat{h}_k^s(X,\hat{\lambda}_k^{(1)},\hat{\lambda}_k^{(2)}) \geq \hat{h}_j^s(X,\hat{\lambda}_j^{(1)},\hat{\lambda}_j^{(2)}), \\ & \qquad \exists j \neq k\ \hat{h}_k^s(X,\hat{\lambda}_k^{(1)},\hat{\lambda}_k^{(2)}) = \hat{h}_j^s(X,\hat{\lambda}_j^{(1)},\hat{\lambda}_j^{(2)})\Big)\Big\} + \varepsilon \ ,\end{aligned} \tag{12}$$

with $u_k^s \in [0,1]$. Applying Lemma C.1, we observe that the second term in r.h.s is such that

$$\begin{aligned}0 \leq \hat{\mathbb{P}}_{X|S=s}\Big(\forall j \neq k\ \hat{h}_k^s(X,\hat{\lambda}_k^{(1)},\hat{\lambda}_k^{(2)}) \geq \hat{h}_j^s(X,\hat{\lambda}_j^{(1)},\hat{\lambda}_j^{(2)}), \\ \exists j \neq k\ \hat{h}_k^s(X,\hat{\lambda}_k^{(1)},\hat{\lambda}_k^{(2)}) = \hat{h}_j^s(X,\hat{\lambda}_j^{(1)},\hat{\lambda}_j^{(2)})\Big) \leq \frac{K-1}{N_{\min}} \ .\end{aligned} \tag{13}$$

Hereafter, we follow the same reasoning as in the proof of Theorem 2.3. We use Corallary A.4 and consider the following cases for $k \in [K]$.

- if $\hat{\lambda}_k^{(1)} = 0$, and $\hat{\lambda}_k^{(2)} = 0$, we deduce that

$$\left| \sum_s s\hat{\mathbb{P}}_{X|S=s} \left( \forall j \neq k \ \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)}) \right) \right| \leq \varepsilon + \frac{2(K-1)}{N_{\min}}. \tag{14}$$

- if there exists $l \in \{1, 2\}$ such that $\hat{\lambda}_k^{(l)} \neq 0$, then $\hat{\mathfrak{h}}_k^l = 0$.

Let us now deal with the unfairness of $\hat{g}_\varepsilon$, recalled in (10). Bounding this quantity is a direct implication of the above lines. On the one hand, let $k \in [K]$ such that $\hat{\lambda}_k^{(1)} = 0$, and $\hat{\lambda}_k^{(2)} = 0$, then from Equation (14), we have

$$\begin{aligned} \left| \sum_{s \in \mathcal{S}} s\mathbb{P}_{X|S=s} \left( \hat{g}_\varepsilon(X, S) = k \right) \right| &= \left| \sum_{s \in \mathcal{S}} s\mathbb{P}_{X|S=s} \left( \forall j \neq k \ \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)}) \right) \right| \\ &\leq \left| \sum_{s \in \mathcal{S}} s \left( \mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s} \right) \left( \forall j \neq k \ \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)}) \right) \right| + \varepsilon + \frac{2(K-1)}{N_{\min}}. \end{aligned} \tag{15}$$

On the other hand, if for $k \in [K]$ there exists $l \in \{1, 2\}$ such that $\hat{\mathfrak{h}}_k^l = 0$ then in view of Equation (12), we also deduce that

$$\begin{aligned} &\left| \sum_{s \in \mathcal{S}} s\mathbb{P}_{X|S=s} \left( \hat{g}_\varepsilon(X, S) = k \right) \right| \\ &\qquad \leq \left| \sum_{s \in \mathcal{S}} s \left( \mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s} \right) \left( \forall j \neq k \ \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)}) \right) \right| + \varepsilon + \frac{2(K-1)}{N_{\min}}. \end{aligned}$$

