Set to Be Fair: Demographic Parity Constraints for Set-Valued Classification

Eyal H. Cohen
Sorbonne Université

Christophe Denis
Université Paris 1 Panthéon-Sorbonne

Mohamed Hebiri
Université Gustave Eiffel

Abstract

Set-valued classification is used in multiclass settings where confusion between classes can occur and lead to misleading predictions. However, its application may amplify discriminatory bias motivating the development of set-valued approaches under fairness constraints. In this paper, we address the problem of set-valued classification under demographic parity and expected size constraints. We propose two complementary strategies: an oracle-based method that minimizes classification risk while satisfying both constraints, and a computationally efficient proxy that prioritizes constraint satisfaction. For both strategies, we derive closed-form expressions for the (optimal) fair set-valued classifiers and use these to build plug-in, data-driven procedures for empirical predictions. We establish distribution-free convergence rates for violations of the size and fairness constraints for both methods, and under mild assumptions we also provide excess-risk bounds for the oracle-based approach. Empirical results demonstrate the effectiveness of both strategies and highlight the efficiency of our proxy method.

1 INTRODUCTION

Set-valued classifiers are powerful tools for handling ambiguity between class labels in multiclass classification problems. Their popularity grew with the advent of conformal prediction (Vovk et al., 2005) and has become increasingly important in large-scale settings. Numerous set-valued frameworks now coexist, each offering different trade-offs and applications (Denis and Hebiri, 2017; Chzhen et al., 2021; Sadinle et al., 2019). In parallel, the rapid expansion of machine learning and deep learning in critical and sensitive domains such as medicine (Celard et al., 2023), hiring (Chen, 2023), criminal justice (Taylor, 2023), and banking (Sadok et al., 2022; Amato et al., 2024), has made algorithmic fairness a central concern in the statistical and machine learning communities (Hardt et al., 2016; Agarwal et al., 2018; Chzhen et al., 2019; Paulus and Kent, 2020; Hobson et al., 2023; Yang et al., 2023; Chen et al., 2023; Cameron et al., 2024). The main issue addressed by algorithmic fairness is the mitigation of learned biases and discrimination arising from sensitive attributes such as gender, ethnicity, or socioeconomic status. A wide range of methods attempt to implement fairness through pre-, in-, or post-processing, targeting either exact fairness or approximate fairness (also called ϵ-fairness), the latter allowing for improved trade-offs with predictive performance (Zemel et al., 2013; Lum and Johndrow, 2016; Calders et al., 2009b; Feldman et al., 2015; Zafar et al., 2017; Barocas et al., 2018; Chzhen et al., 2019; Jiang et al., 2020; Gordaliza et al., 2019; Hardt et al., 2016; Dwork et al., 2012). See (Alves et al., 2023) for a recent review. Approximate fairness offers a flexible way to control fairness while limiting the accuracy drop. However, it requires a predefined level of unfairness, which can be difficult to interpret and calibrate in practice.

In this paper, we consider a set-valued classification problem involving multiple classes and a sensitive attribute. The goal is to build a classifier that outputs a subset of classes while ensuring fairness – in a sense to be specified later – with respect to the sensitive attribute and controlling the average size of the output to limit information disclosure. We focus on exact fairness, but allow a compromise on the interpretability of the output in order to reduce the classification risk. Specifically, we adopt the framework of set-valued classification (Lapin et al., 2016; Denis and Hebiri, 2017; Sadinle et al., 2019), which includes ideas related to the conformal prediction setting (Vovk et al., 2005). In this

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

framework, the classifier may output multiple candidate labels, and the misclassification risk is naturally defined as the probability that the true label is not included in the predicted set. A key advantage of set-valued classifiers under an expected size constraint is that, by allowing larger outputs in ambiguous cases, one can reduce the overall misclassification risk. However, it has been observed that in some applications where set-valued classification is particularly appropriate – such as image classification – biases in the data may lead to systematic misclassification (Besse et al., 2018). This highlights the need to develop fair set-valued classifiers. In this work, we incorporate demographic parity (DP) as a fairness constraint to ensure that the classifier does not discriminate based on the sensitive attribute.

1.1 Related Work

Fairness in classification has been extensively studied under various criteria such as demographic parity (Calders et al., 2009a), equalized odds, and equal opportunity (Hardt et al., 2016). Two scenarios are typically considered: awareness and unawareness (Agarwal et al., 2018; Chzhen et al., 2019; Wang et al., 2022; Gaucher et al., 2023) – whether we have access to the sensitive attribute at prediction time or not. With exact fairness being the concept of calibrating the algorithms to completely remove biases with respect to a given sensitive attribute, a relaxed version, known as approximate fairness or ϵ-fairness allows for a trade-off between accuracy and fairness (Agarwal et al., 2018; Denis et al., 2024). While appealing from a performance standpoint, ϵ-fairness is often less interpretable, as it relies on empirically chosen thresholds for acceptable unfairness. To improve interpretability, α-fairness has been proposed (Chzhen and Schreuder, 2022), which seeks predictions that are at most α times as unfair as an unconstrained baseline, providing a clearer and more intuitive fairness guarantee.

Conformal prediction offers a natural framework for set-valued classification by providing calibrated prediction sets with coverage guarantees (Gibbs et al., 2023; Vovk et al., 2005). Recent work has extended this framework to incorporate fairness constraints, such as adaptively selecting features and equalizing coverage across groups (Zhou and Sesia, 2025), or by combining conformal prediction with quantile regression and fairness adjustments (Romano et al., 2019; Liu et al., 2022). The key difference between conformal prediction and our framework lies in the type of constraint being imposed. While conformal prediction enforces an α-coverage constraint, for some specific $\alpha \in (0, 1)$, we rather impose a constraint on the expected size bounded by a parameter $\beta \in (0, K)$. We refer to Section 6 for a discussion on the differences between conformal prediction framework and our approach.

More broadly, set-valued predictors have been widely used to address class ambiguity in multiclass problems (see (Chzhen et al., 2021) for a review) but has not been explored from the fairness perspective.

Our work focuses on set-valued classification under fairness and size constraints. We provide an explicit solution of the fair set-valued classifier along with theoretical guarantees on constraint violations and excess risk. We also show that, while using a post-processing approach, the constraint violations guarantees are independent of the quality of the underlying estimators.

1.2 Main contributions

Our work focuses on the set-valued classification problem and the demographic fairness constraint. Our main contributions are the following: **i)** we extend the notion of demographic parity to the set-valued classification setting; **ii)** we exhibit a closed-form expression of the optimal fair set-valued classifier under an expected size constraint and deduce from its expression a data-driven procedure based on the plug-in principle. A key feature of the method is its post-processing nature: any preliminary estimator of the conditional probabilities can be used to build a fair set-valued classifier with the prescribed size, using *unlabeled* data only, making it attractive in practice. We provide theoretical controls on the risk, the unfairness and the size of the proposed set-valued classifier. Notably, both guarantees on the constraints are distribution-free. **iii)** we propose a computationally efficient alternative to the optimal approach that avoids the need for solvers. Although not optimal, this proxy satisfies the same constraint violation guarantees, making it a practical alternative. **iv)** we conduct numerical comparisons on both synthetic and real data, demonstrating the relevance of both approaches in practice.

1.3 Paper Outline

Section 2 presents the optimal fair classifier and Section 3 is devoted to analyse statistically its data-driven counterpart. Section 4 introduces a computationally simpler two-step procedure. Section 5 provides empirical results. Finally, several aspects of our approach are discussed in Section 6 followed by a conclusion in Section 7.

2 GENERAL FRAMEWORK

2.1 Statistical setting

We begin with some useful notation. Let $K \geq 2$ be an integer and write $[K]$ to denote the set $\{1, \ldots, K\}$. Let

$(X, S, Y) \in \mathcal{X} \times \mathcal{S} \times [K]$ be a random tuple with distribution $\mathbb{P}$, respectively denoting by X the covariates, S the sensitive attribute, and Y the class label. A set-valued classifier is a function mapping $\mathcal{X} \times \mathcal{S}$ to the power set of classes $2^{[K]}$. Let $\mathbf{\Gamma}$ denote the collection of all set-valued classifiers. For any $\Gamma \in \mathbf{\Gamma}$, two quantities are of interest: the expected size $\mathcal{T}(\Gamma) = \mathbb{E}\left[|\Gamma(X, S)|\right]$ and the risk $R(\Gamma) = \mathbb{P}\left(Y \notin \Gamma(X, S)\right)$. These two objectives are typically in tension: larger sets tend to reduce the risk but increase the size. For every (x, s, k) in $\mathcal{X} \times \mathcal{S} \times [K]$, we denote by $p_k(x, s) = \mathbb{P}\left(Y = k | X = x, S = s\right)$ the conditional class probabilities. The marginal distribution of the sensitive attribute S is denoted by $\pi_s := \mathbb{P}\left(S = s\right)$ for each $s \in \mathcal{S}$. A central tool in our analysis is the use of cumulative distribution functions (cdf) and their general inverses. For each $k \in [K]$ and $s \in \mathcal{S}$, we denote by F_k (respectively $F_{k,s}$) the cdf of $p_k(X, S)$ under the distribution $\mathbb{P}_{(X,S)}$ of (X, S) (respectively the conditional distribution $\mathbb{P}_{X|S=s}$ of X given $S = s$). Moreover, for any real-valued random variable U, we define $\overline{F}_U = 1 - F_U$. Finally, we introduce the function G defined by $G(t) := \sum_{k=1}^{K} \overline{F}_k(t)$ for $t \in \mathbb{R}$ and denote by G^{-1} its generalized inverse.

DP-fair set-valued classifier. We address the fairness problem within the Demographic Parity (DP) framework adapted to the set-valued setting. This leads to the following definition:

Definition 2.1 (DP-constraint)**.** A set-valued classifier $\Gamma \in \mathbf{\Gamma}$ is said to be DP-fair if, for all $k \in [K]$ and $s \in \mathcal{S}$

$$\mathbb{P}_{X|S=s}\left(k \in \Gamma(X, s)\right) = \mathbb{P}_{X,S}\left(k \in \Gamma(X, S)\right) \quad .$$

We denote by $\mathbf{\Gamma}_{\text{fair}}$ the set of all classifiers satisfying the DP constraint.

This definition is a direct extension of the notion of DP in classification (Calders et al., 2009a) to the set-valued setting. Our goal here is to build a set-valued classifier that minimizes the risk under the DP constraint and that has a bounded expected size. More formally, for a fixed limiting size $\beta > 0$, we aim to solve the following constrained optimization problem:

$$\Gamma^*_\beta \in \arg\min\left\{R(\Gamma) \ : \ \Gamma \in \mathbf{\Gamma}_{\text{fair}}, \ \mathcal{T}(\Gamma) \leq \beta\right\} \quad . \quad (1)$$

2.2 Optimal Predictor

One convenient way to get a closed form expression of the optimal predictor Γ^*_β is to lies under the following assumption.

Assumption 1 (Continuity)**.** For each $k \in [K]$ and $s \in \mathcal{S}$, the cdf $F_{k,s}$ is continuous.

Theorem 2.2. *Suppose Assumption 1 is verified. Then the β-specific oracle Γ^*_β is:*

$$\Gamma^*_\beta(x, s) = \left\{k \in [K] : p_k(x, s) \geq \lambda^* + \frac{\gamma^*_{k,s}}{\pi_s}\right\} \quad ,$$

with $\gamma^*_{k,s} = \alpha^*_{k,s} - \pi_s \sum_s \alpha^*_{k,s}$ and λ^* and $\alpha^* = (\alpha_{k,s})_{k \in [K], s \in \mathcal{S}}$ are the Lagrangian multiplier that are characterized as

$$(\lambda^*, \alpha^*) \in \underset{\substack{(\lambda,\alpha) \in \mathbb{R}^{K|\mathcal{S}|+1} \\ \lambda \geq 0}}{\operatorname{argmin}} \ H(\lambda, \alpha) + \lambda\beta, \quad (2)$$

with

$$H(\lambda, \alpha) = \sum_{k,s} \mathbb{E}_{X|S=s}\left[\left(\pi_s\left(p_k(X, s) - \lambda + \sum_s \alpha_{k,s}\right) - \alpha_{k,s}\right)_+\right],$$

where $(\cdot)_+$ stands for the positive part.

The above result shows that, under Assumption 1, the optimal predictor can be characterized as a thresholding rule applied to the conditional probabilities p_k. This threshold is composed of two components: the first, λ^*, is a Lagrange multiplier associated with the expected size constraint, and is therefore responsible for calibrating the average size of the predictor Γ^*_β. The second component adjusts λ^* in a class- and group-specific manner to enforce the fairness constraint. Notably, this characterization extends the one derived in Denis and Hebiri (2017), where only the expected size constraint is considered. In their setting, the threshold involves a single parameter that does not depend neither on the class-label nor on the sensitive feature.

