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]Université Gustave Eiffel, [2]ENSAE-CREST

September 27, 2021

Abstract

Algorithmic Fairness is an established area of machine learning, willing to reduce the influence of biases in the data. Yet, despite its wide range of applications, very few works consider the multi-class classification setting from the fairness perspective. We address this question by extending the definition of *Demographic Parity* to the multi-class problem while specifying the corresponding expression of the optimal fair classifier. This suggests a plug-in data-driven procedure, for which we establish theoretical guarantees. Specifically, we show that the enhanced estimator mimics the behavior of the optimal rule, both in terms of fairness and risk. Notably, fairness guarantee is distribution-free. We illustrate numerically the quality of our algorithm. The procedure reveals to be much more suitable than an alternative approach enforcing fairness constraints on the score associated to each class. This shows that our method is empirically very effective in fair decision making on both synthetic and real datasets.

1 Introduction

Algorithmic fairness has become very popular during the last decade [1–4, 6, 7, 12, 17, 24, 26] because it helps addressing an important social problem: mitigating historical bias contained in the data. This is a crucial issue in many applications such as loan assessment, health care, or even criminal sentencing. The common objective in algorithmic fairness is to reduce the influence of a sensitive attribute on a prediction. Several notions of fairness have already been considered in the literature for the binary classification problem [3, 25]. All of them impose some independence condition between the sensitive feature and the prediction. However, in some applications such as loan agreements, this independence is desired on some or all values of the label space, *e.g., Equality of odds* or *Equal opportunity* [15]. In this paper, we focus on the well established *Demographic Parity* (DP) [4] that requires the independence between the sensitive feature and the prediction function, while not relying on labels. DP has a recognized interest in various applications; this constraint could be typically imposed in loan agreement without gender attributes or in the context of crime prediction without ethnicity discrimination.

All the previously mentioned references focus either on the regression or the binary classification framework. However, many (modern) applications fall within the scope of multi-class classification, *e.g.,* image recognition, text categorization. Moreover, most real world applications can be tackled from the multi-class perspective. For instance, criminal recidivism is often treated as a binary

problem, while it may be more suitable to distinguish between the different stratum of the problem and provide thinner description of the criminal behavior. However, extension of previous works to the multi-class setting is tricky, in particular since the adequate notion of fairness in this framework is not clearly defined and should be handled with caution.

Up to our knowledge, imposing fairness constraint in the multi-class problem has only been briefly discussed in [23], while focusing on Support Vector Machine (SVM) fair prediction. However, their fairness approach relies on properly selecting a subset of the data that is unbiased from the fairness perspective, and hereby differs significantly from ours that aims at enforcing fairness using the dataset as a whole. Besides, from a high-level perspective, the procedure described in [23] chooses to impose 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. However, since the decision rule in the multi-class setting relies on the maximizer over scores, we do not adopt this quite unnatural approach and rather choose to impose fairness directly on the maximizer itself. Our main contributions are three folds:

- We provide an optimal solution for the multi-class classifier problem under the DP constraint;
- We build a data-driven procedure that mimics the performance of the optimal rule both in terms of risk and fairness. Notably, our fairness guarantee is *distribution-free*;
- We illustrate numerically the robustness of our approach on synthetic datasets with various bias levels, as well as on several real datasets. In particular, our approach outperforms those strategies that impose independence on *each score* separately.

Related works. There are mainly three ways to build fair prediction: i) *pre-processing* methods mitigate bias in the data before applying classical Machine Learning algorithms; *in-processing* methods reduce bias during training; and iii) *pro-processing* methods enforce fairness after fitting. The present work falls within the last category of methods. In closely related study [7], the authors derive an expression for fair binary classifiers under *Equal Opportunity* constraint. In contrast, we focus on the multi-class setting, while dealing with the *Demographic Parity* constraints.

Another line of works deals with algorithmic fairness from an optimal transport perspective [6, 9, 13, 14]. There, the authors build a fair prediction based on Wasserstein barycenters of conditional distributions with respect to the sensitive feature. The argumentation extends with little effort to a multi-class setting, even though the proper fairness definition in this context remains a question to investigate. This approach provides a fair classifier based on emprirical risk minimization (ERM) with fairness constraints on each underlying score. Our numerical conclusions show below that the latter is outperformed by our approach requiring fairness on solely the maximizer.

Up to our knowledge, fairness in the multi-class setting has not been considered in earlier works, except [23]. There, the authors build an SVM-type classifier under DP constraint. However their strategy differs significantly from ours, since an in-processing method enforces fairness by sub-sample selection. In contrast, we keep the whole sample and enforce fairness in a post-processing manner.

2 Fair multi-class classification

2.1 Statistical setting

Let (X, S, Y) be a random tuple with distribution $\mathbb{P}$, where $X \in \mathcal{X}$ a subset of $\mathbb{R}^d$, $S \in \mathcal{S} := \{-1, 1\}$, 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 $\min_{s\in\mathcal{S}} \pi_s > 0$. A classification rule g is a function mapping $\mathcal{X} \times \{-1, 1\}$ onto $[K]$, whose performance is evaluated through the misclassification risk

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

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

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

2.2 Multi-class classification with demographic parity

In the present study, we consider muti-class classification problems under DP fairness constraint [4], that requires the independence of the prediction function from the sensitive feature S.

Definition 2.1 (Demographic parity)**.** *We say that a classifier $g \in \mathcal{G}$ (denoted $g \in \mathcal{G}_{\text{fair}}$) with respect to the distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times [K]$ if*

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

The above definition naturally extends to the multi-class setting the DP considered in binary classification [2, 6, 13, 16, 20] . Intuitively, when fairness is required, two important aspects of a classifier need to be controlled: the misclassification risk $\mathcal{R}(\cdot)$ and the unfairness, evaluated as follows.

Definition 2.2 (Unfairness)**.** *The unfairness of a classifier $g \in \mathcal{G}$ is quantified by*

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

Naturally, taking into account the definition above, a classifier g is fair if and only if $\mathcal{U}(g) = 0$.

An alternative definition could consider the maximal unfairness over labels (simply replacing the summation by a maximum over k in the above definition). However, summing over all possible labels is more informative and appears more naturally when controlling the prediction risk (see Thm. 3.2).

2.3 Optimal fair classifier

In this section, we provide an explicit formulation of the optimal fair classifiers *w.r.t.* the misclassification risk under DP constraint. An optimal fair classifier is a solution of

$$\min_{g\in\mathcal{G}_{\text{fair}}} \mathcal{R}(g) \ .$$

The difficulty of obtaining an optimal fair classifier consists in properly balancing the misclassification risk together with the fairness criterion. Hence, we introduce for $\lambda = (\lambda_1, \ldots, \lambda_K) \in \mathcal{R}^K$,

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

We call this measure *fair-risk*. In order to be able to derive a characterization of the optimal fair classifier, we require the following technical assumption on the random variables $p_k(X, s)$.

Assumption 2.3 (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* $s \in \mathcal{S}$*.*

This assumption requires 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 of g^*_{fair}. It may appear that this assumption is unusual, however we observe that in the binary case ($K = 2$), the above assumption simply boils down to the one considered in [7] that requires the continuity of $t \mapsto \mathbb{P}\left(p_k(X,S) \leq t | S = s\right)$. It is, however, clear that in the general case $K \geq 3$ these two conditions describe different sets of distributions. Hence, Assumption 2.3 appears as a condition tailored for the multi-class problem.

We are now in a position to provide a characterization of optimal fair classification.