Therefore, from the above inequalities, taking the maximum over $k \in [K]$, we deduce from Lemma C.2 (point 1.) that conditional on $\mathcal{D}_n$ and on $(S_1, \ldots, S_N)$,

$$\mathbb{E}\left[\mathcal{U}(\hat{g}_\varepsilon)\right] \leq \varepsilon + \left( \frac{KC_1}{\sqrt{N_{\min}}} + \frac{2(K-1)}{N_{\min}} \right) \mathbb{1}_{\{N_{\min} \geq 1\}} + \mathbb{E}\left[\mathcal{U}(\hat{g}_\varepsilon) \mathbb{1}_{\{N_{\min} = 0\}}\right] \leq \varepsilon + \frac{c_1 \mathbb{1}_{\{N_{\min} \geq 1\}}}{\sqrt{N_{\min}}} + C_K \mathbb{P}\left(N_{\min} = 0\right),$$

for some non negative constants $c_1$ and $C_K$ that depend on $K$. Now, we observe that

$$\mathbb{P}\left(N_{\min} = 0\right) = \mathbb{P}\left(N_1 = 0\right) + \mathbb{P}\left(N_{-1} = 0\right)) \leq \exp(\log(1 - \pi_1)N) + \exp(\log(1 - \pi_{-1})N).$$

Therefore, applying Lemma A.2, we deduce that

$$\mathbb{E}\left[\mathcal{U}(\hat{g}_\varepsilon)\right] \leq \frac{C}{\sqrt{N \min(\pi_{-1}, \pi_1)}}.$$

**Unfairness control in the case of exact fairness** Along this proof, we need to adjust the notation as in the case of the optimal rule, *c.f.* (8). As in Lemma C.1, we first introduce, for $s \in \mathcal{S}$ and $k \in [K]$,

$$\hat{h}_k^s : (x, \beta) \mapsto \hat{\pi}_s \bar{p}_k(x, s) - s\beta \ .$$

By construction, the estimator $\bar{p}_k(X, S)$ satisfies Assumption 2.2, therefore for all $s \in \mathcal{S}$ and $k \in [K]$

$$\mathbb{P}_{X|S=s}\left(\hat{g}(X, S) = k\right) = \mathbb{P}_{X|S=s}\left( \forall j \neq k, \hat{h}_k^s(X, \hat{\beta}_k) > \hat{h}_j^s(X, \hat{\beta}_j) \right) \ .$$

Considering the first order optimality conditions for $\hat{\beta}$, we can show that, for all $k \in [K]$ and $s \in \mathcal{S}$, there exists $\alpha_k^s \in [-1, 1]$ such that

$$\begin{aligned} s\hat{\mathbb{P}}_{X|S=s} \left( \forall j \neq k, \hat{h}_k^s(X, \hat{\beta}_k) > \hat{h}_j^s(X, \hat{\beta}_j) \right) + \\ \alpha_k^s \hat{\mathbb{P}}_{X|S=s} \left( \forall j \neq k, \hat{h}_k^s(X, \hat{\beta}_k) \geq \hat{h}_j^s(X, \hat{\beta}_j), \ \exists j \neq k, \hat{h}_k^s(X, \hat{\beta}_k) = \hat{h}_j^s(X, \hat{\beta}_j) \right) = 0 \ . \end{aligned}$$

From the above equation, we deduce that

$$\begin{aligned}\mathcal{U}(\hat{g}) = \max_{k=1\ldots,K} &\left|\mathbb{P}_{X|S=1}\left(\hat{g}(X,S)=k\right) - \mathbb{P}_{X|S=-1}\left(\hat{g}(X,S)=k\right)\right| \\ &\leq \max_{k=1,\ldots,K}\sum_{s\in\mathcal{S}}\left|\left(\mathbb{P}_{X|S=s}-\hat{\mathbb{P}}_{X|S=s}\right)\left(\forall j\neq k,\ \hat{h}_k^s(X,\hat{\beta}_k) > \hat{h}_j^s(X,\hat{\beta}_j)\right)\right| \\ &\qquad + \max_{k=1,\ldots,K}\sum_{s\in\mathcal{S}}\hat{\mathbb{P}}_{X|S=s}\left(\exists j\neq k,\ \hat{h}_k^s(X,\hat{\beta}_k) = \hat{h}_j^s(X,\hat{\beta}_j)\right)\ .\end{aligned}$$