An important issue that remains is the resolution of the optimization problem in Equation (2). The Lagrange multipliers obtained are not unique: the fairness-related parameters α^* can be shifted by a common constant without affecting the resulting classifier Γ^*_β. To address this, and in light of the definition of the optimal fairness parameter γ^*, which satisfies $\sum_{s \in \mathcal{S}} \gamma^*_{k,s} = 0$, the optimization problem can be reparameterized as follows:

$$(\lambda^*, \gamma^*) \in \underset{\substack{(\lambda,\gamma) \in \mathbb{R}^{K|\mathcal{S}|+1} \\ \lambda \geq 0 \\ \sum_{s \in \mathcal{S}} \gamma_{k,s} = 0}}{\operatorname{argmin}} \ J(\lambda, \gamma) \ + \ \lambda\beta \ , \quad (3)$$

with

$$J(\lambda, \gamma) = \sum_{k,s} \mathbb{E}_{X|S=s}\left[\left(\pi_s\left(p_k(X, s) - \lambda\right) - \gamma_{k,s}\right)_+\right].$$

While this reparameterization does not ensure the uniqueness of the pair (λ^*, γ^*), it simplifies the optimization landscape and the construction of the oracle

predictor Γ^*_β. In particular, it allows for easier verification of key properties of the objective function, such as coercivity. We now state several properties of the β-specific oracle Γ^*_β, which will facilitate the analysis in the following sections.

Risk measure. The next result provides an important characterization of the optimal predictor.

Proposition 2.3. Let Γ^*_β be the optimal predictor. Under Assumption 1 the predictor Γ^*_β is DP-fair, and has expected size β. Furthermore, the following holds. $\Gamma^*_\beta \in \arg\min_{\Gamma\in\mathbf{\Gamma}} \mathcal{R}_{\lambda^*,\gamma^*}(\Gamma)$, with

$$\mathcal{R}_{\lambda^*,\gamma^*}(\Gamma) = R(\Gamma) + \lambda^*\left(\mathbb{E}_X\left[|\Gamma(X,S)|\right] - \beta\right) + \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \gamma^*_{k,s}\mathbb{P}_{X|S=s}\left(k\in\Gamma(X,s)\right) \quad .$$

The above proposition shows that the optimal predictor achieves the prescribed expected size β and can be characterized as the minimizer, over all set-valued classifiers, of the Lagrangian objective $\mathcal{R}_{\lambda^*,\gamma^*}$. In particular, this highlights that $\mathcal{R}_{\lambda^*,\gamma^*}$ serves as a relevant surrogate risk in our framework, as it naturally balances three competing objectives: classification accuracy, expected size, and fairness. Moreover, this characterization allows us to derive a closed-form expression for the excess risk of any classifier $\Gamma \in \mathbf{\Gamma}$ relative to the optimal fair predictor.

Corollary 2.4. *Let (λ^*,γ^*) be a solution of (3). Then for each $\Gamma\in\mathbf{\Gamma}$, we have that*

$$\mathcal{R}_{\lambda^*,\gamma^*}(\Gamma) - \mathcal{R}_{\lambda^*,\gamma^*}\left(\Gamma^*_\beta\right) = \sum_{k,s}\mathbb{E}_{X|S=s}\left[\mathbb{1}_{\left\{k\in\Gamma(X,s)\Delta\Gamma^*_\beta(X,s)\right\}}\left|H_k(X,s)\right|\right] \quad ,$$

with $H_k(X,s) = \left|\pi_s\left(p_k(X,s)-\lambda^*\right) - \gamma^*_{k,s}\right|$, and where Δ stands for the symmetric difference of two sets.

A direct consequence of the above result is that, under Assumption 1, the optimal predictor Γ^*_β is *a.s.* unique – this follows from the expression of the excess risk, which involves the symmetric difference between Γ and Γ^*_β on the right-hand side. In particular, if $\widetilde{\Gamma}$ is any solution to the minimization problem in Equation (1), then $\widetilde{\Gamma} = \Gamma^*_\beta$ *a.s.*

On the uniqueness of the optimal predictor. We have shown that the optimal predictor is almost surely unique. Under a more structural assumption, we can further establish the uniqueness of the parameters (λ^*,γ^*) from (3) that characterize Γ^*_β. To that end, we strengthen Assumption 1 with the following condition:

Assumption 2 (Positive density)**.** For each $s\in\mathcal{S}$, the random variables $p_k(X,S)$ admit a strictly positive and continuous density *w.r.t.* $\mathbb{P}_{X|S=s}$.

This assumption ensures that both $F_{k,s}$ and F_k are bijective. In particular, we obtain the following result:

Proposition 2.5. Suppose that Assumption 2 holds. Then, the optimal parameters (λ^*,γ^*) are unique; the predictor Γ^*_β admits a unique parametrization:

$$\Gamma^*_\beta(x,s) = \left\{k\in[K] : p_k(x,s) \geq \bar{F}^{-1}_{k,s}\left(\beta^*_k\right)\right\},$$

with $\beta^*_k = \mathbb{P}\left(k\in\Gamma^*_\beta(X,S)\right)$.

Under the positive density assumption on $p_k(X,S)$, the expression of the threshold in Theorem 2.2 simplifies. In particular, we have $\bar{F}^{-1}_{k,s}\left(\beta^*_k\right) = \lambda^* + \frac{\gamma^*_{k,s}}{\pi_s}$. In Remark 4.1, we leverage this expression to highlight key optimality properties.

3 DATA-DRIVEN PROCEDURE

3.1 Procedure

Our estimation procedure aims to recover the β-specific DP-fair set-valued classifier Γ^*_β introduced in Theorem 2.2, following the plug-in principle. The overall strategy consists in estimating the unknown components involved in the expression of Γ^*_β. Notably, some of these components do not depend on the full data distribution $\mathbb{P}$, which enables a semi-supervised estimation approach – reminiscent of the approach proposed in Denis et al. (2024).

More formally, let $n, N > 1$ be two integers. We assume access to two independent datasets: a first labeled dataset denoted by $\mathcal{D}_n = \{(X_i,S_i,Y_i), i=1,\dots,n\}$, and a second unlabeled dataset $\mathcal{D}_N = \{(X_i,S_i), i=n+1,\dots,n+N\}$. Based on $\mathcal{D}_n$, for each $k\in[K]$, we construct an estimator $\widetilde{p}_k$ of the regression function p_k. This estimation step is standard and has been extensively studied in the literature. In practice, any suitable machine learning method can be employed, such as kernel-based estimators or random forests. To derive theoretical guarantees on the excess risk, we require that the estimated scores satisfy a continuity property analogous to Assumption 1. To enforce this, we introduce a (small) perturbation: let $\epsilon\sim\mathcal{U}([0,10^{-\eta}])$ be an independent random noise, independent from all other data. This additive noise ensures that ties in the estimated probabilities occur with probability zero, without impacting the validity of the procedure. We thus define the final estimator of the class-conditional probabilities as $\widehat{p}_k(x,s) = \widetilde{p}_k(x,s)+\epsilon$.

In a second step, based on the unlabeled dataset $\mathcal{D}_N$, we estimate both the distribution of the sensitive attribute

S and the Lagrangian parameters (λ^*, γ^*). For each $s \in \mathcal{S}$, we build the subset $\mathcal{D}_{N_s} = \{(X_i, S_i) \in \mathcal{D}_N : S_i = s\}$, with corresponding size $N_s = \sum_{i=n+1}^{n+N} \mathbb{1}_{\{S_i=s\}}$. The distribution $(\pi_s)_{s\in\mathcal{S}}$ is estimated using the empirical frequencies $(\widehat{\pi}_s)_{s\in\mathcal{S}}$ with $\widehat{\pi}_s = \frac{N_s}{N}$. Next, inspired by the Lagrangian formulation in Equation (3), we define the empirical parameters $(\widehat{\lambda}, \widehat{\gamma})$ as the solution of the following convex optimization problem:

$$(\widehat{\lambda}, \widehat{\gamma}) \in \underset{\substack{(\lambda,\gamma)\in\mathbb{R}^{K|\mathcal{S}|+1} \\ \lambda\geq 0 \\ \sum_{s\in\mathcal{S}}\gamma_{k,s}=0}}{\operatorname{argmin}} \hat{J}(\lambda, \gamma) + \lambda\beta, \tag{4}$$

with

$$\hat{J}(\lambda, \gamma) = \sum_{k,s} \frac{1}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \left(\widehat{\pi}_s \left(\widehat{p}_k(X_i, s) - \lambda\right) - \gamma_{k,s}\right)_+.$$

The final predictor $\widehat{\Gamma}_\beta$ is then defined pointwise, using these estimated parameters

$$\widehat{\Gamma}_\beta(x, s) = \left\{k \in [K] : \widehat{p}_k(x, s) \geq \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right\} .$$

3.2 Rates of Convergence

The previous section introduced the plug-in set-valued predictor. We now turn to its theoretical performance, focusing on finite-sample guarantees in terms of expected size and fairness constraint violation. To this end, we quantify fairness violation through the unfairness measure $\mathcal{U}(\Gamma)$ defined as

$$\max_{k,s,s'} \left|\mathbb{P}_{X|S=s}\left(k \in \Gamma(X, s)\right) - \mathbb{P}_{X|S=s'}\left(k \in \Gamma(X, s')\right)\right|.$$

This definition extends the fairness measure introduced in Denis et al. (2024) for single-output multiclass classifiers to the set-valued prediction setting. It captures the largest discrepancy, across all class labels and sensitive groups, in the probability that a class is selected by the predictor. We can now state our first result, which shows that the proposed plug-in predictor derived satisfies both fairness and expected size constraints at a controlled rate:

Theorem 3.1 (Fairness and Expected size controls)**.** *Let $\widehat{\Gamma}_\beta$ be the empirical DP-fair set-valued classifier resulting from the procedure. Then, for any data-generating distribution $\mathbb{P}$ and any estimators $\widetilde{p}_k$ of the class-conditional probabilities, the following bounds hold:*

$$\mathbb{E}\left[\mathcal{U}(\widehat{\Gamma}_\beta)\right] \leq \frac{CK}{\sqrt{N}}, \qquad \mathbb{E}\left[\left|\mathcal{T}\left(\widehat{\Gamma}_\beta\right) - \beta\right|\right] \leq \frac{CK}{\sqrt{N}},$$

where $C > 0$ is a universal constant.

The above result provides distribution-free guarantees: it holds uniformly over all distributions $\mathbb{P}$ and all base estimators $(\widetilde{p}_k)_{k=1}^K$. It shows that the proposed method closely mimics the oracle β-specific DP-fair set-valued classifier Γ^*_β – which satisfies both constraints exactly — at a parametric rate. These bounds combine and extend the results of Denis and Hebiri (2017) and Denis et al. (2024), by simultaneously addressing both the fairness constraint and the expected size constraint in the more general set-valued classification setting. Overall, for both bound, we get a linear cost in K for handling multiple outputs and a convergence rate proportional to $1/\sqrt{N}$ with respect to the size of the unlabeled dataset.

We now turn to bounding the risk of the empirical set-valued classifier and compare it to the optimal fair predictor satisfying both the fairness and size constraints.

Theorem 3.2 (excess-risk control)**.** *Let $\widehat{\Gamma}_\beta$ be the empirical DP-fair set-valued classifier. Let $\mathcal{R}_{\lambda^*,\alpha^*}(\cdot)$ the set-valued risk from Proposition 2.3. Then we have*

$$\mathcal{R}_{\lambda^*,\alpha^*}(\widehat{\Gamma}_\beta) - \mathcal{R}_{\lambda^*,\alpha^*}(\Gamma^*_\beta) \leq C_{K,\mathcal{S}}\left(\frac{1}{\sqrt{N}} + \max_{s\in\mathcal{S}} ||\widehat{p} - p||_{\infty,\mathbb{P}_{X|S=s}}\right) ,$$

where $C_{K,\mathcal{S}} > 0$ depends only on K and $|\mathcal{S}|$ and $||\widehat{p} - p||_{\infty,\mathbb{P}_{X|S=s}} = \mathbb{E}_{\mathcal{D}_n}\sup_{x\in\mathcal{X}} |\widehat{p}_k(x, s) - p_k(x, s)|$ for all $s \in \mathcal{S}$ with $\mathbb{E}_{\mathcal{D}_n}$ being the expectation w.r.t. *the law of $\mathcal{D}_n$.*

This bound is composed of two terms: the first, of order $1/\sqrt{N}$, reflects the impact of estimating the Lagrange multipliers based solely on unlabeled data and governs the control of constraint violations; the second term corresponds to the estimation error of the conditional class probabilities $(p_k)_{k\in[K]}$. In particular, the result shows that the plug-in estimator $\widehat{\Gamma}_\beta$ performs nearly as well as the oracle predictor Γ^*_β, provided the class probability estimators converge uniformly, *i.e.,* $\max_{s\in\mathcal{S}} ||\widehat{p} - p||_{\infty,\mathbb{P}_{X|S=s}}$ tends to 0 as the number of labeled data n tends to ∞.

4 A TWO-STEP PROCEDURE: SIZE-TO-FAIRNESS SET-VALUED CLASSIFIER

In this section, we introduce a simpler alternative to the above plug-in approach that avoids optimizing (3). The proposed *size-to-fairness* classifier starts from a size-constrained but potentially unfair predictor and corrects it to ensure fairness.

4.1 Description of the procedure

The method builds upon the characterization given in Proposition 2.5. Let $\widetilde{\Gamma}$ denote a set-valued classifier with expected size β, and define, for each $k \in [K]$, the marginal inclusion rate $\beta_k = \mathbb{P}\left(k \in \widetilde{\Gamma}(X,S)\right)$. We define the associated *fair* set-valued classifier $\widetilde{\Gamma}_{\text{fair}}$ by thresholding each conditional probability using the quantiles of the stratum-specific distributions:

$$\widetilde{\Gamma}_{\text{fair}}(x,s) = \left\{k \in [K],\ \ p_k(x,s) \geq \bar{F}_{k,s}^{-1}(\beta_k)\right\} \ .$$

Under Assumption 1, for every $k \in [K]$ and $s \in \mathcal{S}$, we have

$$\mathbb{P}_{X|S=s}\left(k \in \widetilde{\Gamma}_{\text{fair}}(X,S)\right) = \bar{F}_{k,s}\left(\bar{F}_{k,s}^{-1}(\beta_k)\right) = \beta_k \ ,$$

which ensures that this new predictor satisfies the Demographic Parity (DP) constraint. Moreover, the expected size of $\widetilde{\Gamma}_{\text{fair}}$ remains β, since the transformation preserves marginal inclusion rates. Both fairness and size constraints are thus simultaneously met.