Proposition 2.4. *Let Assumption 2.3 be satisfied and define* $\lambda^* \in \mathbb{R}^K$ *by*

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

Then, $g^*_{\text{fair}} \in \arg\min_{g \in \mathcal{G}_{\text{fair}}} \mathcal{R}(g)$ *if and only if* $g^*_{\text{fair}} \in \arg\min_{g \in \mathcal{G}} \mathcal{R}_{\lambda^*}(g)$*.*

In other words, the optimum of the risk $\mathcal{R}(g)$ over the class of fair classifiers is also maximizing the fair-risk $\mathcal{R}_{\lambda^*}$. By construction, $\mathcal{R}_{\lambda^*}$ is a risk measure which optimally balances both classification accuracy and unfairness. Proposition 2.4 directly implies that $\mathcal{R}_{\lambda^*}(g) \geq \mathcal{R}_{\lambda^*}(g^*_{\text{fair}}) = \mathcal{R}(g^*_{\text{fair}}) \geq 0$, for $g \in \mathcal{G}$. We now quantify the performance of any classifier $g \in \mathcal{G}$ through its *excess fair-risk*

$$\mathcal{E}_{\text{fair}}(g) := \mathcal{R}_{\lambda^*}(g) - \mathcal{R}_{\lambda^*}(g^*_{\text{fair}}) \ .$$

Furthermore, Prop. 2.4 directly implies a closed form expression of optimal fair classifiers, which is the bedrock of our procedure. Any optimal fair classifier is simply maximizing scores, which are obtained by shifting the original conditional probabilities in a optimal manner.

Corollary 2.5. *Under Assumption 2.3, an optimal fair classifier is characterized by*

$$g^*_{\text{fair}}(x,s) \in \arg\max_k \left(\pi_s p_k(x,s) - s\lambda^*_k\right) \quad , \quad (x,s) \in \mathcal{X} \times \mathcal{S} \ .$$

3 Data-driven procedure with statistical guarantees

While the previous section focused on optimal fairness, we provide now a plug in estimator for the optimal fair classifier g^*_{fair}. We propose an algorithm that enjoys strong theoretical guarantees both in terms of fairness and risk. In particular, we exhibit in Section 3.2 distribution-free fairness guarantee.

3.1 Plug-in estimator

We are given two datasets. The first *labeled* one $\mathcal{D}_n = \{(X_i, S_i, Y_i),\ i = 1, \ldots, n\}$ consists of *i.i.d.* samples from the distribution $\mathbb{P}$. This is the classical dataset used for training estimators $(\hat{p}_k)_k$ of the conditional probabilities $(p_k)_k$, *e.g.,* Random Forest, SVM, *etc.* The second *unlabeled* dataset $\mathcal{D}'_N$ consists of N *i.i.d.* copies of (X,S). the sample $\mathcal{D}'_N$ is collected and split in the following way: we set $(S_1, \ldots, S_N)$ the *i.i.d.* sample of sensitive features used to compute empirical frequencies $(\hat{\pi}_s)_{s \in \mathcal{S}}$ for estimating $(\pi_s))_{s \in \mathcal{S}}$. The number of observations corresponding to $S = s$ is denoted N_s, for $s \in \mathcal{S}$.

Of course $N_{-1} + N_1 = N$. For $s \in \mathcal{S}$, the feature vector in $\mathcal{D}'_N$ denoted $X_1^s, \ldots, X_{N_s}^s$ is composed by *i.i.d.* data from $\mathbb{P}_{X^s}$, the distribution of $X|S = s$. All samples are assumed independent.

In order to derive consistency results on the excess fair-risk and the unfairness of our plug-in rule, we require continuity conditions on the random variables $\hat{p}_k(X, S)$, in the spirit of Assumption 2.3 (conditional on the learning sample). However, such property is automatically satisfied whenever perturbing the $(\hat{p}_k)_k$ with a small random noise. To this end, we introduce $\bar{p}_k(X, S, \zeta_k) := \hat{p}_k(X, S) + \zeta_k$, for a given uniform perturbation ζ_k on $[0, u]$.

Given $(\zeta_k)_{k\in[K]}$ and $(\zeta_{k,i}^s)$ independent copies of a Uniform distribution on $[0, u]$, we define the randomized fair classifier $\hat{g}$ by plug-in as

$$\hat{g}(x, s) = \arg\max_{k\in[K]} \left(\hat{\pi}_s \bar{p}_k(x, s, \zeta_k) - s\hat{\lambda}_k \right) , \quad \text{for all } (x, s) \in \mathcal{X} \times \mathcal{S} \ , \tag{2}$$

with $\hat{\lambda} \in \mathbb{R}^K$ given as

$$\hat{\lambda} \in \arg\min_{\lambda} \sum_{s\in\mathcal{S}} \frac{1}{N_s} \sum_{i=1}^{N_s} \left[\max_{k\in[K]} \left(\hat{\pi}_s \bar{p}_k(X_i^s, s, \zeta_{k,i}^s) - s\lambda_k \right) \right] \ . \tag{3}$$

Note that the construction of the plug-in rule $\hat{g}$ relies on (x, s) but also on the perturbations ζ and $\zeta_{k,i}^s$ for $k \in [K]$, $i \in N_s$ and $s \in \mathcal{S}$.

3.2 Statistical guarantees

We are now in position to derive fairness and consistency guarantees of our plug-in procedure.

Universal fairness guarantee. We first focus on fairness assessment and prove that the plug-in estimator $\hat{g}$ is asymptotically fair. The convergence rate on 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 3.1. *For any distribution $\mathbb{P}$, there exists a constant $C > 0$ which only depends on K and $\min_{s\in\mathcal{S}} \pi_s$, such that for any estimators $\hat{p}_k$ we have,*

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

Consistency of the excess fair-risk. We now consider the misclassification risk of the estimator $\hat{g}$. We define the L_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 $\mathbf{p} := (p_1, \ldots, p_K)$ as $\|\hat{\mathbf{p}} - \mathbf{p}\|_1 = \sum_{k\in[K]} |\hat{p}_k(X, S) - p_k(X, S)|$.

Theorem 3.2. *Let Assumption 2.3 be satisfied, then the following holds*

$$\mathbb{E}\left[\mathcal{E}_{\text{fair}}(\hat{g})\right] \leq C \left(\mathbb{E}\left[\|\hat{\mathbf{p}} - \mathbf{p}\|_1\right] + \sum_{s\in\mathcal{S}} \mathbb{E}\left[|\hat{\pi}_s - \pi_s|\right] + \mathbb{E}\left[\mathcal{U}(\hat{g})\right] + u \right) \ .$$

The above result highlights that the excess fair-risk of $\hat{g}$ depends on 1) the quality of the estimators of the conditional probabilities through its L_1-risk; 2) the quality of the estimators of $(\pi_s)_{s\in\mathcal{S}}$; 3) the unfairness of the classifier; and 4) the upper-bound u on the regularizing perturbations. Consequently, $\hat{g}$ is consistent *w.r.t.* the excess-fair risk as soon as the estimator $\hat{\mathbf{p}}$ is consistent in L_1-norm.

Corollary 3.3. *If* $\mathbb{E}\left[\|\hat{\mathbf{p}} - \mathbf{p}\|_1\right] \to 0$ *and* $u = u_n \to 0$ *when* $n \to \infty$*, we have*

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

We emphasize that Theorem 3.1 and Corollary 3.3 directly imply that $\hat{g}$ performs asymptotically as well as g^*_{fair} in terms of both fairness and accuracy.

3.3 ERM estimation of $\hat{\mathbf{p}}$

We have established theoretical guarantees on risk and fairness for our plug-in procedure, when using any off-the-shelf consistent conditional probability estimator. We now study more closely the classical setting where the conditional probabilities estimation step is provided by ERM.