Observe that for all $k\in[K]$

$$\begin{aligned}&\left|\left(\mathbb{P}_{X|S=s}-\hat{\mathbb{P}}_{X|S=s}\right)\left(\forall j\neq k,\ \hat{h}_k^s(X,\hat{\beta}_k) > \hat{h}_j^s(X,\hat{\beta}_j)\right)\right| = \\ &\qquad\left|\left(\mathbb{P}_{X|S=s}-\hat{\mathbb{P}}_{X|S=s}\right)\left(\forall j\neq k,\ \bar{p}_k(X,s)-\bar{p}_j(X,s) \geq \frac{s(\hat{\beta}_k-\hat{\beta}_j)}{\hat{\pi}_s}\right)\right| \\ &\qquad\qquad \leq \sum_{j=1}^{K}\sup_{t\in\mathbb{R}}\left|\left(\mathbb{P}_{X|S=s}-\hat{\mathbb{P}}_{X|S=s}\right)\left(\bar{p}_k(X,s)-\bar{p}_j(X,s)\geq t\right)\right|\ .\end{aligned}$$

Therefore, from the Dvoretzky-Kiefer-Wolfowitz Inequality conditional on $\mathcal{D}_n$ and on $(S_1,\ldots,S_N)$, we deduce that, for each $s\in\mathcal{S}$ and $k\in[K]$

$$\mathbb{E}\left[\left|\left(\mathbb{P}_{X|S=s}-\hat{\mathbb{P}}_{X|S=s}\right)\left(\forall j\neq k, \hat{h}_k^s(X,\hat{\beta}_k) > \hat{h}_j^s(X,\hat{\beta}_j)\right)\right|\right] \leq C\sqrt{\frac{1}{N_s}}\ .$$

Applying Lemma C.1, we then get that, conditional on $\mathcal{D}_n$ and on $(S_1,\ldots,S_N)$, we have that

$$\mathbb{E}\left[\mathcal{U}(\hat{g})\right] \leq C\sum_{s\in\mathcal{S}}\sqrt{\frac{1}{N_s}}\ ,$$

for some positive constant $C$ that depends in $K$. Since $N_s$ is a binomial random variable with parameters $N$ and $\pi_s$, we get

$$\mathbb{E}\left[\mathcal{U}(\hat{g})\right] \leq C\sqrt{\frac{1}{N}},$$

where $C$ depends on $K$ and $\min(\pi_{-1},\pi_1)$. □

*Proof of Theorem 3.3.* Let $0<\delta<1$ and let $k\in[K]$. From Equations (12), and (13) and using Corallary A.4, we deduce that if $\lambda_k^{(1)}\neq 0, \lambda_k^{(2)}\neq 0$, then

$$\begin{aligned}\sum_{s\in\mathcal{S}} s\hat{\mathbb{P}}_{X|S=s}\left(\forall j\neq k\ \hat{h}_k^s(X,\hat{\lambda}_k^{(1)},\hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X,\hat{\lambda}_j^{(1)},\hat{\lambda}_j^{(2)})\right) &\geq \varepsilon - \frac{2(K-1)}{N_{\min}} \\ \sum_{s\in\mathcal{S}} s\hat{\mathbb{P}}_{X|S=s}\left(\forall j\neq k\ \hat{h}_k^s(X,\hat{\lambda}_k^{(1)},\hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X,\hat{\lambda}_j^{(1)},\hat{\lambda}_j^{(2)})\right) &\leq -\varepsilon + \frac{2(K-1)}{N_{\min}}\ .\end{aligned}$$

Therefore, since $\dfrac{C_\delta}{\sqrt{N_{\min}}} \geq \dfrac{2(K-1)}{N_{\min}}$, we deduce that on $\mathcal{A}_{\min} = \left\{\varepsilon > \dfrac{C_\delta}{\sqrt{N_{\min}}}\right\}$

$$0 < \sum_{s\in\mathcal{S}} s\hat{\mathbb{P}}_{X|S=s}\left(\forall j\neq k\ \hat{h}_k^s(X,\hat{\lambda}_k^{(1)},\hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X,\hat{\lambda}_j^{(1)},\hat{\lambda}_j^{(2)})\right) < 0\ ,$$

which leads to a contradiction. Therefore, on the event $\mathcal{A}_{\min}$, we necessary have $\hat{\lambda}_k^{(1)}\hat{\lambda}_k^{(2)}=0$ and $\hat{\lambda}_k^{(1)}+\hat{\lambda}_k^{(2)}\geq 0$. Note that on the event $\mathcal{A}_{\min}$, we have $N_{\min}\geq 1$.