To specify $\widetilde{\Gamma}$, we follow the construction proposed in Denis and Hebiri (2017) and define

$$\widetilde{\Gamma}(X,S) = \left\{k \in [K],\ \ p_k(X,S) \geq G^{-1}(\beta)\right\} \ ,$$

where $G^{-1}(\beta)$ is the global threshold ensuring $\mathcal{T}(\widetilde{\Gamma}) = \beta$. This classifier is known to solve

$$\widetilde{\Gamma} \in \arg\min\{R(\Gamma) \ \ s.t. \ \ \mathcal{T}(\Gamma) \leq \beta\} \ ,$$

and is thus optimal among all size-constrained predictors, albeit not necessarily fair. Furthermore, we have $\mathbb{P}\left(k \in \widetilde{\Gamma}(X,S)\right) = \bar{F}_k\left(G^{-1}(\beta)\right)$. Combining the above steps, we obtain the *size-to-fair* set-valued classifier $\widetilde{\Gamma}_{\beta 2\text{DP}}$, defined as:

$$\widetilde{\Gamma}_{\beta 2DP}(x,s) = \left\{k \in [K],\ \ p_k(x,s) \geq \bar{F}_{k,s}^{-1}\left(\bar{F}_k\left(G^{-1}(\beta)\right)\right)\right\} \ .$$

We emphasize that although the predictor $\widetilde{\Gamma}_{\beta 2DP}$ achieves both the expected size and demographic parity constraints by construction, it may not be optimal in terms of risk minimization. This stems from the fact that the thresholds defining $\widetilde{\Gamma}_{\beta 2DP}$ might differ from those of the optimal DP-fair predictor Γ^*_β, which explicitly minimizes the risk under fairness and size constraints.

The following remark illustrates this discrepancy under additional assumptions ensuring the uniqueness of the Lagrange multipliers and the oracle predictor. It makes explicit the gap between $\widetilde{\Gamma}_{\beta 2DP}$ and Γ^*_β, and justifies the suboptimality (in risk) of the two-step procedure.

Remark 4.1. Assume that for all $k \in [K]$, the random variable $p_k(X,S)$ admits a strictly positive continuous density – that is Assumption 2 – and recall that in this case the oracle thresholds $\bar{F}_{k,s}^{-1}(\beta_k^*)$ with $\beta_k^* = \mathbb{P}(k \in \Gamma^*_\beta(X,S))$ from Proposition 2.5 are uniquely defined. Due to the non-linearity of the quantile operator, the composition of quantiles does not commute, so in general, we have $\beta_k^* \neq \bar{F}_k(G^{-1}(\beta))$. As a consequence, the thresholds used to define the fair correction in the two-step predictor $\widetilde{\Gamma}_{\beta 2DP}$ differ from those of the oracle predictor Γ^*_β.

Despite this potential gap in risk, the size-to-fairness predictor $\widetilde{\Gamma}_{\beta 2DP}$ retains strong advantages: it is easily implementable, requires no constrained optimization, and offers robust constraint satisfaction.

Two-step plug-in predictor. We now construct a plug-in estimator $\widehat{\Gamma}_{\beta 2DP}$ of the size-to-fairness predictor $\widetilde{\Gamma}_{\beta 2DP}$, based on the labeled and unlabeled data. To this end, we consider the same estimators $\widehat{p}_k$ built from the labeled dataset $\mathcal{D}_n$ as in Section 3.1. Then, using the unlabeled dataset $\mathcal{D}_N$, we define the empirical cumulative distribution functions for each $k \in [K]$ and $s \in \mathcal{S}$ as:

$$\widehat{F}_{k,s}(\cdot) = \frac{1}{N_s}\sum_{i=1}^{N} \mathbb{1}_{\{S_i=s, \widehat{p}_k(X_i,S_i) \leq \cdot\}}, \quad \widehat{F}_k = \sum_{s \in \mathcal{S}} \widehat{\pi}_s \widehat{F}_{k,s},$$

where $N_s = \sum_{i=1}^{N} \mathbb{1}_{\{S_i=s\}}$ and $\widehat{\pi}_s = N_s/N$. We denote the associated empirical survival functions by $\widehat{\bar{F}}_{k,s} = 1 - \widehat{F}_{k,s}$ and $\widehat{\bar{F}}_k = 1 - \widehat{F}_k$. We also define the empirical version of the function G as $\widehat{G}(\cdot) = \sum_{k=1}^{K} \widehat{\bar{F}}_k(\cdot)$. With this notation, we define the size-to-fairness plug-in predictor as:

$$\widehat{\Gamma}_{\beta 2DP}(x,s) = \left\{k \in [K] : \widehat{p}_k(x,s) \geq \widehat{\bar{F}}_{k,s}^{-1}\left(\widehat{\bar{F}}_k(\widehat{G}^{-1}(\beta))\right)\right\} \ . \quad (5)$$

4.2 Statistical properties

The previous discussion highlights that, since the population-level predictor $\widetilde{\Gamma}_{\beta 2DP}$ is not necessarily risk-optimal, our main focus lies in assessing whether its plug-in estimator $\widehat{\Gamma}_{\beta 2DP}$ satisfies the desired size and fairness constraints.

Constraint guarantees. We first establish that $\widehat{\Gamma}_{\beta 2DP}$ achieves control over the expected size and demographic parity constraints, up to a deviation of order $1/\sqrt{N}$. Notably, these results are non-asymptotic and distribution-free.

Theorem 4.2. *Let $\widehat{\Gamma}_{\beta 2DP}$ be the empirical predictor defined by Equation (5). There exists a constant $C > 0$ such that*

$$\mathbb{E}\left[\mathcal{U}(\widehat{\Gamma}_{\beta 2DP})\right] \leq \frac{CK}{\sqrt{N}}, \text{ and}$$
$$\mathbb{E}\left[\left|\mathcal{T}\left(\widehat{\Gamma}_{\beta 2DP}\right) - \beta\right|\right] \leq \frac{CK}{\sqrt{N}}.$$

These convergence rates are of the same order as those obtained for the optimal plug-in predictor $\widehat{\Gamma}_\beta$ studied in Section 3.2. In particular, the $1/\sqrt{N}$ rate reflects the statistical error in estimating the cumulative distribution functions $F_{k,s}$ from the unlabeled dataset. Importantly, this means the method benefits from unlabeled data alone, which is advantageous in practice.

5 EXPERIMENTS AND NUMERICAL RESULTS

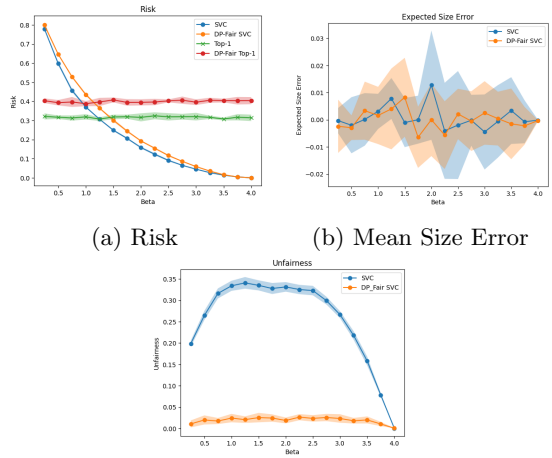
5.1 Implementation

There are two main steps involved in constructing the fair classifiers $\widehat{\Gamma}_\beta$ (referred to as `DP-fair SVC` in the figures) and $\widehat{\Gamma}_{\beta 2DP}$ (referred to as `Two-Step method`): (1) estimating the class-conditional probabilities, and (2) deriving the final classifier using either optimization (solving Equation (4)) or plug-in estimates of quantiles and CDFs.

All experiments are implemented in Python. We use three datasets for each run: a training set composed of n labeled samples and N unlabeled samples, and a test set of T labeled data used solely for evaluation. In the case of real-world data – which only includes labeled observations – we split the data into 80% train / 20% test, and then split the training portion again (50%/50%) to produce D_n and D_N, where labels are dropped from D_N to simulate unlabeled data. The same proportions are applied to synthetic datasets.

Conditional probability estimation. When the probabilities p_k are not available (*i.e.,* except in the idealized synthetic setting), we estimate them using a gradient boosting algorithm (`GradientBoostingClassifier` from `sklearn.ensemble` with 50 estimators). This estimation step uses only the labeled dataset D_n.

Second step. The final classifier is built using only the unlabeled dataset D_N. For $\widehat{\Gamma}_\beta$, we solve Equation (4) using the `L-BFGS-B` optimizer via `scipy.optimize.minimize`. For the two-step method $\widehat{\Gamma}_{\beta 2DP}$, we compute the empirical CDFs and quantiles $\widehat{\widehat{F}}_k$ and $\widehat{\widehat{F}}_k^{-1}$ using `numpy`'s built-in functions.



(a) Risk (b) Mean Size Error

(c) Unfairness

Figure 1: Results on synthetic data with estimated class-conditional probabilities (50 estimators).

5.2 Synthetic Data

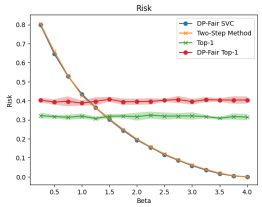
We consider the case of $K = 4$ classes and binary sensitive attributes $s \in \{-1, 1\}$. We generate 10,000 samples from a Gaussian mixture model. The details of the data generation is postponed to the Supplementary material. We explore two scenarios. For the first one, we assume access to the true conditional probabilities p_k — all corresponding plots are deferred to the Supplementary material (see Appendix A). For the second one, probabilities p_k are unknown and must be estimated from data.

Comparison with the Unfair Baseline. Figure 1 shows the results for the second scenario above. Subplot (b) confirms that both fair and unfair classifiers satisfy the size constraint. This aligns with our theoretical guarantees in Theorem 3.1, which match those in (Denis and Hebiri, 2017). Subplot (c) highlights the rise in unfairness when β increases under the unfair classifier – a phenomenon caused by the intrinsic bias amplification of set-valued outputs. In contrast, $\widehat{\Gamma}_\beta$ consistently yields low unfairness, close to zero. Subplot (a) shows that enforcing fairness comes with a slight increase in classification risk. However, given the large fairness improvements (reductions of up to 0.8 in unfairness), this trade-off is acceptable and expected.

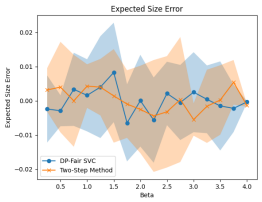
Two-Step vs. Optimizer. We now compare the two fair classifiers $\widehat{\Gamma}_\beta$ (via optimizer) and $\widehat{\Gamma}_{\beta 2DP}$ (two-step), both in terms of performance and runtime. A comparison between both algorithms' properties (time and space complexity) is presented in Appendix A.

Figure 2 shows that the two methods yield comparable performance across risk, size, and unfairness. Although

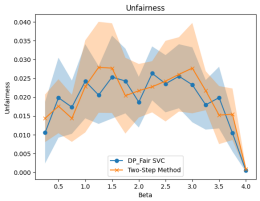
$\widehat{\Gamma}_{\beta 2DP}$ can exhibit slightly higher unfairness, remaining within acceptable bounds. The runtime advantage of the two-step method becomes clear in Figure 3: for large values of K, significantly outperforming the optimizer.



(a) Risk

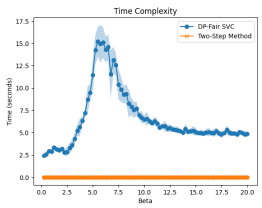


(b) Mean Size Error

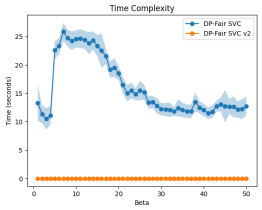


(c) Unfairness

Figure 2: Two-Step vs. Optimizer. Comparison with estimated probabilities (50 estimators).



(a) $K = 20$



(b) $K = 50$

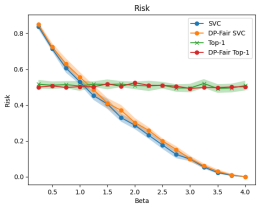
Figure 3: Runtime comparison for increasing K.

5.3 Real Data

We now evaluate our models on the DRUG dataset[1], which contains demographic and personality data for 1,885 individuals, along with drug use behavior. The task is to predict cannabis usage levels. Following (Denis et al., 2024), we reduce the number of classes from 7 to 4: never used, not used in the past year, used in the past year but not today, and used in the past day. The sensitive attribute is binary, indicating whether the respondent has a college degree.

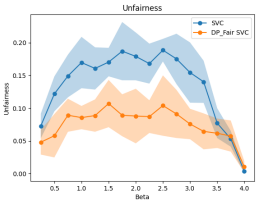
As shown in Figure 4, our DP-fair classifier $\widehat{\Gamma}_\beta$ matches the performance of the unfair baseline in terms of risk and size accuracy, while achieving near-zero unfairness. We also perform our procedure on two additional datasets (ADULT, and CRIME). The results presented

[1] https://www.kaggle.com/datasets/obeykhadija/drug-consumptions-uci



(a) Risk

(b) Mean Size Error



(c) Unfairness

Figure 4: Results on the DRUG dataset (50 estimators, gradient boosting).

in the Supplementary Materials lead to similar conclusion.

