Fair learning with Wasserstein barycenters for non-decomposable performance measures

Solenne Gaucher
CNRS, LMO, Inria
Université Paris-Saclay

Nicolas Schreuder
MaLGa, DIBRIS
Università di Genova

Evgenii Chzhen
CNRS, LMO
Université Paris-Saclay

Abstract

This work provides several fundamental characterizations of the optimal classification function under the demographic parity constraint. In the awareness framework, akin to the classical unconstrained classification case, we show that maximizing accuracy under this fairness constraint is equivalent to solving a fair regression problem followed by thresholding at level $1/2$. We extend this result to linear-fractional classification measures (e.g., F-score, AM measure, balanced accuracy, etc.), highlighting the fundamental role played by regression in this framework. Our results leverage recently developed connection between the demographic parity constraint and the multi-marginal optimal transport formulation. Informally, our result shows that the transition between the unconstrained problem and the fair one is achieved by replacing the conditional expectation of the label by the solution of the fair regression problem. Finally, leveraging our analysis, we demonstrate an equivalence between the awareness and the unawareness setups for two sensitive groups.

1 INTRODUCTION

Our experience of life is increasingly and insidiously being influenced by algorithmic predictions. It is now well accepted that such predictions might replicate or even amplify societal biases and discrimination because of machine learning algorithms' training process (Barocas et al., 2019). A key difficulty in overcoming the effect of those biases is the lack of a precise understanding of how statistical algorithms make predictions: these algorithms are often designed to minimize a user-specified data-dependent loss and yield a highly complex prediction rule, leaving practitioners—and theoreticians—unable to understand and explain the issued predictions. Our goal is to provide a sound and simple mathematical characterization of the prediction process in the presence of fairness constraints.

In this paper we study the demographic parity fairness constraint (Calders et al., 2009) in the *awareness* framework—allowing the prediction rules to explicitly take the sensitive attribute as an input. Even though this constraint is relatively well understood from an algorithmic perspective in both classification (Agarwal et al., 2018; Menon and Williamson, 2018; Zeng et al., 2022; Schreuder and Chzhen, 2021; Yang et al., 2020; Jiang et al., 2020; Chiappa et al., 2020; Feldman et al., 2015; Gordaliza et al., 2019) and regression (Chzhen et al., 2020b,a; Le Gouic et al., 2020; Jiang et al., 2020; Agarwal et al., 2019; Chiappa and Pacchiano, 2021), the connection between the two setups remains opaque. The main goal of the current paper is to unveil it.

In contrast, in the traditional unconstrained learning setup, the relation between classification and its regression counterpart is well understood and can be found in all standard books on the subject (see, e.g., Hastie et al., 2009; Devroye et al., 2013; James et al., 2013; Mohri et al., 2018). For instance, a standard result illustrating this connection states that if η minimizes the squared risk, the classifier $g^*(\cdot) = \mathbf{1}\left(\eta(\cdot) \geq 1/2\right)$ minimizes the misclassification error. Such results form the first building block of many theoretical and practical studies (see, e.g., Audibert and Tsybakov, 2007; Yang, 1999; Massart and Nédélec, 2006; Biau et al., 2008). More recently, the connection between regression and classification was pushed even further. For instance, replacing the misclassification error by the F-score (Van Rijsbergen, 1974; Chinchor, 1992), Zhao et al. (2013) showed that an F-score maximizer can be obtained by properly thresholding the minimizer of the squared risk η. Moreover, a recent thread of results establish this fundamental relation for a large variety of performance measures including AM measure, the Jaccard similarity coefficient, and G-mean, to name a few (Menon et al., 2013; Koyejo et al., 2014, 2015; Yan et al., 2018). Again, akin to the standard minimization of misclassification error problem, all these developments led to many theoretical and practical advances (see,

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

e.g., Jasinska et al., 2016; Chzhen, 2020; Narasimhan et al., 2015; Kotlowski and Dembczyński, 2016; Bascol et al., 2019; Boughorbel et al., 2017). Interestingly, some works that consider group fairness constraints do report F-score as a performance measure in their empirical studies without actually tailoring an algorithm to optimize it directly (see, e.g., Biswas and Rajan, 2020, 2021; Chen et al., 2022; Wang and Singh, 2021; Dablain et al., 2022; Wick et al., 2019). A possible cause of this is the absence of characterization of fair (F-score) optimal classifiers in the fairness literature. In this paper we fill this gap for the demographic parity constraint and a large class of performance measures.