The remaining of the proof consists in dealing with the two sub-cases when $\hat{\lambda}_k^{(1)}\hat{\lambda}_k^{(2)}=0$ and $\hat{\lambda}_k^{(1)}+\hat{\lambda}_k^{(2)}\geq 0$. First, let us consider for $k\in[K]$, the case where $\hat{\lambda}_k^{(1)}\neq 0$, and $\hat{\lambda}_k^{(2)}=0$ (the case $\hat{\lambda}_k^{(1)}=0$, and $\hat{\lambda}_k^{(2)}\neq 0$ follows in the same way). We observe that since $\hat{\mathfrak{h}}_k^1=0$, on the event $\mathcal{A}_{\min}$

$$0 \leq \varepsilon - \frac{2(K-1)}{N_{\min}} \leq \sum_s s\hat{\mathbb{P}}_{X|S=s}\left(\forall j\neq k\ \hat{h}_k^s(X,\hat{\lambda}_k^{(1)},\hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X,\hat{\lambda}_j^{(1)},\hat{\lambda}_j^{(2)})\right) \leq \varepsilon + \frac{2(K-1)}{N_{\min}}\ . \quad (16)$$

Moreover, we have that for each $k \in [K]$ such that $\hat{\lambda}_k^{(1)} \neq 0$, and $\hat{\lambda}_k^{(2)} = 0$

$$\left| \left| \sum_s s \mathbb{P}_{X|S=s} \left( \hat{g}_\varepsilon(X,S) = k \right) \right| - \varepsilon \right| =$$
$$\left| \left| \sum_s s \mathbb{P}_{X|S=s} \left( \forall j \neq k \ \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)}) \right) \right| -$$
$$\left| \sum_s s \hat{\mathbb{P}}_{X|S=s} \left( \forall j \neq k \ \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)}) \right) \right| +$$
$$\left| \sum_s s \hat{\mathbb{P}}_{X|S=s} \left( \forall j \neq k \ \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)}) \right) \right| - \varepsilon \Bigg| \ .$$

Using first the triangle inequality and then the reverse triangle inequality, we get from Equation (16)

$$\left| \left| \sum_s s \mathbb{P}_{X|S=s} \left( \hat{g}_\varepsilon(X,S) = k \right) \right| - \varepsilon \right| \leq$$
$$\left| \sum_{s \in \mathcal{S}} s \left( \mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s} \right) \left( \forall j \neq k \ \hat{h}_k^s(X, \hat{\lambda}_k^{(1)}, \hat{\lambda}_k^{(2)}) > \hat{h}_j^s(X, \hat{\lambda}_j^{(1)}, \hat{\lambda}_j^{(2)}) \right) \right| + \frac{2(K-1)}{N_{\min}} \ , \quad (17)$$

which yields together with Lemma C.2 (point 2.),

$$\left| \left| \sum_s s \mathbb{P}_{X|S=s} \left( \hat{g}_\varepsilon(X,S) = k \right) \right| - \varepsilon \right| \leq \left( K \sqrt{\frac{2 \log(\frac{4K}{\delta})}{N_{\min}}} + \frac{2(K-1)}{N_{\min}} \right) \leq \frac{C_\delta}{\sqrt{N_{\min}}} \ . \quad (18)$$

Now, observe that for the second sub-case when $\hat{\lambda}_k^{(1)} = \hat{\lambda}_k^{(2)} = 0$, we get using Equation (15), applying again Lemma C.2 (point 2.) the following bound

$$\left| \sum_s s \mathbb{P}_{X|S=s} \left( \hat{g}_\varepsilon(X,S) = k \right) \right| \leq \varepsilon + \frac{C_\delta}{\sqrt{N_{\min}}} \ .$$

Combining these two bounds, we conclude that on the event $\mathcal{A}(\delta) = \mathcal{A}_{\min} \bigcap \left( \cap_{k \in [K]} \mathcal{A}_k(\delta) \right)$

$$\mathcal{U}(\hat{g}_\varepsilon) < \varepsilon + \frac{C_\delta}{\sqrt{N_{\min}}} \ ,$$

which concludes the main part of the proof. Let us now focus on the particular case where on $\mathcal{A}(\delta)$ we have

$$\mathcal{U}(\hat{g}_\varepsilon) < \varepsilon - \frac{C_\delta}{\sqrt{N_{\min}}} \ .$$