Comparison with the top-1 estimation. In Figures 1, 4 and 5 to 7 and their counterparts about the two-step method's performance, we include the top-1 classifier — gradient boosting classifiers with 50 estimators — and a fair analogue. This fair predictor is computed using the procedure in Denis et al. (2024) in the exact fairness case. As expected, we achieve similar risks with a set-valued classifier for $\beta = 1$ and with the top-1 classifier for both the "unfair" and fair pairs.

6 DISCUSSION

Links with the conformal prediction framework. The review (Chzhen et al., 2021) elaborates on the similarities and differences between the conformal prediction setting and the set-valued framework where a size-constraint is imposed. They showed that there is a one-to-one correspondence between the resulting sets-valued at the population level. However, things change as soon as we consider the empirical counterparts of these sets — by plug-in for instance. In particular, a major drawback in the conformal prediction setting is the amount of label data needed to ensure convergence. After the present work, it appears obvious — even though non-trivial — that conformal prediction can also be extended to fairness framework. A standard procedure for building such method consists in splitting the data into training and calibration samples and enforcing the constraints on the latter. Importantly, calibrating coverage requires labeled data, whereas calibrating expected size only requires unlabeled data — see (Chzhen et al., 2021) for more details.

Recent contributions introduce fair conformal prediction (García-Galindo et al., 2025) and (Liu et al., 2022) enforcing risk-wise fairness and dealing with quantile regression respectively. These differ significantly from the present work and appear as complementary works. We conclude this paragraph with two key arguments in favor of our framework compared to conformal prediction. First, our approach provides theoretical guarantees on both the expected size and the fairness constraints, extending existing work on fair conformal methods in multiclass settings. In particular, we establish control of the excess risk of our algorithm, whereas the conformal prediction literature typically does not address the expected size of the resulting set-valued predictor. Second, and maybe the most important point, the parameter β in our setting has a concrete, interpretable meaning, which we leverage to address a major issue in fairness. Indeed, enforcing exact fairness typically comes with a loss in prediction accuracy. By allowing a slightly larger value of β, we can mitigate this degradation in accuracy — *c.f.,* Figure 1(c). The parameter α controlling the error rate in conformal prediction does not allow for such interpretability.

According to the fairness notion. The present contribution extends the set-valued framework to the fairness setting by enforcing *Demographic Parity (DP)*. The choice of DP is motivated by it being the most standard and illustrative notion of fairness — DP is probably the most commonly used notion to introduce fairness constraints in prediction tasks. However, the extension of DP notion to the set-valued framework is an important question. In the present work we propose Definition 2.1 that highlights the probability of each k to belong to the set-valued prediction. A more restrictive DP-fairness could require that the probability of predicting each possible set be identical across groups. This (stronger) DP notion enforces that the distribution of the predictor $\Gamma(X, S)$ is independent of the sensitive attribute. However, this leads to a number of constraints equal to the size of the power set $2^{[K]}$, exponential in the number of classes and therefore not tractable in practice. Definition 2.1 is a less restrictive yet fully meaningful definition of DP for the classification setting. It still guarantees that no label is assigned unfavorably to any sensitive group, which is precisely the fairness requirement that matters in our context.
Beyond DP, one might be interested in considering other notions of fairness such as *Equal opportunity*. We wish to emphasize that our post-processing method is flexible: the methodology could — modulo some technical adjustments — be extended to other fairness constraints, provided that a closed-form expression of the optimal solution is available. One of the main strengths of our contribution is not tied specifically to the DP constraint, but rather to showing that fair set-valued classification is achievable with clear theoretical guarantees, and that the set-valued framework is particularly well-suited for enforcing exact fairness.

7 CONCLUSION

In this work, we introduce a novel framework for learning *fair set-valued classifiers* under Demographic Parity constraints. Unlike standard multi-class predictors, our approach outputs subsets of labels, thereby enabling a flexible control of prediction uncertainty while enforcing fairness across sensitive groups. We characterize the optimal trade-off between accuracy, fairness, and output size via an oracle construction, and propose two practical algorithms: a plug-in estimator based on constrained optimization, and a computationally efficient two-step correction procedure. Both methods rely solely on unlabeled data for enforcing fairness, making them appealing for real-world applications where labeling might be expensive or sensitive.
Our framework offers an alternative to popular fairness relaxations such as ε-fairness, where the choice of the tolerance parameter ε is often arbitrary and lacks interpretability. In contrast, the set-valued formulation enables a direct and meaningful control of the predictor's output size, which provides both interpretability and tunability from a practitioner's perspective. This makes our approach a compelling and principled substitute for unconstrained or approximately constrained fairness objectives.

Beyond empirical performance and constraint guarantees, the set-valued perspective opens promising research directions. In particular, future work could investigate how to extend fairness-aware prediction to structured output problems, such as hierarchical classification. Moreover, extending the fairness constraint to more general criteria (such as equalized odds or individual fairness) in the set-valued prediction setting is an open challenge.

Overall, our results suggest that fair set-valued prediction is a versatile and powerful tool for bridging the gap between predictive performance, fairness, and interpretability.

References

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

G. Alves, F. Bernier, M. Couceiro, K. Makhlouf, C. Palamidessi, and S. Zhioua. Survey on fairness notions and related tensions. *EURO Journal on Decision Processes*, 11:100033, January 2023.

A. Amato, J. R. Osterrieder, and M. R. Machado. How can artificial intelligence help customer intelligence for credit portfolio management? a systematic literature review. *International Journal of Information Management Data Insights*, 4(2):100234, November 2024.

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

P. Besse, C. Castets-Renard, A. Garivier, and J.M. Loubes. Can everyday ai be ethical. fairness of machine learning algorithms. *ArXiv*, abs/1810.01729, 2018.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, page 13–18, December 2009a.

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

J. Cameron, J. Cheong, M. Spitale, and H. Gunes. Multimodal Gender Fairness in Depression Prediction: Insights on Data from the USA & China . In *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Los Alamitos, CA, USA, September 2024. IEEE Computer Society.

P. Celard, E. L. Iglesias, J. M. Sorribes-Fdez, R. Romero, A. Seara Vieira, and L. Borrajo. A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. *Neural Computing and Applications*, 35(3): 2291–2323, 2023.

R. Chen, J. Wang, D. Williamson, T. Chen, J. Lipkova, M. Lu, S. Sahai, and F. Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6): 719–742, 2023.

Z. Chen. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*, 10(1):1–12, 2023.

E. Chzhen and N. Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, 2022.

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, and T. Lorieul. Set-valued classification – overview via a unified framework. *arXiv:2102.12318*, 2021. URL http://arxiv.org/abs/2102.12318.

C. Denis and M. Hebiri. Confidence sets with expected sizes for multiclass classification. *Journal of Machine Learning Research*, 18(102):1–28, 2017.

C. Denis, R. Elie, M. Hebiri, and F. Hu. Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research*, 25 (130):1–46, 2024.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

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

Alberto García-Galindo, Marcos López-De-Castro, and Rubén Armañanzas. Fair prediction sets through multi-objective hyperparameter optimization. *Machine Learning*, 114(1):27, January 2025. ISSN 1573-0565. doi: 10.1007/s10994-024-06721-w.

S. Gaucher, N. Schreuder, and E. Chzhen. Fair learning with wasserstein barycenters for non-decomposable performance measures. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

I. Gibbs, J. Cherian, and E. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2023.

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

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Radu Herbei and Marten H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 34(4): 709–721, 2006. ISSN 0319-5724.

Z. Hobson, J. A. Yesberg, B. Bradford, and J. Jackson. Artificial fairness? trust in algorithmic police decision-making. *Journal of Experimental Criminology*, 19(1):165–189, 2023.

R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. *Uncertainty in Artificial Intelligence Conference*, 2020.

M. Lapin, M. Hein, and B. Schiele. Loss functions for top-k error: Analysis and insights. In *CVPR*, 2016.

M. Liu, L. Ding, D. Yu, W. Liu, L. Kong, and B. Jiang. Conformalized fairness via quantile regression. *NeurIPS*, 2022.

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

Y. Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, 88: 10–11, 2012.

J. Paulus and D. Kent. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *npj Digital Medicine*, 3(1):1–8, 2020.

Y. Romano, E. Patterson, and E. Candès. Conformalized quantile regression. *NeurIPS*, 2019.

M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114 (525):223–234, 2019.

H. Sadok, F. Sakka, and M. El Maknouzi. Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance*, 10(1):2023262, 2022.

I. Taylor. Justice by algorithm: The limits of ai in criminal sentencing. *Criminal Justice Ethics*, 42(3): 193–213, 2023.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

X. Wang, Y. Zhang, and R. Zhu. A brief review on algorithmic fairness. *Management System Engineering*, 1(1):7, 2022.

J. Yang, A. Soltan, D. Eyre, and D. Clifton. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nature Machine Intelligence*, 5(8):884–894, 2023.

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

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

Y. Zhou and M. Sesia. Conformal classification with equalized coverage for adaptively selected groups. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, 2025.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] In Section 2 we introduce the setting and assumptions, in Sections 3 and 4 we present the procedures to implement our approaches, and in Section 5 we detail the implementation, model and data used.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] Provided in the supplementary material in Appendix A.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The source code is provided in the supplemental material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes] The proofs are provided in the supplementary material.
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] A notebook is provided as supplemental material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Set to Be Fair: Demographic Parity Constraints for Set-Valued Classification: Supplementary Materials

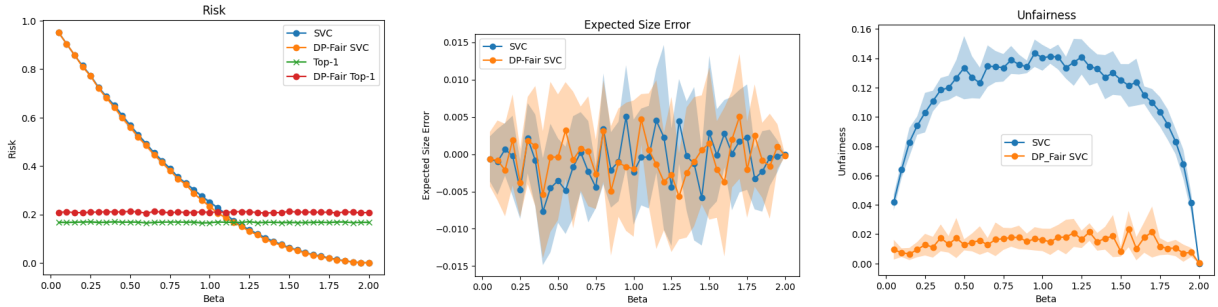
This appendix consists mainly in two parts. A first one is devoted to additional numerical results (Appendix A). In a second part (Appendix B), we gather the proof of our results.

A ADDITIONAL NUMERICAL CONSIDERATIONS

A.1 Additional results on real datasets.

We first present the result for the dataset ADULT which is set of 32561 individuals where the task is to predict if they earn more or less than $50k$ a year using census data consisting of 14 features. We use the gender as sensitive attribute. While this is a binary classification setting, it is still interesting to note that getting close to $\beta = 0$ or to $\beta = 2$ is similar to allowing the model to "refuse to answer" similarly as in Herbei and Wegkamp (2006) which also explains the symmetry around 1 in terms of unfairness. The obtained results are displayed in Figure 5, and exhibit similar results as for DRUG dataset (see Figure 4).

We also apply our procedure to the CRIME dataset which is a set of 1994 american cities, where the task is to predict the amount of violent crime per capita — separated in $K = 5$ classes, with the 0.2, 0.4, 0.6 and 0.8 quantiles as separations — using 123 features — after removing the state, county, community, communityname and fold features since they act as identifiers — with the sensitive attribute indicating whether the neighbourhood has a majorly white population. The results are displayed in Figure 6. Here again, we observe similar performance as for the other datasets. In particular, the set-valued classifier achieves the same risk as the the top-1 classifier in both the "unfair" and fair case, and the fair set-valued classifier allows us to do as good as the "unfair" top-1 classifier by choosing $\beta = 1.25$, illustrating how fairness can be maintained by relaxing the size constraint by a relatively low amount. We can notice this pattern being approximately true for other values of β, where the DP-fair set-valued classifier performs as well as the "unfair" one by raising the value of β by less than 0.25.

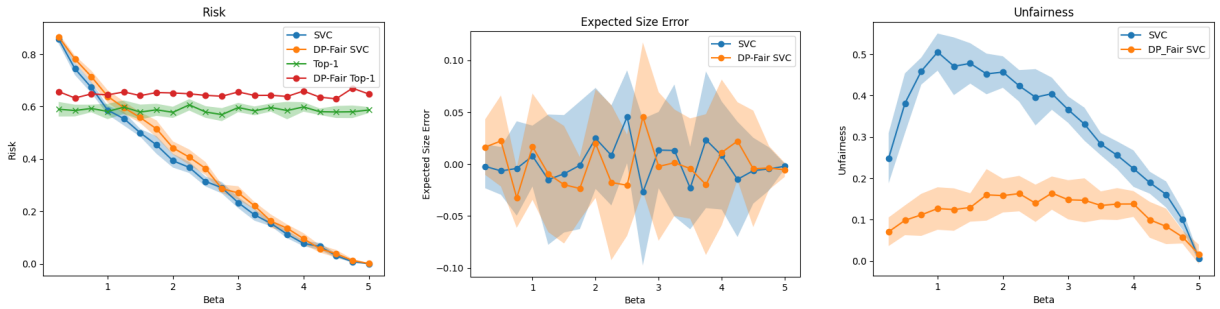


(a) Risk (b) Mean Size Error (c) Unfairness

Figure 5: Results on the ADULT dataset (50 estimators, gradient boosting).