For a given (measurable) score function $\mathbf{f}(\cdot) = (f_1(\cdot), \ldots, f_K(\cdot))$ mapping $\mathcal{X} \times \{-1, 1\}$ onto $\mathbb{R}^K$, we define the induced classification rule[1]

$$g_{\mathbf{f}}(\cdot) \in \arg\max_{k \in [K]} f_k(\cdot) \ .$$

For the sake of simplicity, we focus on the L_2-risk[2]

$$R_2(\mathbf{f}) := \mathbb{E}\left[\sum_{k=1}^{K} \left(Z_k - f_k(X, S)\right)^2\right] \ , \quad \text{where} \ \ Z_k := 2\mathbb{1}_{\{Y=k\}} - 1.$$

The optimal score function $\mathbf{f}^*$ *w.r.t.* R_2 is given by

$$\mathbf{f}^* := \arg\min_{\mathbf{f}} R_2(\mathbf{f}) \ ,$$

where the infimum is taken over all measurable functions that map $\mathcal{X} \times \{-1, 1\}$ onto $\mathbb{R}^K$. The optimum $\mathbf{f}^*$ satisfies the relation $f_k^*(X, S) = 2p_k(X, S) - 1$. Zhang's Lemma [27] implies that $\mathbb{E}\left[\mathcal{R}(g_{\mathbf{f}}) - \mathcal{R}(g^*)\right] \leq \left(\mathbb{E}\left[R_2(\mathbf{f}) - R_2(\mathbf{f}^*)\right]\right)^{1/2}$, for any score function $\mathbf{f}$. This inequality highlights the connection between the usual misclassification risk of $g_{\mathbf{f}}$ and the L_2-risk of the score function $\mathbf{f}$.

In addition, the empirical counterpart of the L_2-risk R_2 is given for any function $\mathbf{f}$ by

$$\hat{R}_2(\mathbf{f}) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(Z_k^i - f_k(X_i, S_i)\right)^2, \quad \text{with} \ \ Z_k^i := 2\mathbb{1}_{\{Y_i=k\}} - 1 \, .$$

The empirical risk minimizer $\hat{f}$ over a given convex set $\mathcal{F}$ of functions is given by

$$\hat{\mathbf{f}} \in \arg\min_{\mathbf{f} \in \mathcal{F}} \hat{R}_2(\mathbf{f}) \ .$$

In view of the expression for the optimal score function $\mathbf{f}^*$, we naturally set $\hat{p}_k := \frac{\hat{f}_k + 1}{2}$ as an estimator of the conditional probability p_k. Given $\hat{p}$, our plug-in procedure given in Eq. (2) provides a fair classifier that we denote by $g_{\hat{\mathbf{f}}}$. According to the theoretical study conducted in the previous section, Theorem 3.1 ensures that the classifier $g_{\hat{\mathbf{f}}}$ is asymptotically fair.

Lastly, we invoke Eq. (2) in order to provide a fair classifier that we denote by $g_{\hat{\mathbf{f}}}$. These steps allow us to use the theoretical results established in the previous section. In particular, Theorem 3.1

[1] Whenever the maximum is reached at multiple indices, we set by convention $g_{\mathbf{f}}$ as the smallest index in $[K]$.

[2] Since the $0-1$ loss lacks convexity, we consider the square loss as a convex surrogate to avoid computational issues.

states that the classifier $g_{\hat{\mathbf{f}}}$ is asymptotically fair. Furthermore, the convergence of the excess fair-risk of $g_{\hat{\mathbf{f}}}$ to zero follows immediately from the consistency of the empirical risk minimizer $\hat{\mathbf{f}}$ *w.r.t.* to the L_2-risk. (due to Zhang's Lemma.)

We conclude this section by specifying rates of convergence for the excess fair-risk under additional assumptions on the class of functions $\mathcal{F}$.

Assumption 3.4. *The set $\mathcal{F}$ satisfies the following:*

1. *There exists $B > 0$ s.t. $\|\mathbf{f}\|_\infty := \max_{k\in[K]} \sup_{x\in\mathcal{X}} |f_k(x)| \leq B$, for each $\mathbf{f} \in \mathcal{F}$;*
2. *For $\varepsilon > 0$, there exists an ε-net $\mathcal{F}_\varepsilon \subset \mathcal{F}$ w.r.t. $\|.\|_\infty$ s.t. $\log(|\mathcal{F}_\varepsilon|) \leq C_{\mathcal{F}} \log(\varepsilon^{-1})$.*

Note that Assumption 3.4 covers classical parametric classes among others. This structural assumption on the set of models $\mathcal{F}$ enables to control the rate of convergence of the excess fair-risk.

Proposition 3.5. *Assume that $\mathbf{f}^* \in \mathcal{F}$ and $u = u_n = \frac{1}{n}$. If Assumptions 2.3 and 3.4 hold, then*

$$\mathbb{E}\left[\mathcal{E}_{\text{fair}}\left(g_{\hat{\mathbf{f}}}\right)\right] \leq C\left(\left(\frac{\log(n)}{n}\right)^{1/2} + N^{-1/2}\right) \ .$$

Under classical assumptions on the complexity of $\mathcal{F}$, Proposition 3.5 induces in particular a parametric rate of convergence for the excess fair-risk of $g_{\hat{f}}$.

4 Implementation of the algorithm

The fair classifier $\hat{g}$ is given by (2)-(3). The main steps of its implementation are described by the pseudo-code provided in Algorithm 1. In this section we briefly discuss two aspects of this algorithm.

Algorithm 1 Plug-in fair classifier

Input: new data point (x, s), base estimators $(\bar{p}_k)_k$, unlabeled sample $\mathcal{D}'_N$, $(\zeta_k)_k$ and $(\zeta^s_{k,i})_{k,i,s}$ collection of i.i.d uniform perturbations 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 $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_K)$ as a solution of Eq. (3);

`Acceleration scheme of Section 4 can be used for this step.`

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

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

First of all, base estimators $(\bar{p}_k)_k$ are needed as input of the algorithm. For this step we can fit any off-the-shelf estimators by using the labeled dataset $\mathcal{D}_n$. In our numerical study, we consider Random Forest (RF), SVM, logistic regression.

Once we have computed the $(\bar{p}_k)_k$, the fair classifier $\hat{g}$ relies on the estimator $\hat{\lambda}$ computed in **Step 2.** of the algorithm. It requires solving the minimization problem given by Eq. (3). The corresponding objective function is convex but non-smooth due to the evaluation of the function (hard) *max*. One classical way to regularize the objective function is to replace hard-max by a soft-max (also known as LogSumExp). Namely, for β a positive real number designating the temperature parameter and $x \in \mathbb{R}^K$, we have

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

Whenever $\beta \to 0$ the soft-max reduces to the classical max function. Problem (3) with the soft-max relaxation is regular enough to be solved by a gradient-based optimization method, such as e.g. accelerated gradient descent [18, 19]. Empirical study shows that $\beta = 0.005$ enables a good accuracy of our algorithm, without deviating too much from the original solution (with the max function).

Instead of the regularizing the objective function, one can also use sampling methods such as cross-entropy optimization [21] directly on the original objective function. Despite the precision of such algorithm, its downside is its computational complexity which increases much faster than the one of its smooth counterpart with the dimension of the problem. For such reason, the smooth formulation of the problem has been preferred in the following numerical study.

5 Numerical Evaluation

In this section, we discuss several numerical aspects[3] of the proposed procedure. As a benchmark, we introduce an alternative approach that enforces fairness on each individual score. Then, we illustrate the efficiency on our procedure to build fair reliable predictions on synthetic and real world datasets.

5.1 Benchmark alternative approach for fair multi-class classification

The procedure developed above enforces the score maximizer to be fair. An alternative approach [23] consists in imposing fairness at the level of each score function instead of their maximizer.