Then for all $k \in [K]$ we have

$$\left| \sum_s s \mathbb{P}_{X|S=s} \left( \hat{g}_\varepsilon(X,S) = k \right) \right| < \varepsilon - \frac{C_\delta}{\sqrt{N_{\min}}} \ .$$

Hence on the set $\mathcal{A}(\delta)$, we deduce that the case related to (18) is not possible and then for each $k$, we necessary have $\hat{\lambda}_k^{(1)} = \hat{\lambda}_k^{(2)} = 0$ and then

$$\hat{g} = \hat{g}_\varepsilon \ .$$

To conclude the proof, we observe that

$$\mathbb{P}\left(\mathcal{A}(\delta)^c\right) = \mathbb{P}\left(\mathcal{A}_{\min}^c\right) + \sum_{k=1}^K \mathbb{P}\left(\mathcal{A}_k^c(\delta)\right) \leq \mathbb{P}\left( N_1 \leq \frac{C_\delta^2}{\varepsilon^2} \right) + \mathbb{P}\left( N_- 1 \leq \frac{C_\delta^2}{\varepsilon^2} \right) + K\delta.$$

But, from Lemma A.1, we have for each $s \in \mathcal{S}$,

$$\mathbb{P}\left(N_s \leq \frac{C_\delta^2}{\varepsilon^2}\right) \leq \exp\left(-2N\left(\pi_s - \frac{C_\delta^2}{\varepsilon^2 N}\right)^2\right) \leq \exp\left(-N\frac{\pi_s^2}{2}\right) \leq \delta \ ,$$

provided that $\pi_s > \dfrac{2C_\delta^2}{N\varepsilon^2}$, and $N \geq 2\frac{\log(1/\delta)}{\pi_{\min}^2}$. Since $\varepsilon > \frac{\sqrt{2}C_\delta}{\sqrt{\pi_{\min}N}}$, and $N \geq 2\frac{\log(1/\delta)}{\pi_{\min}^2}$, the latter conditions are satisfied. Therefore, we deduce that

$$\mathbb{P}\left(\mathcal{A}(\delta)^c\right) \leq (K+2)\delta \ .$$

□

*Proof of Theorem 3.4.* We only consider the case $\varepsilon > 0$. The proof in the case of exact fairness relies on similar arguments and then it is omitted. To ease the notation, we write $\hat{g}$ instead of $\hat{g}_\varepsilon$.

The proof goes conditional on the training data. First, let us decompose the *excess fair-risk* of the classifier $\hat{g}$ in a convenient way for our analysis

$$\begin{aligned}\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(\hat{g}) - \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g^*_{\varepsilon-\text{fair}}) &= \left(\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(\hat{g}) - \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g})\right) \\ &\quad + \left(\mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g}) - \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}\right)\right) \ . \qquad (19)\end{aligned}$$

According to the first term, we have

$$\begin{aligned}\left(\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(\hat{g}) - \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g})\right) &= \sum_{k=1}^K \left(\lambda_k^{*(1)} - \hat{\lambda}_k^{(1)}\right)\left[\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\hat{g}(X,S)=k\right) - \varepsilon\right] \\ &\quad + \sum_{k=1}^K \left(\lambda_k^{*(2)} - \hat{\lambda}_k^{(2)}\right)\left[-\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\hat{g}(X,S)=k\right) - \varepsilon\right].\end{aligned}$$

Let $\delta = 1/N$. If $\varepsilon \leq \frac{\sqrt{2}C_\delta}{\sqrt{\pi_{\min}N}}$, since parameters $\lambda_k^{*(l)}$ and $\hat{\lambda}_k^{(l)}$ are bounded, we deduce from the above equation and Theorem 3.2 that

$$\mathbb{E}\left[\left(\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(\hat{g}) - \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g})\right)\right] \leq C\min\left(\varepsilon, \frac{\sqrt{2}C_\delta}{\sqrt{\pi_{\min}N}}\right) + \frac{C}{\sqrt{N}} \leq C\frac{\log(N)}{\sqrt{N}} \ , \qquad (20)$$

where $C > 0$ is a constant which depends on $\pi_{\min}$ and $K$. If $\varepsilon > \frac{\sqrt{2}C_\delta}{\sqrt{\pi_{\min}N}}$, we apply Theorem 3.3. We have on the event $\mathcal{A}(1/N)$ that