A.2 Computational considerations.

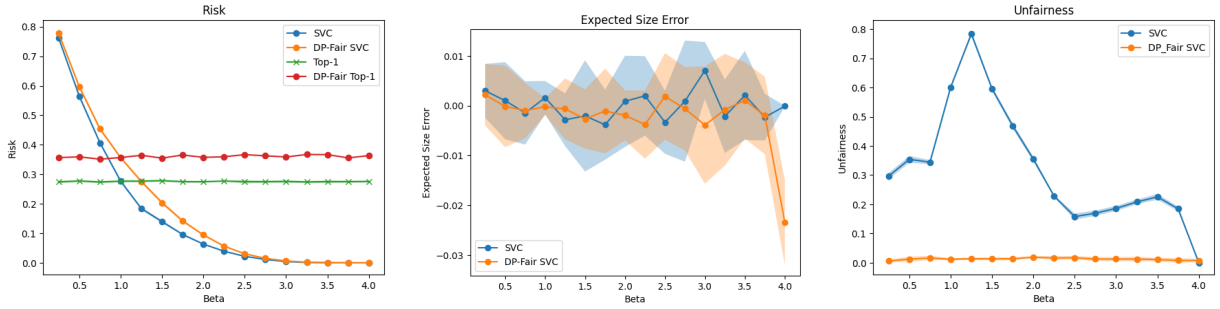
The two-step approach $\widehat{\Gamma}_{\beta 2DP}$ introduced in Section 4 does not benefit from the optimality guarantees in terms of risk. However, it offers strong computational advantages over the estimator $\widehat{\Gamma}_\beta$ presented in Section 3. Both procedures rely on preliminary estimators $(\widehat{p}_k(X_i, S_i))_{i=1}^N$, so we do not include the cost of this estimation step in the complexity analysis. Let M denote the cost of one step of the numerical optimizer and T the number



(a) Risk (b) Mean Size Error (c) Unfairness

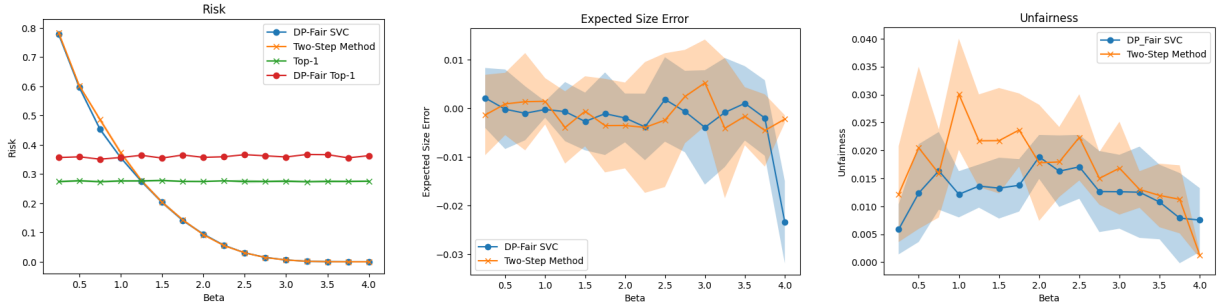
Figure 6: Results on the CRIME dataset (50 estimators, gradient boosting).

of iterations needed for convergence. Then: i) the plug-in estimator $\widehat{\Gamma}_\beta$ has overall time complexity of order $\mathcal{O}(MTK|\mathcal{S}|N)$; ii) in contrast, the two-step method $\widehat{\Gamma}_{\beta 2DP}$ requires only $\mathcal{O}(K\,|S|\,N)$ operations, as it reduces to empirical quantile computations. Furthermore, the constant M can in practice grow with K, $|\mathcal{S}|$, N, and even the target size level β (observed empirically for the latter). For instance, using the `BFGS` optimization algorithm, one may encounter complexities of order at least $\mathcal{O}(T(K\,|\mathcal{S}|)^3 N)$. Faster algorithms can be considered – *e.g.* by using smoothing techniques (Nesterov, 2012)) – however, even by accelerating the solving step, the time complexity cannot drop below the two-step procedure's one. Empirically, we observe that the two-step method scales significantly better with the number of classes. On a toy example, increasing K results in a clear speed-up for $\widehat{\Gamma}_{\beta 2DP}$ compared to $\widehat{\Gamma}_\beta$ (see Figure 3 in Section 5). This makes the two-step predictor a promising alternative for large-scale applications.



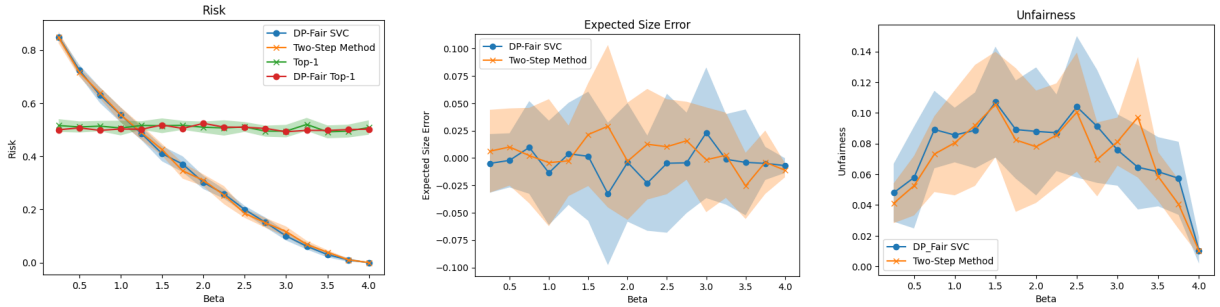
(a) Risk (b) Mean Size Error (c) Unfairness

Figure 7: Results on synthetic data using the true conditional distributions.



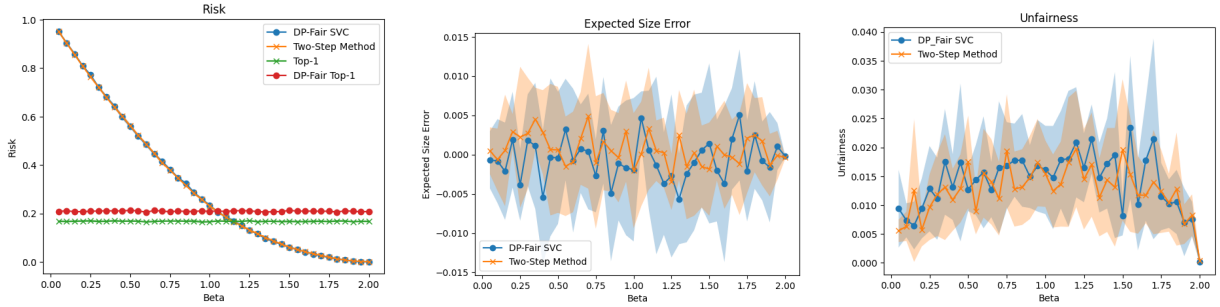
(a) Risk (b) Mean Size Error (c) Unfairness

Figure 8: Comparison between optimizer-based and two-step fair classifiers (true probabilities).



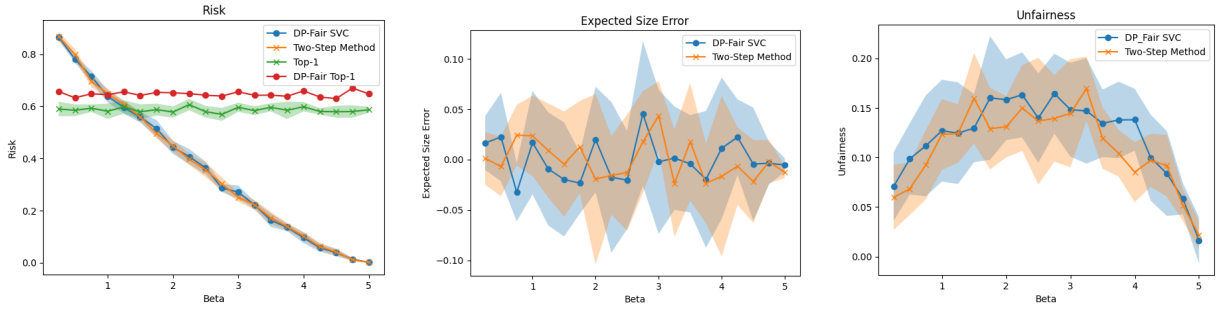
(a) Risk (b) Mean Size Error (c) Unfairness

Figure 9: Two-step vs optimizer on the DRUG dataset.



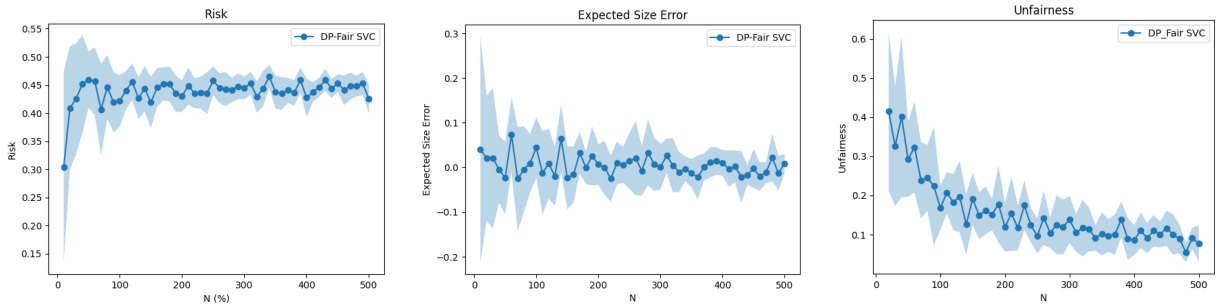
(a) Risk (b) Mean Size Error (c) Unfairness

Figure 10: Two-step vs optimizer on the ADULT dataset.



(a) Risk (b) Mean Size Error (c) Unfairness

Figure 11: Two-step vs optimizer on the CRIME dataset.



(a) Risk (b) Mean Size Error (c) Unfairness

Figure 12: Performances in relation to N.

A.3 Storage considerations.

For both methods, we need to store $K \times |\mathcal{S}|$ cumulative distribution functions as well as $\mathcal{O}(K \times |\mathcal{S}|)$ constants. The space complexity is therefore in $\mathcal{O}(K\,|\mathcal{S}|\,C)$ with C the size of a cdf. Apart from these, we need to store the data and the classifier — or estimator of the p_k's, which are not inherent to the algorithms.

A.4 Additional numerical results.

This section provides all plots related to the the performance of the set-valued classifiers when the class-conditional probabilities are known (see Figure 7, and 8). As compared to the case where the class-conditional probabilities are unknown, we observe that estimating p_k leads to minimal performance degradation, validating our theoretical findings in Theorem 3.2.

In addition, Figures 9 to 11 display the two-step set-valued classifier performance as compared to the optimizer-based approach in the case of the real data. The conclusion are similar to the case of the synthetic data: risks, size, and unfairness are comparable for both approaches.

A.5 Performance related to N.

Figure 12 shows the performances of our approach on generated data using the true class membership probabilities and with $\beta = 1.5$, varying the size of the unlabeled dataset used to learn the parameters λ and γ. We observe that for relatively small values of N the procedure already achieves reasonable performances, dividing by more than 4 the unfairness shown in Figure 7 for the "unfair" set-valued classifier with $\beta = 1.5$, with $N = 200$.

B PROOFS

This appendix is organised as follows: in Section B.1, we state important technical tools that will be used in the main proof section. The remaining of this appendix is devoted to the proofs of the results in Sections 2 to 4 respectively.

We also introduce $\Delta = \{(\lambda, \gamma) \in \mathbb{R}^{K|\mathcal{S}|+1},\ \ \lambda \geq 0,\ ; \sum_{s\in\mathcal{S}} \gamma_{k,s} = 0\}$ the set of parameters of interests. Throughout the proofs, we use the following notation

$$\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{p}_{k,s} \geq \cdot\right) = \frac{1}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\{\widehat{p}_k(X_i,s)\geq\cdot\}}.$$

B.1 Technical Results

Lemma B.1. *Let $Z_1, \ldots, Z_N$ i.i.d random variable with continuous distribution function. Let us denote by $\widehat{F}$ the empirical cumulative distribution function. For each $u \in (0,1)$, we have*

$$0 \leq \widehat{F}(\widehat{F}^{-1}(u)) - u \leq \frac{1}{N}, \quad a.s.$$

Proof. Let σ be the ordering permutation ensuring $Z_{\sigma(i)} < Z_{\sigma(i+1)}$ almost surely for $i \in [N-1]$. Assume that for some $i \in \{2, \dots, N\}$, $u \in [\hat{F}(Z_{\sigma(i-1)}), \hat{F}(Z_{\sigma(i)}))$. Therefore $\hat{F}^{-1}(u) = Z_{\sigma(i)}$ and then $\hat{F}\left(\hat{F}^{-1}(u)\right) = \hat{F}(Z_{\sigma(i)}) = \frac{i}{N}$. By subtracting u, thanks to the continuity assumption, we get

$$\begin{aligned} 0 = \hat{F}(Z_{\sigma(i)}) - \hat{F}(Z_{\sigma(i)}) &\leq \hat{F}\left(\hat{F}^{-1}(u)\right) - u \\ &= \hat{F}(Z_{\sigma(i)}) - u \\ &\leq \hat{F}(Z_{\sigma(i)}) - \hat{F}(Z_{\sigma(i-1)}) = \frac{1}{N} \ \ a.s. \end{aligned}$$

And for $u \in (0, \widehat{F}(Z_{\sigma(1)}))$, similar reasoning holds with $\hat{F}^{-1}(u) = Z_{\sigma(1)}$. □

Lemma B.2. *Let Z a random variable distributed according to a Binomial distribution with parameter N, p. We have that*

$$\mathbb{E}\left[\frac{\mathbb{1}_{\{Z>0\}}}{Z}\right] \leq \frac{2}{(N+1)p} \ .$$

Lemma B.3. *(Proposition 1 of Denis and Hebiri (2017)) Under Assumption 1, the following properties hold:*

- $\forall t \in (0,1), \beta \in (0,K) : G^{-1}(\beta) \leq t \iff \beta \geq G(t)$,
- $\forall \beta \in (0,K), G(G^{-1}(\beta)) = \beta$.