Definition 5.1. *We say that* $\mathbf{f} : \mathcal{X} \times \{-1, 1\} \mapsto \mathbb{R}^K$ *is* score-fair *in demographic parity if each coordinate of* $\mathbf{f}$ *is fair* w.r.t. *the demographic parity notion of fairness.*

Consequently, a possible way to tackle this problem is to consider the following minimization task

$$\mathbf{f}^*_{\text{score}-\text{fair}} \in \operatorname{argmin} \ \{R_2(f) : \mathbf{f} \text{ is } \textit{score-fair}\} \quad .$$

Obviously, *score-fair* DP does not imply DP over the score maximizer. Optimal *score-fair* functions rely on the L_2-risk and be easily characterized following the approach in [9, 14]. In particular, Thm. 2.3 in [9] identifies the distribution of score-fair classifier $\mathbf{f}^*_{\text{score}-\text{fair}}$ as solutions of a Wasserstein barycenter problem. The procedure for estimating $\mathbf{f}^*_{\text{score}-\text{fair}}$ is described by Alg. 2 in Appx. A.1.

5.2 Evaluation on synthetic data

Synthetic data. Conditional on $Y = k$, the feature X comes from a Gaussian mixture, while the sensitive feature S follows a Bernoulli *contamination* :

$$X|Y = k \sim \frac{1}{m}\sum_{i=1}^{m} \mathcal{N}_d(c^k + \mu_i^k, I_d) \ \text{ and } \ S|Y = k \sim 2 \cdot \mathcal{B}\left(p\mathbb{1}_{k \leq \lfloor K/2 \rfloor} + (1-p)\mathbb{1}_{k > \lfloor K/2 \rfloor}\right) - 1,$$

with $c^k \sim \mathcal{U}_d(-1, 1)$, and $\mu_1^k, \ldots, \mu_m^k \sim \mathcal{N}_d(0, I_d)$. Notably, this synthetic data structure enables to challenge different aspects of our algorithm. The parameter p measures the historical bias in the dataset. Specifically, the model becomes fair when $p = 1/2$ and completely unfair when $p \in \{0, 1\}$.

[3] The source of our method can be found at `https://github.com/xxxxxx`.

Simulation scheme. We compare our method to the benchmark *score-fair* alternative algorithm and the baseline unfair approach. We set $u = 10^{-5}$ and the probabilities p_k are estimated by RF with default parameters in `scikit-learn`. For all experiments, we generate $n = 600$ synthetic examples per class and we split the data into three sets (60% training set, 20% hold-out set and 20% unlabelled set). The performance of a classifier g is evaluated by its empirical accuracy $\mathrm{Acc}(g)$ on the hold-out set. The fairness of g is measured on the hold-out set via the empirical counterpart of the unfairness measure $\mathcal{U}(g)$ given in Definition 2.2. We repeat this procedure 30 times in order to report the average performance (accuracy and unfairness) alongside its standard deviation on the hold-out set.

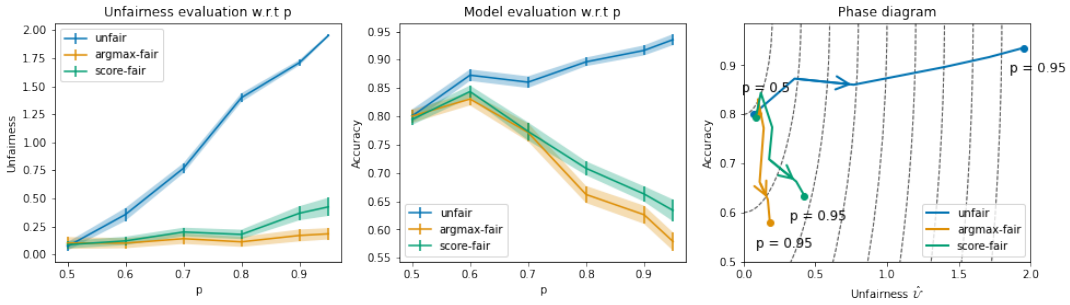

Figure 1: Performance of the classification procedures in terms of accuracy and fairness for the *unfair*, the *argmax-fair*, and the *score-fair* classifiers. Left: evolution of the unfairness *w.r.t.* p; Middle: evolution of the accuracy *w.r.t.* p; Right: (Accuracy, Unfairness) phase diagram that shows the evolution, *w.r.t.* p, of trade-off between accuracy and fairness. The arrows go from fairest to most unfair situations. Top-left corner in the diagram gives the best trade-off.

Fairness versus Accuracy. Fig. 1 displays the fairness and accuracy performances of our algorithm for different levels of historical bias in the dataset (measured by p). Our algorithm outperforms both *s*core-fair and unfair classifier, in terms of fairness efficiency. However, such fairness performance is directly counter-balanced by a weaker accuracy, as visualised on the phase diagram (Unfairness, Accuracy) in Fig. 1-right. Fairness efficiency of our methods is particularly significant for datasets with large historical bias ($p = 0.9$ or 0.95).

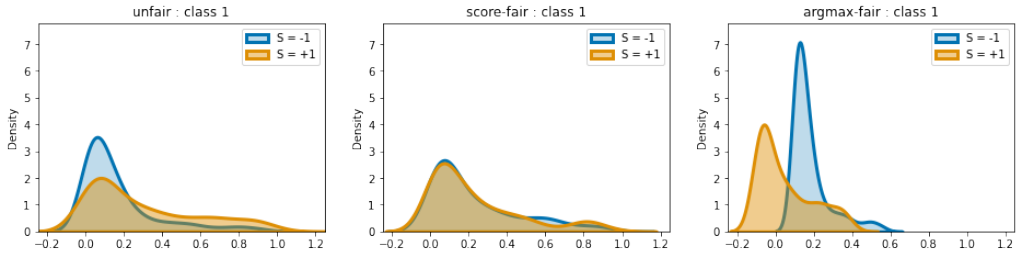

Figure 2: Empirical distribution of the score functions for the class $Y = 1$, conditional to the sensitive feature values $S = \pm 1$. *unfair* (left), *score-fair* middle and *argmax-fair* (right) classifiers.

Fairness at the level of scores. Whereas both *argmax-fair* and *score-fair* approaches succeed to build fair classification, these two methods differ significantly on their impact on scores. Fig. 2 highlights this difference for the specific class $Y = 1$, but similar behavior for other classes is presented in Appx. A.3. The right panel confirms our findings in Proposition 2.5: *argmax-fair* enforces fairness by shifting the conditional probabilities. Also expected is the fairness efficiency of *score-fair* (middle plot), while the resulting scores are easier to interpret: the distributions of the predictions for both sensitive features merge.

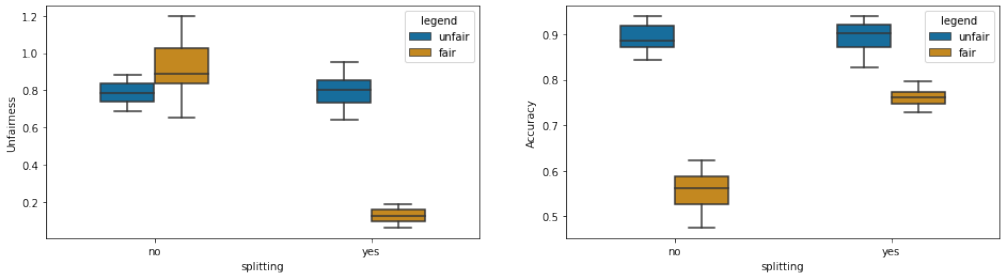

Figure 3: Empirical impact of data splitting on unfairness (Left – the lower the better) and accuracy (Right: accuracy – the higher the better). Boxplots generated over 30 repetitions with $p = 0.75$.

Splitting the sample. Our theoretical study relies on the independence of the datasets $\mathcal{D}_n$ and $\mathcal{D}'_n$. Figure 3 illustrates the importance of such condition for the fairness but also the accuracy of our proposed *argmax-fair* method. Indeed, whenever splitting is not performed (left parts of plots), the fairness performance of the fair algorithm can even be worse than the unfair method. This emphasize that splitting is crucial and enables to avoid over-fitting on the training set.