- either $\hat{\lambda}_k^{(1)} = 0$, and then since $\lambda_k^{*(1)} > 0$ is bounded

$$\left(\lambda_k^{*(1)} - \hat{\lambda}_k^{(1)}\right)\left[\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\hat{g}(X,S)=k\right) - \varepsilon\right] \leq C\left(\mathcal{U}(\hat{g}) - \varepsilon\right)) \leq C\frac{\log(N)}{\sqrt{N_{\min}}} \ .$$

- or $\hat{\lambda}_k^{(1)} > 0$, in this case on $\mathcal{A}(1/N)$, $\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\hat{g}(X,S)=k\right) > 0$. From Equation (18) in the proof of Theorem 3.3, we deduce

$$\left(\lambda_k^{*(1)} - \hat{\lambda}_k^{(1)}\right)\left[\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\hat{g}(X,S)=k\right) - \varepsilon\right] \leq C\frac{\log(N)}{\sqrt{N_{\min}}} \ .$$

Since on $\mathcal{A}(1/N)$, $N_{\min} \geq 1$, we deduce that if $\varepsilon > \frac{\sqrt{2}C_\delta}{\sqrt{\pi_{\min}N}}$

$$\mathbb{E}\left[\sum_{k=1}^K \left(\lambda_k^{*(1)} - \hat{\lambda}_k^{(1)}\right)\left[\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\hat{g}(X,S)=k\right) - \varepsilon\right]\right] \leq C\left(\mathbb{E}\left[\frac{\log(N)\mathbb{1}\{N_{\min}\geq 1\}}{\sqrt{N_{\min}}}\right] + \mathbb{P}\left(\mathcal{A}(1/N)^c\right)\right).$$

According to Lemma C.2 we have $\mathbb{P}\left(\mathcal{A}(1/N)^c\right) \leq \dfrac{K+2}{N}$. Then we deduce form Lemma A.2 that

$$\mathbb{E}\left[\sum_{k=1}^{K}\left(\lambda_k^{*(1)} - \hat{\lambda}_k^{(1)}\right)\left[\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\hat{g}(X,S)=k\right) - \varepsilon\right]\right] \leq C\frac{\log(N)}{\sqrt{N}}.$$

Similar reasoning leads to

$$\mathbb{E}\left[\sum_{k=1}^{K}\left(\lambda_k^{*(2)} - \hat{\lambda}_k^{(2)}\right)\left[-\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\hat{g}(X,S)=k\right) - \varepsilon\right]\right] \leq C\frac{\log(N)}{\sqrt{N}}.$$

Combining the two above inequalities and Equation (20), we obtain for $\varepsilon > 0$ and $N$ large enough

$$\mathbb{E}\left[\left(\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}\left(\hat{g}\right) - \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g})\right)\right] \leq C\frac{\log(N)}{\sqrt{N}}. \tag{21}$$

Then we have shown that the first term in the r.h.s. of Eq. (19) relies on the unfairness of the classifier $\hat{g}$. Now, let us consider the second term in r.h.s. of Equation (19). Our goal will be to show that this term mainly depends on the quality of the base estimators $\hat{p}_k$. Since $\left(\lambda^{*(1)},\lambda^{*(2)}\right)$ is a maximizer of $\mathcal{R}_{(\lambda^{(1)},\lambda^{(2)})}(g^*_{(\lambda^{(1)},\lambda^{(2)})})$ over $(\lambda^{(1)},\lambda^{(2)})$, it is clear that, conditional on the data, $\mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g^*_{\lambda^{*(1)},\lambda^{*(2)}}) \geq \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}})$. (The parameter $\hat{\lambda}$ is seen as fixed conditional on the data.) Therefore, we have

$$\mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g}) - \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}\right) \leq \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g}) - \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}) \ .$$

By introducing $\hat{g}^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}$, we remove the estimation of $\lambda^{*(1)},\lambda^{*(2)}$ from the study of $\mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g}) - \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}\right)$. At this point, it becomes clear that bounding this term does not relies on the unlabeled sample sizes $N_s$. Let us recall the definition of $g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}$: conditional on the data

$$g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}} \in \arg\min_{g\in\mathcal{G}} \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(g) \ .$$