Literature that treats group fairness notions is typically distinguished by two features: fairness definition and access to the sensitive attribute at prediction time. While this work focuses on demographic parity, we discuss both awareness and unawareness setups—allowing or not the access to the sensitive attribute at prediction time respectively. Unlike the case of awareness, in which a significant understanding has been achieved from a theoretical perspective, the case of unawareness remains opaque with contributions mainly focusing on algorithmic constructions (see e.g., Agarwal et al., 2018, 2019; Oneto et al., 2020; Michele et al., 2017; Narasimhan, 2018). A notable work of Lipton et al. (2018) puts forward several empirical evidences highlighting critical issues arising in the unawareness setup. Our work makes a step towards a more explicit and transparent description of the optimal classifier under the demographic parity constraint with unawareness by introducing a simple theoretical reduction scheme to the awareness setup for binary protected attribute. Consequently, our results support theoretically the empirical claims made by Lipton et al. (2018).

Contributions. The goal of this work is to establish a link between regression and classification problems under the demographic parity constraint. We make the following contributions to the study of algorithmic fairness: **1)** we show that, under mild assumptions, if f^* minimizes the squared risk under the demographic parity constraint, then $\mathbf{1}\left(f^* \geq 1/2\right)$ minimizes the probability of misclassification under the same constraint; **2)** we extend the above result to a large family of performance measures introduced in Koyejo et al. (2015) for unconstrained classification; **3)** in the case of a binary sensitive attribute, we provide a simple reduction scheme that transforms, in a optimal way, the *unawareness* setup into the *awareness* one.

The first two contributions show the fundamental role played by regression in the context of demographic parity constraint and are built using basic tools from univariate optimal transport theory. As an interesting consequence of our analysis, we show that the notion of strong demographic parity introduced by Jiang et al. (2020) is equivalent to the usual demographic parity when a performance measure is minimized. The latter indicates that the post-hoc or the downstream threshold will never harm the demographic parity constraint. The last contribution constitutes a step towards the theoretical treatment of the unawareness setup—a problem that still remains open. Importantly, even though our results are stated in the fair learning setting, they imply new results in the general learning setting. In particular, our results characterize the optimal unconstrained classifier for a large class of classification performance measures.

2 PROBLEM SETUP

Consider a triplet $(\boldsymbol{X}, S, Y) \in \mathcal{X} \times [K] \times \{0,1\}$, following some joint distribution $\mathbb{P}$, consisting of the nominally non-sensitive and sensitive features, and the label, respectively. Classifiers are functions of the form $g : \mathcal{X} \times [K] \mapsto \{0,1\}$ and score functions take the form $f : \mathcal{X} \times [K] \mapsto [0,1]$. The set of all classifiers is denoted by $\mathcal{G}$ and the set of all score functions is denoted by $\mathcal{F}$. We set $\eta(\boldsymbol{X}, S) \triangleq \mathbb{E}[Y \mid \boldsymbol{X}, S]$ and recall that η minimizes the squared risk without any constraint. For each $s \in [K]$, we define $p_s \triangleq \mathbb{P}(S = s)$. The central object of this work is the optimal fair score function

$$f^* \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \left\{\mathbb{E}(Y - f(\boldsymbol{X}, S))^2 \,:\, f(\boldsymbol{X}, S) \perp\!\!\!\perp S\right\} \enspace . \quad (1)$$

An explicit expression for f^* under standard assumptions was derived in (Chzhen et al., 2020b; Le Gouic et al., 2020) using the univariate optimal transport theory and the reduction of the problem in Eq. (1) to a multi-marginal optimal transport formulation. In particular, they showed that, under mild assumptions, there is a one-to-one correspondence between the problem in Eq. (1) and the problem

$$\min_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{K} p_s \mathsf{W}_2^2\left(\mathrm{Law}(\eta(\boldsymbol{X}, S) \mid S = s),\, \nu\right) \enspace ,$$

where W_2 is the Wasserstein-2 distance (Villani, 2009, Definition 6.1) and $\mathcal{P}_2(\mathbb{R})$ denotes the space of univariate probability measures with finite second moment. Denoting by $\nu^\star$ the solution of the above problem, it was shown that

$$f^*(\boldsymbol{x}, s) = T_{\mathrm{Law}(\eta(\boldsymbol{X}, S) \mid S = s) \to \nu^\star}\left(\eta(\boldsymbol{x}, s)\right) \enspace ,$$

where $T_{\mathrm{Law}(\eta(\boldsymbol{X}, S) \mid S = s) \to \nu^\star}$ is the optimal transport map from $\mathrm{Law}(\eta(\boldsymbol{X}, S) \mid S = s)$ to $\nu^\star$. Up until now, unlike in the regression setting, it was not clear if a direct link between optimal transport and the fair binary classification problem existed–or even made sense. Our work shows that such a connection exists and that it is fundamental.