5.3 Application to real datasets

Methods. We compare our method *argmax-fair* and the alternative approach *score-fair* for both linear and non-linear multi-class classification. For linear models, we consider the one-versus-all logistic regression (reglog) and the SVM with linear kernel (linearSVC); for non-linear models: SVM model with Gaussian kernel (GaussSVC) and RF. Hyperparemeters are provided in Appx. A.2.

Datasets. The performance of our method is evaluated on four benchmark datasets : CRIME, LAW, WINE and CMC. Hereafter, we provide a short description of these datasets.

Communities&Crime (CRIME) 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 $K = 7$ balanced classes based on equidistant quantiles.Following [5] and [10] the binary sensitive feature is the percentage of black population.

Law School Admissions (LAW) dataset [22] presents national longitudinal bar passage data and has 20649 examples. The task is to predict a students GPA divided into $K = 4$ classes based on equidistant quantiles. The sensitive attribute is the race (white versus non-white).

Wine Quality (WINE) dataset [11] reports the description of 6497 wines and the task is to predict the quality graded by the experts. The quality is between 3 (bad) and 9 (good) but we

consider only $K = 5$ classes (4 to 8) due to a too low frequency of the class 3 and 9 (resp. 5 and 30 examples). The sensitive attribute is the color (red versus white).

Contraceptive Method Choice (CMC) dataset is about 1987 National Indonesia Contraceptive Prevalence Survey. The problem is to predict the contraceptive method choice of a woman (no use, long-term or short-term methods) based on her demographic and socio-economic characteristics. The sensitive feature is the religion (Islam versus Non-Islam).

DATA / METHOD	CRIME, K = 7		LAW, K = 4		WINE, K = 5		CMC, K = 3	
	Accuracy	Unfairness	Accuracy	Unfairness	Accuracy	Unfairness	Accuracy	Unfairness
reglog + unfair	0.34 ± 0.02	1.12 ± 0.07	0.43 ± 0.01	0.89 ± 0.05	0.54 ± 0.01	0.47 ± 0.05	0.52 ± 0.02	0.78 ± 0.16
reglog + score-fair	0.33 ± 0.01	0.78 ± 0.09	0.42 ± 0.01	0.09 ± 0.02	0.54 ± 0.01	0.08 ± 0.03	0.51 ± 0.02	0.25 ± 0.1
reglog + argmax-fair	0.28 ± 0.01	0.26 ± 0.07	0.42 ± 0.01	0.05 ± 0.02	0.54 ± 0.02	0.04 ± 0.01	0.52 ± 0.02	0.19 ± 0.1
linearSVC + unfair	0.36 ± 0.02	1.12 ± 0.07	0.43 ± 0.01	0.97 ± 0.07	0.53 ± 0.01	0.27 ± 0.05	0.51 ± 0.02	0.63 ± 0.22
linearSVC + score-fair	0.31 ± 0.02	0.88 ± 0.05	0.42 ± 0.01	0.1 ± 0.03	0.53 ± 0.01	0.1 ± 0.07	0.53 ± 0.02	0.26 ± 0.16
linearSVC + argmax-fair	0.29 ± 0.02	0.25 ± 0.08	0.42 ± 0.01	0.04 ± 0.02	0.53 ± 0.01	0.06 ± 0.04	0.52 ± 0.02	0.2 ± 0.12
GaussSVC + unfair	0.36 ± 0.02	1.4 ± 0.13	0.43 ± 0.01	1.04 ± 0.04	0.53 ± 0.01	0.28 ± 0.06	0.51 ± 0.02	1.0 ± 0.17
GaussSVC + score-fair	0.35 ± 0.02	1.02 ± 0.07	0.42 ± 0.01	0.16 ± 0.04	0.55 ± 0.01	0.12 ± 0.04	0.51 ± 0.02	0.16 ± 0.09
GaussSVC + argmax-fair	0.3 ± 0.02	0.22 ± 0.05	0.42 ± 0.01	0.10 ± 0.03	0.55 ± 0.01	0.06 ± 0.03	0.5 ± 0.03	0.2 ± 0.08
RF + unfair	0.37 ± 0.02	1.02 ± 0.04	0.40 ± 0.01	0.65 ± 0.04	0.66 ± 0.01	0.31 ± 0.05	0.55 ± 0.02	0.35 ± 0.18
RF + score-fair	0.34 ± 0.02	0.67 ± 0.06	0.39 ± 0.01	0.11 ± 0.05	0.66 ± 0.01	0.09 ± 0.03	0.52 ± 0.03	0.21 ± 0.08
RF + argmax-fair	0.3 ± 0.02	0.33 ± 0.11	0.39 ± 0.01	0.07 ± 0.02	0.66 ± 0.01	0.08 ± 0.02	0.55 ± 0.02	0.22 ± 0.13

Table 1: Performance (accuracy & unfairness) of the methods for all datasets and classifiers. We report the means and standard deviations over the 30 repetitions. Colored values highlight fairness.

Performance. Results are presented in Table 1 and highlight the effectiveness of our method. As an example, for the LAW dataset and the GaussSVC with *argmax-fair*, the unfairness is divided by almost 25 (0.97 to 0.04). Furthermore, the *argmax-fair* procedure outperforms the *unfair* and the *score-fair* algorithms for the datasets CRIME, LAW and WINE in terms of unfairness: However, we observe a small decrease of the models accuracy (relatively small compared to the gain in fairness). Note that for the dataset CMC, *score-fair* and *argmax-fair* achieve similar performance.

6 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 Alg. 1). Our approach achieves distribution-free fairness and can be applied on top of any probabilistic base estimator. We illustrate the proficiency of our procedure on various synthetic and real datasets, notably in comparison to the *score-fair* approach suggested in [23]. The efficiency of our algorithm in terms of fairness is particularly salient for datasets with large historical bias.

However, our numerical study also outlines the downside of fairness proficiency in terms of classification accuracy. One should hereby be very cautious when using classifiers with strong fairness guarantee, as it possibly degrades the classification quality. This calls for an analysis of classification problems with fairness constraints from a multi-objective perspective and paves the way for characterizing the Pareto front between fairness and accuracy objectives. This, together with considering (convex)-ERM when the fairness is enforced on the full vector $\hat{\mathbf{f}}$, is left for further research.

References

[1] 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.

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

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

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

[5] 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.

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

[7] 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.

[8] 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*, 2020.

[9] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with wasserstein barycenters. In *Advances in Neural Information Processing Systems*, 2020.

[10] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression via plug-in estimator and recalibration with statistical guarantees. *https://hal.archives-ouvertes.fr/hal-02501190*, 2020.

[11] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

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

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

[14] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

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

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

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

[18] Y. Nesterov. A method for solving the convex programming problem with convergence rate O(1/kˆ2). In *Doklady Akademii Nauk SSSR*, 1983.

[19] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

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

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

[22] L. F. Wightman and H. Ramsey. *LSAC national longitudinal bar passage study*. Law School Admission Council, 1998.

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

[24] 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.

[25] 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.

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

[27] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32, 2004.

Supplementary material

The supplementary material consists of two parts. Appendix A deals with additional numerical considerations while Appendix B contains all proofs of our results.

A Algorithmic considerations

In this section we provide a pseudo-code for the alternative fair classifier as well as additional numerical results.