Then using similar arguments as those leading to Eq. (6) implies that

$$g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(x,s) \in \arg\max_{k\in[K]}\left(\pi_s p_k(x,s) - s(\hat{\lambda}_k^{(1)} - \hat{\lambda}_k^{(2)}\right) \ .$$

As a consequence, using the writing of the fair-risk provided by Lemma B.1

$$\begin{aligned}
&\mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g}) - \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}) = \\
&\sum_{s\in\mathcal{S}}\mathbb{E}_{X|S=s}\left[\max_{k\in[K]}\left(\pi_s p_k(X,s) - s(\hat{\lambda}_k^{(1)} - \hat{\lambda}_k^{(2)})\right) - \sum_{k=1}^{K}\left(\pi_s p_k(X,s) - s(\hat{\lambda}_k^{(1)} - \hat{\lambda}_k^{(2)})\right)\mathbb{1}_{\{\hat{g}(X,s)=k\}}\right] \ .
\end{aligned} \tag{22}$$

Because of the indicator function, there is only one non-zero element in the inner sum. Then we observe that for each $s \in \mathcal{S}$

$$\begin{aligned}
&\left|\max_{k\in[K]}\left(\pi_s p_k(X,s) - s(\hat{\lambda}_k^{(1)} - \hat{\lambda}_k^{(2)})\right) - \sum_{k=1}^{K}\left(\pi_s p_k(X,S) - s(\hat{\lambda}_k^{(1)} - \hat{\lambda}_k^{(2)})\right)\mathbb{1}_{\{\hat{g}(X,s)=k\}}\right| \\
&\qquad \leq 2\max_{k\in[K]}\left|(\pi_s p_k(X,s) - s(\hat{\lambda}_k^{(1)} - \hat{\lambda}_k^{(2)})) - (\hat{\pi}_s\bar{p}_k(X,s) - s(\hat{\lambda}_k^{(1)} - \hat{\lambda}_k^{(2)}))\right| \\
&\qquad \leq 2\left(\max_{k\in[K]}|p_k(X,s) - \bar{p}_k(X,s)| + |\pi_s - \hat{\pi}_s|\right) \ ,
\end{aligned}$$

where the last inequality is due to the fact that $\pi_s, \hat{\pi}_s, p_k$, and $\bar{p}_k$ are all in $[0,1]$. Therefore, recalling that $\bar{p}_k$ is a randomized version of $\hat{p}_k$ we can write

$$\mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g}) - \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}) \leq C\left(\|\hat{\mathbf{p}} - \mathbf{p}\|_1 + \sum_{s\in\mathcal{S}} |\hat{\pi}_s - \pi_s| + u\right) \ ,$$

and obtain the bound

$$\mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g}) - \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}\left(g^*_{\lambda^{*(1)},\lambda^{*(2)}}\right) \leq C\left(\|\hat{\mathbf{p}} - \mathbf{p}\|_1 + \sum_{s\in\mathcal{S}} |\hat{\pi}_s - \pi_s| + u\right) \ .$$

In view of Equation (20), the above inequality together with Equation (21) yield the desired result. □

*Proof of Theorem 3.8.* Let us remind the reader that for each $k \in [K]$, and $s \in \mathcal{S}$

$$h^s_k(X, \hat{\lambda}^{(1)}_k, \hat{\lambda}^{(2)}_k) := \left(\pi_s p_k(X,s) - s(\hat{\lambda}^{(1)}_k - \hat{\lambda}^{(2)}_k)\right).$$

We start the proof with Equation (22),

$$\mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g}) - \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}) = \sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_{k\in[K]} h^s_k(X, \hat{\lambda}^{(1)}_k, \hat{\lambda}^{(2)}_k)) - \sum_{k=1}^{K} h^s_k(X, \hat{\lambda}^{(1)}_k, \hat{\lambda}^{(2)}_k)\mathbb{1}_{\{\hat{g}(X,S)=k\}}\right] \ .$$

Furthermore, we have that

$$g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(X,s) \in \arg\max_{k\in[K]} h^s_k\left(X, \hat{\lambda}^{(1)}_k, \hat{\lambda}^{(2)}_k\right) \ .$$