Lemma B.4. *(Dvoretzky-Kiefer-Wolfowitz Inequality) Let $Z_1, \dots, Z_N$ i.i.d. real-valued random variable with common distribution function F. We denote by $\widehat{F}_N$ the corresponding empirical cumulative distribution function. The following holds*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|\widehat{F}_N(x) - F(x)\right| \geq \varepsilon\right) \leq 2\exp(-2n\varepsilon^2), \qquad \forall \varepsilon > 0 \ .$$

Lemma B.5. *Let $\beta > 0$, and let us define the function H that maps Δ onto $\mathbb{R}_+$ defined as follows*

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

The function H is convex and coercive.

Proof. The convexity of H is straightforward. Let us focus on the second point of the lemma. Consider $(\lambda_m, \gamma_m)_{m\geq 0} \in \Delta$ with ℓ_2 norm $\|(\lambda_m, \gamma_m)\| \to +\infty$. We observe that if $\lambda_m \to +\infty$, and since

$$H(\lambda, \gamma) \geq \lambda\beta \ ,$$

we have $H(\lambda_m, \gamma_m) \to +\infty$ as $m \to +\infty$. Otherwise, $\|\gamma_m\| \to +\infty$. In this case since for each $k \in [K]$ we have the condition $\sum_{s\in\mathcal{S}} \gamma_{k,s} = 0$, we deduce that there exists $k_0 \in [K]$, and $s_0 \in \mathcal{S}$ such that

$$\gamma^m_{k_0,s_0} \to -\infty \ \text{ as } \ m \to +\infty \ .$$

Therefore, we have, with $m \to +\infty$, that

$$H(\lambda_m, \gamma_m) \geq \mathbb{E}_{X|S=s}\left[\left(\pi_s\left(p_k(X,s) - \lambda_m\right) - \gamma^m_{k_0,s_0}\right)_+\right] + \mathcal{O}(1) \to +\infty \ ,$$

where $\mathcal{O}(1)$ is a negligible term (bounded by a constant). It then shows that H is coercive. □

B.2 Proof of Section 2

Proof of Theorem 2.2. In both cases, we apply the weak duality principle. More precisely, we first solve the max-min problem

$$\max_{\lambda,\boldsymbol{\alpha}} \min_{\Gamma} \mathcal{L}\left(\Gamma, \lambda, \boldsymbol{\alpha}\right) \ ,$$

where $\mathcal{L}(\Gamma, \lambda, \boldsymbol{\alpha})$ is the Lagrangian associated to the minimization problem. Then, we show that the solution of the max-min is an optimal fair set-valued classifier.

We first write the Lagrangian associated to our minimization problem. For $\lambda, \boldsymbol{\alpha}, \Gamma$, we have that

$$\mathcal{L}(\Gamma, \lambda, \boldsymbol{\alpha}) = R(\Gamma) + \lambda \left(\mathbb{E}_X \left[|\Gamma(X,S)|\right] - \beta\right) + \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \alpha_{k,s} \left(\mathbb{P}_{X|S=s}\left(k \in \Gamma(X,s)\right) - \mathbb{P}_{X,S}\left(k \in \Gamma(X,S)\right)\right) \ .$$

We observe that $\mathcal{L}$ can be expressed as follows

$$\begin{aligned}\mathcal{L}(\Gamma, \lambda, \boldsymbol{\alpha}) = 1 &- \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\mathbb{1}_{\{k\in\Gamma(X,s)\}}\left(\pi_s p_k(X,s)\right)\right] \\ &+ \lambda\left(\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\pi_s \mathbb{1}_{\{k\in\Gamma(X,s)\}}\right] - \beta\right) \\ &+ \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\alpha_{k,s}\mathbb{1}_{\{k\in\Gamma(X,s)\}}\right] - \mathbb{E}_{X|S=s}\left[\bar{\alpha}_k \mathbb{1}_{\{k\in\Gamma(X,s)\}}\pi_s\right] \ ,\end{aligned}$$

$\bar{\alpha}_k = \sum_{s\in\mathcal{S}} \alpha_{k,s}$. Therefore, we get

$$\mathcal{L}(\Gamma, \lambda, \boldsymbol{\alpha}) = 1 - \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\mathbb{1}_{\{k\in\Gamma(X,s)\}}\left(\pi_s\left(p_k(X,s) - \lambda + \bar{\alpha}_k\right) - \alpha_{k,s}\right)\right] - \lambda\beta \ .$$

From the above equation, it is not difficult to see that

$$\Gamma^*_{\lambda,\boldsymbol{\alpha}} \in \arg\min_{\Gamma} \mathcal{L}(\Gamma, \lambda, \boldsymbol{\alpha}) \ , \tag{6}$$

is also characterized pointwise as

$$\Gamma^*_{\lambda,\boldsymbol{\alpha}}(x,s) = \left\{k \in [K], \ \ p_k(x,s) \geq \lambda + \frac{\alpha_{k,s}}{\pi_s} - \bar{\alpha}_k\right\} \ .$$

Furthermore, injecting to the Lagrangian we have that

$$\mathcal{L}\left(\Gamma^*_{\lambda,\boldsymbol{\alpha}}, \lambda, \boldsymbol{\alpha}\right) = 1 - \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\left(\pi_s\left(p_k(X,s) - \lambda + \bar{\alpha}_k\right) - \alpha_{k,s}\right)_+\right] - \lambda\beta \ .$$

Next, it remains to optimize in $(\lambda, \boldsymbol{\alpha})$. We have that

$$(\lambda^*, \boldsymbol{\alpha}^*) \in \arg\max_{\lambda,\boldsymbol{\alpha}} \mathcal{L}\left(\Gamma^*_{\lambda,\boldsymbol{\alpha}}, \lambda, \boldsymbol{\alpha}\right) \ , \tag{7}$$

is characterized as

$$(\lambda^*, \boldsymbol{\alpha}^*) \in \arg\min_{\substack{(\lambda,\alpha)\in\mathbb{R}^{K|\mathcal{S}|+1} \\ \lambda\geq 0}} \widetilde{H}(\lambda, \alpha) \ ,$$

with $\widetilde{H}(\lambda, \boldsymbol{\alpha}) = \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\left(\pi_s\left(p_k(X,s) - \lambda + \bar{\alpha}_k\right) - \alpha_{k,s}\right)_+\right] + \lambda\beta$. We observe that the above minimization problem can be reformulated as follows

$$(\lambda^*, \gamma^*) \in \arg\min_{\substack{(\lambda,\gamma)\in\mathbb{R}^{K|\mathcal{S}|+1} \\ \lambda\geq 0 \\ \sum_{s\in\mathcal{S}}\gamma_{k,s}=0}} \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\left(\pi_s\left(p_k(X,s) - \lambda\right) - \gamma_{k,s}\right)_+\right] + \lambda\beta \ , \tag{8}$$

with the introduced reparameterization $\gamma_{k,s} = \alpha_{k,s} - \pi_s \sum_s \alpha_{k,s}$. Let us then denote by H the objective function defined as

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

From Lemma B.5, H is convex and coercive. Therefore there exists a global minimizer $(\lambda^*, \boldsymbol{\gamma}^*)$ that belongs to a compact subset of Δ. Therefore, it implies that the function $\widetilde{H}$ admits also a global minimizer $(\lambda^*, \boldsymbol{\alpha}^*)$. Furthermore, thanks to Assumption 1, the function $\widetilde{H}$ is differentiable *w.r.t.* $(\lambda, \boldsymbol{\gamma})$. Therefore, we deduce from the first order condition that $\mathbf{0} \in \partial \widetilde{H}(\lambda^*, \boldsymbol{\gamma}^*)$. We then have that

$$\partial_\lambda \widetilde{H} \left(\lambda^*, \boldsymbol{\alpha}^* \right) = - \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \pi_s \mathbb{P}_{X|S=s} \left(p_k(X,s) \geq \lambda^* + \frac{\alpha^*_{k,s} - \pi_s \bar{\alpha}^*_k}{\pi_s} \right) + \beta = 0 \ ,$$

which means that Γ^*_β as expected size β. Furthermore

$$\partial_{\alpha_{k,s}} \widetilde{H} \left(\lambda^*, \boldsymbol{\alpha}^* \right) = \sum_{s' \in \mathcal{S}} \pi_{s'} \mathbb{P}_{X|S=s'} \left(p_k(X,s') \geq \lambda + \frac{\alpha^*_{k,s'} - \pi_{s'} \bar{\alpha}^*_k}{\pi_{s'}} \right) - \mathbb{P}_{X|S=s} \left(p_k(X,s) \geq \lambda^* + \frac{\alpha^*_{k,s} - \pi_s \bar{\alpha}^*_k}{\pi_s} \right) = 0 \ .$$

This means that the set-valued classifier Γ^*_β defined, with the reparameterization $\gamma^*_{k,s} = \alpha^*_{k,s} - \pi_s \bar{\alpha}^*_k$ given by

$$\Gamma^*_\beta(x,s) = \left\{ k \in [K], \ \ p_k(x,s) \geq \lambda^* + \frac{\gamma^*_{k,s}}{\pi_s} \right\} \ , \tag{9}$$

satisfies the demographic parity constraint. To conclude the proof we observe that Γ^*_β satisfies

$$\Gamma^*_\beta \in \arg\min_{\Gamma} \mathcal{L} \left(\Gamma, \lambda^*, \boldsymbol{\alpha}^* \right) \ . \tag{10}$$

To this end we consider Γ another set-valued classifier that satisfies the demographic parity constraint and $\mathcal{T}(\Gamma) \leq \beta$. We have that

$$\begin{aligned} \mathcal{L} \left(\Gamma^*_\beta, \lambda^*, \boldsymbol{\alpha}^* \right) &= R(\Gamma^*_\beta) \leq \mathcal{L} \left(\Gamma, \lambda^*, \boldsymbol{\alpha}^* \right) \\ &= R(\Gamma) + \underbrace{\lambda^*}_{\geq 0} \underbrace{\left(\mathcal{T}(\Gamma) - \beta \right)}_{\leq 0} + \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \alpha^*_{k,s} \underbrace{\left(\mathbb{P}_{X|S=s} \left(k \in \Gamma(X,s) \right) - \mathbb{P}_{X,S} \left(k \in \Gamma(X,S) \right) \right)}_{=0} \\ &\leq R(\Gamma) \ , \end{aligned}$$

which yields the result.