A.1 Pseudo-code for score-fair algorithm

First of all, we recall the definition of $\mathbf{f}^*_{\text{score}-\text{fair}}$, the *score-fair* function

$$\mathbf{f}^*_{\text{score}-\text{fair}} \in \text{argmin } \{R_2(\mathbf{f}) : \mathbf{f} \text{ is } \textit{score-fair}\} \ . \tag{4}$$

where $R_2(\mathbf{f}) = \mathbb{E}\left[\sum_{k=1}^K (Z_k - f_k(X,S))^2\right]$ and $\mathbf{f}$ is *score-fair* means that for all $k \in [K]$ and for all $t \in \mathbb{R}$ we have

$$\mathbb{P}(f_k(X,S) \leq t | S = -1) = \mathbb{P}(f_k(X,S) \leq t | S = 1) \ .$$

The optimal solution for the problem (4) is separable and can then be solved element-wise. In particular, Theorem 2.3 in [9] applies and allows us to deduce the explicit form $\mathbf{f}^*_{\text{score}-\text{fair}} = (f^*_{\text{sf},1}, \ldots, f^*_{\text{sf},K}) \in \mathbb{R}^K$ such that

$$f^*_{\text{sf},k}(x,s) = \left(\pi_{-s} Q_{f^*_k|-s}\right) \circ F_{f^*_k|s}\left(f^*_k(x,s)\right) \ . \tag{5}$$

where for all $s \in \mathcal{S}$, $F_{f^*_k|s}$ is the Cumulative Distribution Function (CDF) of $f^*_k(X)|S = s$ and $Q_{f^*_k|s} : [0,1] \to \mathbb{R}$ is the corresponding quantile function defined for all $t \in (0,1]$ as $Q_{f^*_k|s}(t) = \inf\left\{y \in \mathbb{R} : F_{f^*_k|s}(y) \geq t\right\}$. Hence, it only remains to estimate each $f^*_{\text{sf},k}$ by plug-in. More precisely, we need to estimate for all $s \in \mathcal{S}$, the proportion π_s, the CDF $F_{f^*_k|s}$ and the quantile function $Q_{f^*_k|s}$. Algorithm 2 proposes a pseudo-code for the implementation of an estimator $\hat{g}_{\text{score}-\text{fair}}$ of the classifier $g^*_{\text{score}-\text{fair}}$ deduced from the *score-fair classifier* given by

$$g^*_{\text{score}-\text{fair}}(x,s) = \arg\max_{k \in [K]} f^*_{\text{sf},k}(x,s), \qquad \forall (x,s) \in \mathcal{X} \times \mathcal{S} \ .$$

We use in Algorithm 2 a close methodology to the one considered in Section 3. In particular, the base estimators $\bar{f}_k$ in the **Input** of the algorithm relies on an estimator $\hat{f}_k$ of the k-th element of the optimal score function $\mathbf{f}^*$ and on a uniform perturbation $(\zeta_k)_k$. Also, in **Step 0-a.** the constructed sample $\mathcal{D}_{\mathcal{X}} = \{X_1, \ldots, X_N\}$ consists only of the covariates from $\mathcal{D}'_N$. In **Step 0-b.** we split the sample $\mathcal{D}_{\mathcal{X}}$ into two sets $\mathcal{D}_{\mathcal{X},1}$ and $\mathcal{D}_{\mathcal{X},2}$ with size[4] $N/2$.

A.2 Hyperparameters of estimators for real datasets

The hyperparamerers are set with default parameters in scikit-learn except the number of trees for RF which is set at 500.

[4]For simplification of the presentation we assume that N is an even integer.

Algorithm 2 ERM Score-fair classifier

Input: new data point (x, s), base estimators $(\bar{f}_k)_k$, unlabeled sample $\mathcal{D}'_N$, $(\zeta_k)_k$ and $(\zeta^{j,s}_{k,i})_{k,i,j,s}$ collection of i.i.d uniform perturbations in $[0, 10^{-5}]$
Step 0-a. Separate $\mathcal{D}'_N$ to construct two samples $(S_1, \ldots, S_N)$ and $\mathcal{D}_{\mathcal{X}} = \{X_1, \ldots, X_N\}$;
Step 0-b. Split $\mathcal{D}_{\mathcal{X}}$ into two samples $\mathcal{D}_{\mathcal{X},1}$ and $\mathcal{D}_{\mathcal{X},2}$ of size $N/2$;
Step 0-c. Split $\mathcal{D}_{\mathcal{X},j}$ into two samples $\mathcal{D}^s_{\mathcal{X},j} = \{X^s_{j,1}, \ldots, X^s_{j,N_s}\}$ with $s \in \mathcal{S}$;
Step 1. Compute the empirical frequencies $(\hat{\pi}_s)_s$ based on $(S_1, \ldots, S_N)$;
for $k = 1$ **to** K **do**
For all $s \in \mathcal{S}$, estimate the CDF $F_{f^*_k|s}$ based on $\mathcal{D}^s_{\mathcal{X},1}$;
`Jittering with` $(\zeta^{1,s}_{k,i})_{k,i,1,s}$ `is needed for this step.`
For all $s \in \mathcal{S}$, estimate the quantile function $Q_{f^*_k|s}$ based on $\mathcal{D}^s_{\mathcal{X},2}$;
`Jittering with` $(\zeta^{2,s}_{k,i})_{k,i,2,s}$ `is needed for this step.`
Compute $\hat{f}_{\mathrm{sf},k}(x, s)$, the estimator of $f^*_{\mathrm{sf},k}(x, s)$ given in Eq. (5) by plug-in;
end for
Output: fair classifier $\hat{g}_{\mathrm{sf},k}(x, s) = \arg\max_{k \in [K]} \hat{f}_{\mathrm{sf},k}(x, s)$ at point (x, s).

A.3 Additional illustrations for synthetic data

We propose additional numerical illustrations that display the effectiveness of our procedure on the synthetic data presented in Section 5.2. We mainly *(i)* justify our choice of the temperature β in Figure 4 ; *(ii)* show the effectiveness of both *score-fair* and *argmax-fair* classifier in terms of unfairness reduction in Figure 5, and *(iii)* illustrate our method's robustness with respect to the number of classes K in Figure 6. By default, as in the main body, we set the number of classes $K = 4$ in all our experiments.

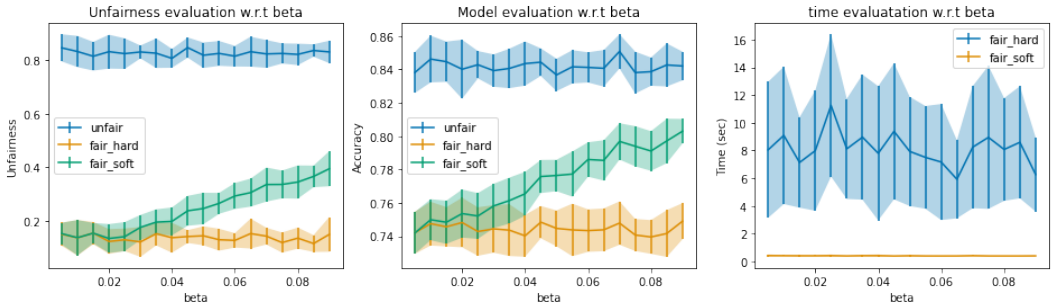

Figure 4: Performance of the classification procedures in terms of accuracy, fairness and time complexity for the *fair_soft* (argmax-fair classifier with soft-max evaluation) classifier obtained from Algorithm 1 with the acceleration scheme in **Step 2**. The *unfair* and the *fair_hard* (*argmax-fair* classifier with hard-max evaluation [21] – obtained by Algorithm 1 without acceleration in **Step 2.**) classifiers are used as baselines and Random Forest is used as base estimator. Left: evolution of the unfairness *w.r.t.* the temperature β; Middle: evolution of the accuracy *w.r.t.* β; Right: evolution of the time complexity *w.r.t.* β. We report the means and standard deviations over 30 simulations. The figure shows that both *fair_soft* and *fair_hard* have the comparable performance in terms of unfairness and accuracy for $\beta \leq 0.01$. However *fair_soft* is considered much faster than *fair_hard* hence *fair_soft* is chosen.