Therefore, we observe that

$$\max_{k\in[K]} h^s_k(X, \hat{\lambda}^{(1)}_k, \hat{\lambda}^{(2)}_k)) - \sum_{k=1}^{K} h^s_k(X, \hat{\lambda}^{(1)}_k, \hat{\lambda}^{(2)}_k)\mathbb{1}_{\{\hat{g}(X,S)=k\}} = \sum_{i=1,k\neq i}^{K} \left|h^s_i(X, \hat{\lambda}^{(1)}_i, \hat{\lambda}^{(2)}_i) - h^s_k(X, \hat{\lambda}^{(1)}_k, \hat{\lambda}^{(2)}_k)\right| \mathbb{1}_{\{g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(X,s)=i\}}\mathbb{1}_{\{\hat{g}(X,s)=k\}}. \quad (23)$$

Moreover, for $k \neq i$ on the event $\left\{g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(X,s) = i, \hat{g}(X,s) = k\right\}$, we have from Equation (22)

$$\left|h^s_i(X, \hat{\lambda}^{(1)}_i, \hat{\lambda}^{(2)}_i) - h^s_k(X, \hat{\lambda}^{(1)}_k, \hat{\lambda}^{(2)}_k)\right| \leq 2\max_{s\in\mathcal{S}}\left(\max_{k\in[K]}\sup_x |p_k(x,s) - \bar{p}_k(x,s)| + |\pi_s - \hat{\pi}_s|\right). \quad (24)$$

Now, we observe that from Assumption 3.7, conditional on the data for each $s \in \mathcal{S}$

$$\mathbb{P}_{X|S=s}\left(\left|h^s_i(X, \hat{\lambda}^{(1)}_i, \hat{\lambda}^{(2)}_i) - h^s_k(X, \hat{\lambda}^{(1)}_k, \hat{\lambda}^{(2)}_k)\right| \leq 2\left(\max_{k\in[K]}\sup_x |p_k(x,s) - \bar{p}_k(x,s)| + |\pi_s - \hat{\pi}_s|\right)\right) \leq C\left(\max_{k\in[K]}\sup_x |p_k(x,s) - \bar{p}_k(x,s)| + |\pi_s - \hat{\pi}_s|\right).$$

Combining the above inequality with Equation (22), Equation (23), and Equation (24), we obtain that

$$\mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(\hat{g}) - \mathcal{R}_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}(g^*_{\hat{\lambda}^{(1)},\hat{\lambda}^{(2)}}) \leq C\sum_{s\in\mathcal{S}}\left(\max_{k\in[K]}\sup_x |p_k(x,s) - \bar{p}_k(x,s)| + |\pi_s - \hat{\pi}_s|\right)^2 \leq C\left(\|\hat{\mathbf{p}} - \mathbf{p}\|^2_\infty + u^2 + \sum_{s\in\mathcal{S}} |\hat{\pi}_s - \pi_s|^2\right).$$

Finally, we deduce again the desired result from the above inequality, Equation (20), and Equation (21). □

# D Additional numerical experiments

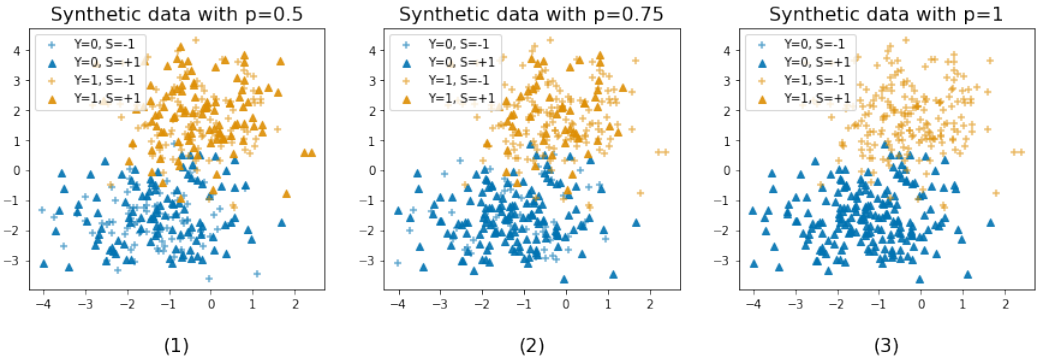


Figure 9: Example of synthetic data in binary case where $d = 2$ and $m = 1$. The level of unfairness is set as follows: *(1)* $p = 0.5$ (no unfairness); *(2)* $p = 0.75$ (unfair dataset); *(3)* $p = 1$ (highly unfair dataset).