□

Proof of Proposition 2.3. The proof of the proposition can be easily deduced from the proof of Theorem 2.2. The first two points of the proposition are already shown in the proof of Theorem 2.2. Let us now proof the last point. Let $(\lambda^*, \boldsymbol{\alpha}^*)$ the Lagrangian parameters defined as in Equation (7). Hence, we have that $(\lambda^*, \boldsymbol{\gamma}^*)$ is a minimizer of Equation (8) with, for each $k \in [K], s \in \mathcal{S}$, $\gamma^*_{k,s} = \alpha^*_{k,s} - \pi_s \bar{\alpha}^*_k$. Furthermore, we observe that for each set-valued classifier Γ

$$\mathcal{L} \left(\Gamma, \lambda^*, \boldsymbol{\alpha}^* \right) = \mathcal{R}_{\lambda^*, \boldsymbol{\gamma}^*}(\Gamma) \ ,$$

which yields the result thanks to the characterization of Γ^*_β (see Equation (10)). □

Proof of Proposition 2.4. Let Γ a set-valued classifier, and $(\lambda^*, \boldsymbol{\gamma}^*)$ defined by Equation (3). We have that

$$\mathcal{R}_{\lambda^*, \boldsymbol{\gamma}^*} \left(\Gamma \right) = R(\Gamma) + \lambda^* \left(\mathbb{E}_X \left[|\Gamma(X,S)| \right] - \beta \right) + \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \gamma^*_{k,s} \mathbb{P}_{X|S=s} \left(k \in \Gamma(X,s) \right) .$$

Since

$$\mathbb{E}_X\left[|\Gamma(X,S)|\right] = \sum_{k\in[K]}\sum_{s\in\mathcal{S}} \pi_s \mathbb{P}_{X|S=s}\left(k\in\Gamma(X,S)\right),$$

and

$$1-R(\Gamma) = \sum_{k\in[K]}\sum_{s\in\mathcal{S}} \pi_s \mathbb{E}_{X|S=s}\left[\mathbb{1}_{\{k\in\Gamma(X,S)\}} p_k(X,S)\right],$$

we deduce that

$$\begin{aligned}\mathcal{R}_{\lambda^*,\gamma^*}(\Gamma) - \mathcal{R}_{\lambda^*,\gamma^*}\left(\Gamma^*_\beta\right) \\ = \sum_{k\in[K]}\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\left(\pi_s\left(p_k(X,S)-\lambda\right)+\gamma^*_{k,s}\right)\left(\mathbb{1}_{\left\{k\in\Gamma^*_\beta(X,S)\right\}} - \mathbb{1}_{\{k\in\Gamma\}}\right)\right].\end{aligned}$$

Now, by definition of Γ^*_β (see Equation (9)), we observe that

$$\pi_s\left(p_k(X,S)-\lambda\right)+\gamma^*_{k,s} \geq 0 \text{ iff } \mathbb{1}_{\left\{k\in\Gamma^*_\beta(X,S)\right\}} = 1.$$

In view of this observation, we get the desired result. □

Proof of Proposition 2.5. We start with the first point of the proposition. Assume that there exists $(\widetilde{\lambda},\widetilde{\gamma})$ such that for each $s\in\mathcal{S}$

$$\Gamma^*_\beta(x,s) = \left\{k\in[K],\ \ p_k(x,s) \geq \widetilde{\lambda} + \frac{\widetilde{\gamma}_{k,s}}{\pi_s}\right\},$$

with for $k\in[K]$, $\sum_{s\in\mathcal{S}}\widetilde{\gamma}_{k,s}=0$. Under Assumption 2, we have that for each (k,s)

$$\mathbb{P}_{X|S=s}\left(p_k(X,S)\geq\lambda^*+\frac{\gamma^*_{k,s}}{\pi_s}\right) = \mathbb{P}_{X|S=s}\left(p_k(X,S)\geq\widetilde{\lambda}+\frac{\widetilde{\gamma}_{k,s}}{\pi_s}\right) \text{ iff } \lambda^*+\frac{\gamma^*_{k,s}}{\pi_s} = \widetilde{\lambda}+\frac{\widetilde{\gamma}_{k,s}}{\pi_s}.$$

Since $\sum_{s\in\mathcal{S}}\pi_s=1$, and $\sum_{s\in\mathcal{S}}\gamma^*_{k,s}=\sum_{s\in\mathcal{S}}\widetilde{\gamma}_{k,s}=0$, we deduce from the above equality that

$$\lambda^* = \widetilde{\lambda},$$

and then $\widetilde{\gamma}_{k,s}=\gamma^*_{k,s}$ for each $(k,s)\in[K]\times\mathcal{S}$.

For the second point, we observe since Γ^*_β satisfies the expected size constraint and the Demographic parity constraint, we deduce that for each $(k,s)\in[K]\times\mathcal{S}$

$$\mathbb{P}\left(p_k(X,S)\geq\lambda^*+\frac{\gamma^*_{k,s}}{\pi_s}\right) = \mathbb{P}\left(k\in p_k(X,S)\right) := \beta^*_k.$$

Therefore, under Assumption 2, we have that

$$\lambda^*+\frac{\gamma^*_{k,s}}{\pi_s} = F^{-1}_{k,s}(\beta^*_k).$$

□

B.3 Proof of Section 3

We start this section with the following lemma.

Lemma B.6. *Let $\widehat{H}$ the function each $(\lambda,\gamma)\in\Delta$*

$$\widehat{H}(\lambda,\gamma) = \sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\frac{1}{N_s}\sum_{i\in\mathcal{D}_{N_s}}\left(\widehat{\pi}_s\left(\widehat{p}_k(X_i,s)-\lambda\right)-\gamma_{k,s}\right)_+ + \lambda\beta.$$

The function $\widehat{H}$ is convex, coercive, and its subgradient is as follows

$$h_\lambda \in \partial_\lambda \widehat{H}(\lambda,\gamma) \text{ iff } \exists \mu \in [0,1],\ \ h_\lambda = -\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \frac{\widehat{\pi}_s}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)>\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}} - \mu \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \frac{\widehat{\pi}_s}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)=\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}} + \beta,$$

and for each $k \in [K]$, and $s \in \mathcal{S}$,

$$h_{\gamma_{k,s}} \in \partial_{\gamma_{k,s}} \widehat{H}(\lambda,\gamma) \text{ iff } \exists \sigma_{k,s} \in [0,1],\ \ h_{\gamma_{k,s}} = -\frac{1}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)>\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}} - \sigma_{k,s}\frac{1}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)=\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}}.$$

Lemma B.7. *For each $k \in [K]$, and $s \in \mathcal{S}$, there exists $C > 0$ such that*

$$\frac{1}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)=\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}} \leq \frac{C}{N_s} \qquad a.s.$$

Proof of Theorem 3.1. We first start with the following decomposition

$$\mathcal{T}\left(\widehat{\Gamma}\right) - \beta = \mathbb{E}\left[\left|\widehat{\Gamma}(X,S)\right|\right] - \beta = \sum_{k\in[K]}\sum_{s\in\mathcal{S}} \pi_s \mathbb{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right) - \beta$$
$$\sum_{k\in[K]}\sum_{s\in\mathcal{S}} (\pi_s - \widehat{\pi}_s)\, \mathbb{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right) + \sum_{k\in[K]}\sum_{s\in\mathcal{S}} \widehat{\pi}_s \mathbb{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right) - \beta \quad (11)$$

The first term in the *r.h.s.* of the above equation can be bounded as follows

$$\left|\sum_{k\in[K]}\sum_{s\in\mathcal{S}} (\pi_s - \widehat{\pi}_s)\, \mathbb{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right)\right| \leq K\,|\mathcal{S}|\max_{s\in\mathcal{S}} |\widehat{\pi}_s - \pi_s|\,. \quad (12)$$

For the second term, conditional on D_n, since $\widehat{p}_k$ satisfies similar assumption as Assumption 1, we observe that

$$\sum_{k\in[K]}\sum_{s\in\mathcal{S}} \widehat{\pi}_s \mathbb{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right) - \beta =$$
$$\sum_{k\in[K]}\sum_{s\in\mathcal{S}} \pi_s \left(\mathbb{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right) - \widehat{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right)\right)$$
$$+ \sum_{k\in[K]}\sum_{s\in\mathcal{S}} \widehat{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right).$$

Now, we observe that from Lemma B.6, since $\widehat{H}$ is convex and coercive, $(\widehat{\lambda},\widehat{\gamma}$ is a global minimizer that satisfies the first order condition. Therefore, $0 \in \partial_\lambda(\lambda^*,\gamma^*)$. Hence, there exists $\mu \in [0,1]$ such that

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \frac{\widehat{\pi}_s}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)>\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}} = \beta - \mu \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \frac{\widehat{\pi}_s}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)=\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}}.$$

Then from Equation 11 and 12, we deduce that on the event $\{N_s \geq 1\}$

$$\left|\mathcal{T}(\widehat{\Gamma}) - \beta\right| \leq K\,|\mathcal{S}| \max_{s\in\mathcal{S}} |\widehat{\pi}_s - \pi_s| + \sum_{k\in[K]}\sum_{s\in\mathcal{S}} \pi_s \left|\mathbb{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right) - \widehat{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right)\right| \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \frac{\widehat{\pi}_s}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)=\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}}.$$

Therefore, from Lemma B.7 and since $\pi_s \in (0,1)$, we deduce that *a.s.*

$$\left|\mathcal{T}(\widehat{\Gamma}) - \beta\right| \leq K\,|\mathcal{S}| \max_{s\in\mathcal{S}} |\widehat{\pi}_s - \pi_s| + \sum_{k,s} \sup_{t\in[0,1]} \left|\widehat{F}_{k,s}(t) - \widehat{F}_{k,s}(t)\right| + \frac{C}{\min_{s\in\mathcal{S}} N_s}. \tag{13}$$

Applying Lemma B.4, we have that conditional D_n and N_s, on the event $\{N_s \geq 1\}$, we have that

$$\mathbb{E}\left[\sup_{t\in[0,1]} \left|\widehat{F}_{k,s}(t) - \widehat{F}_{k,s}(t)\right|\right] \leq \frac{C}{\sqrt{\min_{s\in\mathcal{S}} N_s}}.$$

Therefore, from the above inequality and Equation 13, we deduce that

$$\mathbb{E}\left[\left|\mathcal{T}(\widehat{\Gamma}) - \beta\right|\right] = \mathbb{E}\left[\left|\mathcal{T}(\widehat{\Gamma}) - \beta\right| \mathbb{1}_{\{N_s\geq 1\}} + \mathbb{1}_{\{N_s=0\}}\right] \leq K\,|\mathcal{S}| \max_{s\in\mathcal{S}} \left(\mathbb{E}\left[|\widehat{\pi}_s - \pi_s|\right] + \mathbb{E}\left[\frac{C\mathbb{1}_{\{N_{\min}\geq 1\}}}{\sqrt{N_{\min}}}\right]\right) + \mathbb{E}\left[\frac{C}{N_{\min}}\right] + 2K\mathbb{P}\left(N_{\min} = 0\right).$$

Let us deal now with the proof of the unfairness bound that follows the same lines than for expected size. Let $k \in [K], s, s' \in \mathcal{S}$. The following hold

$$\left|\mathbb{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right) - \mathbb{P}_{X|S=s'}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s'}}{\widehat{\pi}_{s'}}\right)\right| \leq \sup_{t\in\mathbb{R}} \left|F_{k,s}(t) - \widehat{F}_{k,s}(t)\right| + \sup_{t\in\mathbb{R}} \left|F_{k,s'}(t) - \widehat{F}_{k,s'}(t)\right| + \left|\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right) - \widehat{\mathbb{P}}_{X|S=s'}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s'}}{\widehat{\pi}_{s'}}\right)\right|. \tag{14}$$

We consider the function $\widehat{H}$ defined in Lemma B.6. Since, we minimize this function in γ under the constraints that $\sum_{s\in\mathcal{S}} \gamma_{k,s} = 0$ for each $k \in [K]$. we deduce from the KKT conditions that for each $k \in [K], s \in \mathcal{S}$ there exists $\widehat{\nu}_{k\in[K]} \in \mathbb{R}^K$ such that

$$0 \in \partial_{\gamma_{k,s}} \widehat{H}(\widehat{\lambda}, \widehat{\gamma}) + \widehat{\nu}_k, \quad \text{with} \quad \sum_{s\in\mathcal{S}} \widehat{\gamma}_{k,s} = 0.$$

Therefore, from Lemma B.6, we obtain that there exists $\sigma_{k,s} \in [0,1]$ such that

$$0 = -\frac{1}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)>\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}} - \sigma_{k,s}\frac{1}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)=\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}} + \widehat{\nu}_k,$$

that implies that for each $s \neq s' \in \mathcal{S}$

$$\frac{1}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)>\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}} - \frac{1}{N_{s'}} \sum_{i\in\mathcal{D}_{N_{s'}}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s')>\lambda+\frac{\gamma_{k,s'}}{\widehat{\pi}_{s'}}\right\}} = \sigma_{k,s'}\frac{1}{N_{s'}} \sum_{i\in\mathcal{D}_{N_{s'}}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s')=\lambda+\frac{\gamma_{k,s'}}{\widehat{\pi}_{s'}}\right\}} - \sigma_{k,s}\frac{1}{N_s} \sum_{i\in\mathcal{D}_{N_s}} \mathbb{1}_{\left\{\widehat{p}_k(X_i,s)=\lambda+\frac{\gamma_{k,s}}{\widehat{\pi}_s}\right\}}.$$

From the above inequality, we then deduce thanks to Lemma B.7

$$\left|\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right) - \widehat{\mathbb{P}}_{X|S=s'}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s'}}{\widehat{\pi}_{s'}}\right)\right| \leq \frac{C}{N_{\min}}.$$

Combining the above inequality together with Equation 14, Lemma B.2, and Lemma B.4, we easily obtain that for each k, s, s'

$$\mathbb{E}\left[\left|\mathbb{P}_{X|S=s}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s}}{\widehat{\pi}_s}\right) - \mathbb{P}_{X|S=s'}\left(\widehat{p}_k(X,S) > \widehat{\lambda} + \frac{\widehat{\gamma}_{k,s'}}{\widehat{\pi}_{s'}}\right)\right|\right] \leq \sqrt{\frac{C_{k,\mathcal{S}}}{N}},$$

that yields the desired result. □

Proof of Theorem 3.2. For each $(\lambda, \gamma) \in \Delta$, we consider the predictor $\Gamma^*_{\lambda,\gamma}$ defined as

$$\Gamma^*_{\lambda,\gamma}(x,s) = \left\{k \in [K],\ \ p_k(x,s) \geq \lambda + \frac{\gamma_{k,s}}{\pi_s}\right\}.$$

Note that using similar arguments as in the proof of Proposition 2.3, we can show that the predictor Γ^* is optimal with respect to $\mathcal{R}_{\lambda,\gamma}$ defined for a predictor Γ by

$$\mathcal{R}_{\lambda,\gamma}(\Gamma) = R(\Gamma) + \lambda\left(\mathbb{E}_X\left[|\Gamma(X,S)|\right] - \beta\right) + \sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\gamma_{k,s}\mathbb{P}_{X|S=s}\left(k \in \Gamma(X,s)\right) \quad .$$

Hence

$$\Gamma^*_{\lambda,\gamma} \in \arg\min_{\Gamma} \mathcal{R}_{\lambda,\gamma}(\Gamma).$$