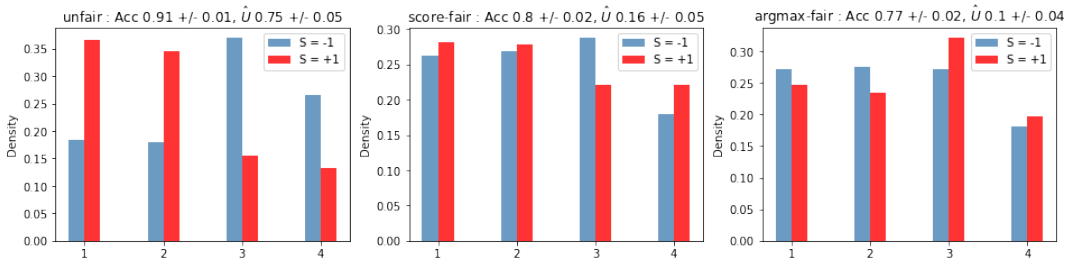

Figure 5: Empirical distribution of the *unfair* (left), the *score-fair* (middle) and the *argmax-fair* (right) classifiers conditional to the sensitive feature $S = \pm 1$. Each performance (accuracy and unfairness) is evaluated over 30 simulations and we consider RF as the base estimator. The histograms display the effectiveness of both *score-fair* and *argmax-fair* in enforcing fairness by rendering the empirical distributions across the two groups ($S = -1$ and $S = +1$) close. As shown by this empirical study, *argmax-fair* outperforms the *score-fair* classifier in terms of fairness.

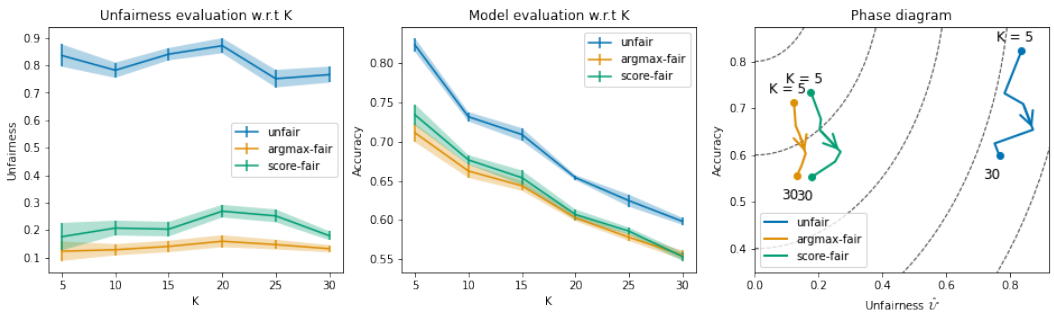

Figure 6: Performance of the classification procedures in terms of accuracy and unfairness for the *unfair*, the *argmax-fair*, and the *score-fair* classifiers. RF is used as base estimator. Left: evolution of the unfairness *w.r.t.* the number of classes K; Middle: evolution of the accuracy *w.r.t.* K; Right: (Accuracy, Unfairness) phase diagram that shows the evolution that highlights a trade-off between accuracy and fairness *w.r.t.* K. The arrows go from fairest to most unfair situations. Top-left corner in the diagram gives the best trade-off. We report the means and standard deviations over 30 simulations. The figure shows that the increase in the number of classes doesn't impact the unfairness of each method and *argmax-fair* remains more effective in fairness than the other two methods.

B Proof of main results

In this section, we gather the proof of our results. In all the sequel, C denotes a generic constant, whose value may vary from line to line.

B.1 Proof for fair optimal rule

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

Lemma B.1. *The fair-risk of a classifier g with balancing parameter $\lambda = (\lambda_1, \ldots, \lambda_K) \in \mathbb{R}^K$ rewrites:*

$$\mathcal{R}_\lambda(g) = \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[\sum_{k=1}^{K} \left(\pi_s p_k(X,S) - s\lambda_k \right) \mathbb{1}_{\{g(X,S) \neq k\}} \right] \quad . \tag{6}$$

Proof of Lemma B.1. Let $\lambda \in \mathbb{R}^K$ and recall the following definition of the fair-risk

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

We have the following decomposition

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

which directly implies (6). □

Proof of Proposition 2.4. Recall that g^*_λ minimizes $\mathcal{R}_\lambda$ on $\mathcal{G}$. Besides, we deduce from Lemma B.1 that

$$\mathcal{R}_\lambda(g^*_\lambda) = 1 - \sum_{s\in\mathcal{S}}\mathbb{E}_{X|S=s}\left[\max_{k\in[K]}\left(\pi_s p_k(X,s) - s\lambda_k\right)\right] \ . \tag{7}$$

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

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

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

$$\begin{aligned}0 = \sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}&\left(\forall j\neq k\ (\pi_s p_k(X,s) - s\lambda^*_k) > (\pi_s p_j(X,s) - s\lambda^*_j)\right)\\ &+ su^s_k\mathbb{P}_{X|S=s}\Big(\forall j\neq k\ (\pi_s p_k(X,s) - s\lambda^*_k) \geq (\pi_s p_j(X,s) - s\lambda^*_j),\\ &\qquad\qquad \exists j\neq k\ (\pi_s p_k(X,s) - s\lambda^*_k) = (\pi_s p_j(X,s) - s\lambda^*_j)\Big) \ ,\end{aligned}$$

with $u^s_k\in[0,1]$ for all $k\in[K]$ and $s\in\mathcal{S}$. Thanks to Assumption 2.3, $p_k(X,s) - p_j(X,s)$ has no atoms for all $s\in\mathcal{S}$ and then the second part of the r.h.s. vanishes. Therefore for all $k\in[K]$

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

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

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

so that $g^*_{\lambda^*}$ 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^*_{fair} with the optimality of λ^* over the family $(R_\lambda(g^*_\lambda))_{\lambda\in\mathbf{R}^K}$, we deduce

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

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

Proof of Corollary 2.5. The proof follows directly from Lemma B.1 and Proposition 2.4. In particular, Eq. (7) implies that

$$g^*_\lambda \in \arg\min_g \mathcal{R}_{\lambda^*}(g),$$

is characterized by

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

□

B.2 Proof of Consistency results

We start this section with two results, Lemmas B.2-B.3 that directly follow from similar arguments as in the proofs of Proposition A.2. and Lemma B.8 in [8] respectively. Their proofs are hence omitted.

Lemma B.2. *The parameter λ^*, and $\hat{\lambda}$ are bounded.*

Lemma B.3. *We have that, for each $s \in \mathcal{S}$ and $k \in [K]$,*

$$\mathbb{E}\left[\frac{1}{N_s}\sum_{i=1}^{N_s} \mathbb{1}\{\exists j \neq k, \hat{h}^s_k(X_i,\hat{\lambda}_k) = \hat{h}^s_j(X_i,\hat{\lambda}_j)\}\right] \leq \frac{C}{N_s} \ ,$$

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

Let us now consider the proofs of Theorem 3.1 and Theorem 3.2.