Besides similarly to Proposition 2.3, we have that

$$\mathcal{R}_{\lambda,\gamma}\left(\Gamma^*_{\lambda,\gamma}\right) - \mathcal{R}_{\lambda,\gamma}(\Gamma) = \sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\mathbb{E}_{X|S=s}\left[\mathbb{1}_{\left\{k\in\Gamma(X,s)\Delta\Gamma^*_{\lambda,\gamma}(X,s)\right\}}\left|\pi_s\left(p_k(X,s)-\lambda\right)-\gamma_{k,s}\right|\right] \quad , \tag{15}$$

Finally, we recall that the optimal parameters satisfy

$$(\lambda^*, \gamma^*) \in \arg\min_{(\lambda,\gamma)\in\Delta} \mathcal{R}_{\lambda,\gamma}(\Gamma^*_{\lambda,\gamma}).$$

Now, we start with the following decomposition

$$\begin{aligned}\mathcal{R}_{\lambda^*,\gamma^*}(\widehat{\Gamma}) - \mathcal{R}_{\lambda^*,\alpha^*}(\Gamma^*_\beta) = &\left(\mathcal{R}_{\lambda^*,\gamma^*}(\widehat{\Gamma}) - \mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\widehat{\Gamma})\right)\\ &+ \left(\mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\widehat{\Gamma}) - \mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\Gamma^*_{\widehat{\lambda},\widehat{\gamma}})\right) + \left(\mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\Gamma^*_{\widehat{\lambda},\widehat{\gamma}}) - \mathcal{R}_{\lambda^*,\gamma^*}(\Gamma^*_{\lambda^*,\gamma^*})\right)\end{aligned} \tag{16}$$

We now act on each of the three terms. For the first term in the *r.h.s.* of the above equation we observe that since for each $k \in [K]$, $\sum_s \widehat{\gamma}_{k,s} = \sum_s \gamma^*_{k,s} = 0$, we have that

$$\begin{aligned}\mathcal{R}_{\lambda^*,\gamma^*}(\widehat{\Gamma}) - \mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\widehat{\Gamma}) = &R(\widehat{\Gamma}_{\widehat{\lambda},\widehat{\gamma}}) - R(\widehat{\Gamma}_{\widehat{\lambda},\widehat{\gamma}})\\ &+ (\lambda^* - \widehat{\lambda})\left(\mathbb{E}_{X,S}\left[\left|\widehat{\Gamma}_{\widehat{\lambda},\widehat{\gamma}}(X,S)\right|\right] - \beta\right)\\ &+ \sum_{s\in\mathcal{S}}\sum_{k=1}^{K}(\gamma^*_{k,s} - \widehat{\gamma}_{k,s})\left(\mathbb{P}_{X|S=s}\left(k \in \widehat{\Gamma}_{\widehat{\lambda},\widehat{\gamma}}(X,s)\right) - \mathbb{P}_{X|S=1}\left(k \in \widehat{\Gamma}_{\widehat{\lambda},\widehat{\gamma}}(X,S)\right)\right).\end{aligned}$$

Therefore, from Theorem 3.1, since parameters $\lambda^*, \widehat{\lambda}, \gamma^*$, and $\widehat{\gamma}$ are bounded, we deduce that

$$\mathcal{R}_{\lambda^*,\gamma^*}(\widehat{\Gamma}) - \mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\widehat{\Gamma}) \leq C_{K,\mathcal{S}}\sqrt{\frac{1}{N}}. \tag{17}$$

For the last term, we use the characterization of (λ^*, γ^*) and then observe that

$$\mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\Gamma^*_{\widehat{\lambda},\widehat{\gamma}}) - \mathcal{R}_{\lambda^*,\gamma^*}(\Gamma^*_{\lambda^*,\gamma^*}) \leq 0 \tag{18}$$

Finally, we consider the second term in the *r.h.s.* of Equation (16). Using Equation (15), we deduce

$$\begin{aligned}\mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\widehat{\Gamma}_{\widehat{\lambda},\widehat{\gamma}}) - \mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\Gamma^*_{\widehat{\lambda},\widehat{\gamma}}) \\ = \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\mathbb{1}_{\left\{k\in\widehat{\Gamma}_{\widehat{\lambda},\widehat{\gamma}}(X,s)\Delta\Gamma^*_{\widehat{\lambda},\widehat{\gamma}}(X,s)\right\}}\left|\pi_s\left(p_k(X,s)-\widehat{\lambda}\right)-\widehat{\gamma}_{k,s}\right|\right]\end{aligned} \tag{19}$$

We observe that $k \in \widehat{\Gamma}_{\widehat{\lambda},\widehat{\gamma}}(X,s)\Delta\Gamma^*_{\widehat{\lambda},\widehat{\gamma}}(X,s)$ implies

$$\begin{aligned}\left|\pi_s\left(p_k(X,s)-\widehat{\lambda}\right)-\widehat{\gamma}_{k,s}\right| &\leq \left|\pi_s\left(p_k(X,s)-\widehat{\lambda}\right)-\widehat{\gamma}_{k,s}-\widehat{\pi}_s\left(\widehat{p}_k(X,S)-\widehat{\lambda}\right)+\widehat{\gamma}_{k,s}\right| \\ &\leq \left|\pi_s(p_k(X,s)-\widehat{p}_k(X,s))+(\widehat{p}_k(X,s)-\widehat{\lambda})(\pi_s-\widehat{\pi}_s)\right|.\end{aligned}$$

Therefore, from Equation 19, since $\widehat{p}_k$, and $\widehat{\lambda}$ are bounded, we deduce that

$$\begin{aligned}\mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\widehat{\Gamma}_{\widehat{\lambda},\widehat{\gamma}}) - \mathcal{R}_{\widehat{\lambda},\widehat{\gamma}}(\Gamma^*_{\widehat{\lambda},\widehat{\gamma}}) &\leq \sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\pi_s\mathbb{E}_{X|S=s}\left[|\widehat{p}_k(X,s)-p_k(X,s)|\right] + C_K\sum_{s\in\mathcal{S}}\mathbb{E}\left[|\widehat{\pi}_s-\pi_s|\right] \\ &\leq C_K\left(\max_{s\in\mathcal{S}}||\widehat{p}-p||_{\infty,\mathbb{P}_{X|S=s}}+\sqrt{\frac{1}{N}}\right)\end{aligned} \tag{20}$$

Combining the results from Equations (17), (18) and (20), we obtain

$$\mathcal{R}_{\lambda^*,\gamma^*}(\widehat{\Gamma}_{\widehat{\lambda},\widehat{\gamma}}) - \mathcal{R}_{\lambda^*,\gamma^*}(\Gamma^*_{\lambda^*,\gamma^*}) \leq C_{K,\mathcal{S}}\frac{1}{\sqrt{N}} + C_K\max_{s\in\mathcal{S}}||\widehat{p}-p||_{\infty,\mathbb{P}_{X|S=s}},$$

which yields the desired result. □

B.4 Proof of Section 4

First of all, applying Lemma B.1, we have, almost surely, that for each $u \in (0,1)$, and $\beta \in (0,K)$

$$0 \leq u - \widehat{\overline{F}}_{k,s}(\widehat{\overline{F}}^{-1}_{k,s}(u)) \leq \frac{1}{N_s}, \text{ and } 0 \leq \beta - \widehat{G}(\widehat{G}^{-1}(\beta)) \leq \frac{K}{N}. \tag{21}$$

Proof of Theorem 4.2. We start the proof the unfairness bound and then establish the bound on the expected size.

Unfairness bound. For each $k \in [K], s \in \mathcal{S}$, we introduce $\widehat{\beta}_k = \widehat{\overline{F}}_k(\widehat{G}^{-1}(\beta))$, $\widehat{\delta}_{k,s} = \widehat{\overline{F}}^{-1}_{k,s}\left(\widehat{\beta}_k\right)$, and $\widehat{h}_{k,s} = \widehat{p}_k(X,s) - \widehat{\delta}_{k,s}$. We have

$$\begin{aligned}\mathcal{U}(\widehat{\Gamma}) = \max_{k,s,s'}\Big(&\left|\mathbb{P}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)-\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)+\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)\right. \\ &\left.-\widehat{\mathbb{P}}_{X|S=s'}\left(\widehat{h}_{k,s'}\geq 0\right)+\widehat{\mathbb{P}}_{X|S=s'}\left(\widehat{h}_{k,s'}\geq 0\right)-\mathbb{P}_{X|S=s'}\left(\widehat{h}_{k,s'}\geq 0\right)\right|\Big)\end{aligned}$$

Then, we deduce that

$$\begin{aligned}\mathcal{U}(\widehat{\Gamma}) \leq 2\max_{k,s}\left(\left|\mathbb{P}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)-\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)\right|\right) \\ + \max_{k,s,s'}\left(\left|\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)-\widehat{\mathbb{P}}_{X|S=s'}\left(\widehat{h}_{k,s'}\geq 0\right)\right|\right)\end{aligned} \tag{22}$$

Noting that for

$$\max_{k,s}\left(\left|\mathbb{P}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)-\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)\right|\right) \leq \sum_{k\in[K]}\sup_t\left|\mathbb{P}_{X|S=s}\left(\widehat{p}_k(X,s)>t\right)-\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{p}_k(X,s)>t\right)\right|.$$

Similarly to the proof of Theorem 3.1, using Lemma B.4 conditionally on D_n and N_s and then by integrating over D_n and N_s, thanks to Lemma B.2, we have that there exists $C>0$ a constant such that:

$$\mathbb{E}\left[\max_{k,s}\left(\left|\mathbb{P}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)-\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)\right|\right)\right]\leq C\frac{K}{\sqrt{N}}.$$

Furthermore, thanks to Lemma B.1, since $\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)=\widehat{\overline{F}}_{k,s}(\widehat{\delta}_{k,s})$. We can write

$$\begin{aligned}\max_{k,s,s'}\left(\left|\widehat{\mathbb{P}}_{X|S=s}\left(\widehat{h}_{k,s}\geq 0\right)-\widehat{\mathbb{P}}_{X|S=s'}\left(\widehat{h}_{k,s'}\geq 0\right)\right|\right) &= \max_{k,s,s'}\left(\left|\widehat{\overline{F}}_{k,s}(\widehat{\delta}_{k,s})-\widehat{\overline{F}}_{k,s'}(\widehat{\delta}_{k,s'})\right|\right)\\ &\leq \max_{k,s,s'}\left(\left|\widehat{\beta}_k+\frac{1}{N_s}-\widehat{\beta}_k+\frac{1}{N_{s'}}\right|\right)\\ &\leq \max_{k,s,s'}\left(\left|\frac{1}{N_s}+\frac{1}{N_{s'}}\right|\right)\\ &\leq \frac{2}{\min_s N_s}.\end{aligned}$$

In view of Equation 22, applying again Lemma B.2, we can now combine all the terms

$$\mathbb{E}\left[\mathcal{U}(\Gamma)\right]\leq\frac{C_{K,\mathcal{S}}}{\sqrt{N}}.$$

Size constraint violation. We can write:

$$\begin{aligned}\left|\mathbb{E}_{X,S}\left[\left|\widehat{\Gamma}(X,S)\right|\right]-\beta\right| &= \left|\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\pi_s\mathbb{P}_{X|S=s}\left(k\in\widehat{\Gamma}(X,s)\right)-\beta\right|\\ &= \left|\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\pi_s\left(\overline{F}_{k,s}\left(\widehat{\delta}_{k,s}\right)\right)-\beta\right|\\ &= \left|\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\pi_s\left(\overline{F}_{k,s}\left(\widehat{\delta}_{k,s}\right)-\widehat{\overline{F}}_{k,s}\left(\widehat{\delta}_{k,s}\right)+\widehat{\overline{F}}_{k,s}\left(\widehat{\delta}_{k,s}\right)\right)-\beta\right|\end{aligned}$$

Then we deduce that

$$\left|\mathbb{E}_{X,S}\left[\left|\widehat{\Gamma}(X,S)\right|\right]-\beta\right|\leq\left|\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\pi_s\left(\overline{F}_{k,s}\left(\widehat{\delta}_{k,s}\right)-\widehat{\overline{F}}_{k,s}\left(\widehat{\delta}_{k,s}\right)\right)\right|+\left|\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\pi_s\widehat{\overline{F}}_{k,s}\left(\widehat{\delta}_{k,s}\right)-\beta\right|.$$

To control the first term in the *r.h.s.* of the above equation we use Lemma B.4. For the second term, we observe that

$$\begin{aligned}&\left|\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\pi_s\widehat{\overline{F}}_{k,s}\left(\widehat{\delta}_{k,s}\right)-\beta\right|\\ &\qquad\leq\left|\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\pi_s\widehat{\overline{F}}_{k,s}\left(\widehat{\delta}_{k,s}\right)-\sum_{k\in[K]}\sum_{s\in\mathcal{S}}\pi_s\widehat{\overline{F}}_k(\widehat{G}^{-1}(\beta))\right|+\left|\sum_{k\in[K]}\sum_{s\in\mathcal{S}}\pi_s\widehat{\overline{F}}_k(\widehat{G}^{-1}(\beta))-\beta\right|.\end{aligned}$$

From Equation 21, we then deduce that almost surely

$$\left| \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \pi_s \widehat{\overline{F}}_{k,s} \left(\widehat{\delta}_{k,s} \right) - \beta \right| \leq \frac{1}{N_s} + \frac{K}{N}.$$

Therefore, applying again Lemma B.2 we deduce the desired result. □