Proof of Theorem 3.1. As in Lemma B.3, we first introduce, for $s \in \mathcal{S}$ and $k \in [K]$,

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

By construction, the estimator $\bar{p}_k(X,S)$ satisfies Assumption 2.3, therefore for all $s \in \mathcal{S}$ and $k \in [K]$

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

Now, let us make use of the optimality of $\hat{\lambda}$. We denote by $\hat{\mathbb{P}}_{X|S=s}$ the empirical measure on the data $\{X^s_1,\dots,X^s_{N_s}\}$. Considering the first order optimality conditions for $\hat{\lambda}$, we can show that, for all $k \in [K]$ and $s \in \mathcal{S}$, there exists $\alpha^s_k \in [-1,1]$ such that

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

From the above equation, we deduce that

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

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

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

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 $k \in [K]$

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

Applying Lemma B.3, we then get that, Conditional on $\mathcal{D}_n$ and on $(S_1, \ldots, S_N)$, we have that

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

Since N_s is a binomial random variable with parameter (π_s, N), 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 3.2. First, let us consider the following writing of the excess risk of $\hat{g}$

$$\mathcal{E}_{\text{fair}}\left(\hat{g}\right) = \left(\mathcal{R}_{\lambda^*}\left(\hat{g}\right) - \mathcal{R}_{\hat{\lambda}}(\hat{g})\right) + \left(\mathcal{R}_{\hat{\lambda}}(\hat{g}) - \mathcal{R}_{\lambda^*}\left(g^*_{\lambda^*}\right)\right) \ . \tag{8}$$

We propose to deal with the two terms in r.h.s. of Equation (8) separately. According the the first term, we have

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

Since, for each $k \in [K]$, the parameters λ_k^* and $\hat{\lambda}_k$ are bounded (see Lemma B.2), we deduce that

$$\mathcal{R}_{\lambda^*}\left(\hat{g}\right) - \mathcal{R}_{\hat{\lambda}}(\hat{g}) \leq C\mathcal{U}\left(\hat{g}\right) \ . \tag{9}$$

Then we have shown that the first term in the r.h.s. of Eq. (8) relies on the unfairness of the classifier $\hat{g}$. Now, let us consider the second term in r.h.s. of Equation (8). Our goal will be to show that this term mainly depends on the quality of the base estimators $\hat{p}_k$. Since λ^* is a maximizer of $\mathcal{R}_\lambda(g^*_\lambda)$

over λ, it is clear that, conditional on the data, $\mathcal{R}_{\lambda^*}(g^*_{\lambda^*}) \geq \mathcal{R}_{\hat{\lambda}}(g^*_{\hat{\lambda}})$. (The parameter $\hat{\lambda}$ is seen as fixed conditional on the data.)Therefore, we have

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

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

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

Then using similar arguments as those leading to Eq. (7) implies that

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

(see also Corollary 2.5.) As a consequence, using the writing of the fair-risk provided by Lemma B.1

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

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

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

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

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

and obtain the bound

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

In view of Equation (8), the above inequality together with Equation (9) yield the desired result. □

B.3 Proof of rate of convergence

Proof of Proposition 3.5. From Theorems 3.1 and 3.2, we have that

$$\mathbb{E}[\mathcal{E}_{\text{fair}}(g_{\hat{\mathbf{f}}})] \quad \leq \quad \mathbb{E}\left\|\hat{\boldsymbol{p}} - \mathbf{p}\right\|_1 + \frac{1}{n} + \frac{C}{\sqrt{N}} \ .$$

Since

$$\mathbb{E}\left\|\hat{\boldsymbol{p}} - \mathbf{p}\right\|_1 \leq \frac{1}{2}\mathbb{E}\left\|\hat{\boldsymbol{f}} - \boldsymbol{f}^*\right\|_1 \leq \frac{1}{2}\sqrt{\sum_{k=1}^{K}\mathbb{E}\left[\left(f_k(X,S) - f_k^*(X,S)\right)^2\right]} \ ,$$

it remains to provide a control on the term $\sum_{k=1}^{K}\mathbb{E}\left[\left(f_k(X,S) - f_k^*(X,S)\right)^2\right]$. For this purpose, let us first prove that for each score function $\mathbf{f} \in \mathcal{F}$, the following holds

$$\sum_{k=1}^{K}\mathbb{E}\left[\left(f_k(X,S) - f_k^*(X,S)\right)^2\right] \leq 2\left(R_2(\mathbf{f}) - R_2(\mathbf{f}^*)\right) \ . \tag{10}$$

Indeed, we observe that

$$\frac{(Z_k - f_k)^2 + (Z_k - f_k^*)^2}{2} - \left(Z_k - \left(\frac{f_k + f_k^*}{2}\right)\right)^2 = \frac{(f_k - f_k^*)^2}{4} \ .$$

From this equality, we then deduce that

$$\frac{1}{4}\sum_{k=1}^{K}\mathbb{E}\left[\left(f_k(X,S) - f_k^*(X,S)\right)^2\right] = \frac{1}{2}\left(R_2(\mathbf{f}) + R_2(\mathbf{f}^*)\right) - R_2\left(\frac{\mathbf{f} + \mathbf{f}^*}{2}\right) \ .$$

Since R_2 is positive, we get Equation (10).

The next step of the proof is to bound $R_2(\hat{\mathbf{f}}) - R_2(\mathbf{f}^*)$. We have by definition of $\hat{\mathbf{f}}$

$$R_2(\hat{\mathbf{f}}) - R_2(\mathbf{f}^*) \leq R_2(\hat{\mathbf{f}}) - R_2(\mathbf{f}^*) - 2\left(\hat{R}_2(\hat{\mathbf{f}}) - \hat{R}_2(\mathbf{f}^*)\right).$$

Furthermore from Assumption 3.4 with $\varepsilon = 1/n$, we have

$$R_2(\hat{\mathbf{f}}) - R_2(\mathbf{f}^*) - 2\left(\hat{R}_2(\hat{\mathbf{f}}) - \hat{R}_2(\mathbf{f}^*)\right) \leq \frac{C}{n} + \sup_{f \in \mathcal{F}_{1/n}}\left\{R_2(\mathbf{f}) - R_2(\mathbf{f}^*)\right\} - 2\left(\hat{R}_2(\hat{\mathbf{f}}) - \hat{R}_2(\mathbf{f}^*)\right) \ . \tag{11}$$

If we denote $\text{Err}(\mathbf{f}) = \text{R}_2(\mathbf{f}) - \text{R}_2(\mathbf{f}^*)$ and $\widehat{\text{Err}}(\mathbf{f}) = \hat{R}_2(\hat{\mathbf{f}}) - \hat{R}_2(\mathbf{f}^*)$, from Bernstein's Inequality together with Assumption 3.4, we have, for all $t > 0$ and $\mathbf{f} \in \mathcal{F}_\varepsilon$,

$$\begin{aligned}
& \mathbb{P}\left(\text{Err}(\mathbf{f}) - 2 \cdot \widehat{\text{Err}}(\mathbf{f}) \geq \text{t}\right) \\
\leq \ & \mathbb{P}\left(2\left(\text{Err}(\mathbf{f}) - \widehat{\text{Err}}(\mathbf{f})\right) \geq t + \text{Err}(\mathbf{f})\right) \\
\leq \ & \exp\left(-\frac{\frac{n}{8} \cdot (t + \mathbb{E}[h(Z, \mathbf{f}(X,S))])^2}{\mathbb{E}[|h(Z, \mathbf{f}(X,S))|^2] + \frac{C}{3} \cdot (t + \mathbb{E}[h(Z, \mathbf{f}(X,S))])}\right)
\end{aligned}$$

where $h(Z, \mathbf{f}(X,S)) := \sum_{k=1}^{K}(|Z - f_k(X,S)|^2 - |Z - f_k^*(X,S)|^2)$. Furthermore, observe that

$$\mathbb{E}[|h(Z, \mathbf{f}(X,S))|^2] \leq C \cdot \mathbb{E}[h(Z, \mathbf{f}(X,S))] \ ,$$

which, plugged in the previous inequality, directly provides

$$\mathbb{P}\left(\mathrm{Err}(\mathbf{f}) - 2 \cdot \widehat{\mathrm{Err}}(\mathbf{f}) \geq \mathrm{t}\right) \quad \leq \quad \exp\left(-Cnt\right) \quad .$$

Hence, combining a union bound argument together with Assumption 3.4 and (11), we compute

$$\mathbb{E}[R_2(\hat{\mathbf{f}}) - R_2(\mathbf{f}^*)] \leq \frac{C}{n} + CC_{\mathcal{F}}\frac{\log(n)}{n} \quad . \tag{12}$$

Plugging this inequality in (10) concludes the proof. □