Notation. Given a real-valued function $f : \mathcal{X} \times [K] \to \mathbb{R}$, we denote by $\mu_s(f)$ the univariate measure defined for all $A \subset \mathbb{R}$ as $\mu_s(f)(A) \triangleq \mathbb{P}(f(\boldsymbol{X}, S) \in A \mid S = s)$. For any univariate measure μ, we denote by F_μ its cumulative distribution. For any $x \in \mathbb{R}$ we set $(x)_+ \triangleq \max\{x,\, 0\}$.

3 THE MISCLASSIFICATION RISK: A WARM-UP

In this section, we begin by tackling the classical minimization of the misclassification risk problem and highlight the main novelties and advances with respect to previous works. To this end, we consider the following optimal (in terms of the misclassification risk) fair classifier

$$g^* \in \arg\min_{g \in \mathcal{G}} \left\{\mathbb{P}(Y \neq g(\boldsymbol{X}, S)) \,:\, g(\boldsymbol{X}, S) \perp\!\!\!\perp S\right\} \enspace . \tag{2}$$

We work under the following assumption.

Assumption 3.1. *For $s \in [K]$, let $\mathrm{Law}(\eta(\boldsymbol{X}, S) \mid S = s)$ be continuous and supported on an interval.*

A slightly modified version of the above was used in the context of fairness in (Chzhen et al., 2020b,a; Le Gouic et al., 2020; Jiang et al., 2020) and also also in the classical unconstrained classification with generalized performance measures (Yan et al., 2018). In Section A, we relax the above assumption and provide a proof that unifies the awareness case considered just below with the unawareness case presented in Section 5, Theorem 5.2.

The first warm-up result is reminiscent of those recently obtained in (Zeng et al., 2022; Schreuder and Chzhen, 2021). The proof based on the $\min\max$ duality and is very similar to the classical Neyman-Pearson lemma. While it does not allow to immediately reach our goals, it gives several fundamental insights that were already invoked in previous works on the demographic parity constraint (Lipton et al., 2018; Hardt et al., 2016).

Theorem 3.2. *Under Assumption 3.1, $g^* : \mathcal{X} \times [K] \to \{0, 1\}$ defined in (2) can be expressed as*

$$g^*(\boldsymbol{x}, s) = \mathbf{1}\left(2\eta(\boldsymbol{x}, s) - 1 \geq \lambda^*_s / p_s\right) \enspace ,$$

where $\boldsymbol{\lambda}^* = (\lambda^*_1, \ldots, \lambda^*_K)^\top \in \mathbb{R}^K$ is a solution of

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^K} \left\{\mathbb{E}\left[\left|2\eta(\boldsymbol{X}, S) - 1 - \frac{\lambda_S}{p_S}\right|\right] \,:\, \mathbb{E}\left[\frac{\lambda_S}{p_S}\right] = 0\right\} \enspace .$$

The main takeaway message from the above theorem is: under the stated assumption, the optimal fair classifier can be derived as a group-wise thresholding of the regression function η, with thresholds eventually depending on the sensitive groups. For a similar statement without the continuity assumption, we refer the reader to Zeng et al. (2022) who derived optimal randomized classifiers using the Neyman-Pearson lemma. Let us now provide a novel characterization of an optimal fair classifier.

Theorem 3.3 (Wasserstein based fair optimal classifier)**.** *Under Assumption 3.1, $g^* : \mathcal{X} \times [K] \to \{0, 1\}$ defined in (2) can be expressed as*

$$g^*(\boldsymbol{x}, s) = \mathbf{1}\left(f^*(\boldsymbol{x}, s) \geq 1/2\right) \text{ with } f^* \text{ defined in (1)} \enspace .$$

Discussion. The above result is instructive on its own—one can solve binary classification under the demographic parity constraint by solving the corresponding regression problem. We recall that (Chzhen et al., 2020b; Le Gouic et al., 2020) built a statistically consistent algorithm for the estimation of the latter. Furthermore, they showed that,

$$f^*(\boldsymbol{x}, s) = \underbrace{\left(\sum_{\sigma=1}^K p_\sigma F^{-1}_{\mu_\sigma(\eta)}\right) \circ F_{\mu_s(\eta)}}_{\text{transport to the barycenter}} \circ\, \eta(\boldsymbol{x}, s) \enspace .$$

Feldman et al. (2015) proposed to transport the group-wise distribution of $\eta(\boldsymbol{X}, S)$ towards their common barycenter as a disparity removal strategy. Yet, a theoretical justification was missing and this approach remained a heuristic until the work of Gordaliza et al. (2019) who provided an upper bound on the excess risk in terms of the Wasserstein barycenter objective. Later, Jiang et al. (2020) relied on the barycenter formulation involving the Earth Mover distance (Rachev and Rüschendorf, 1998) and showed that a transport-based prediction results in a minimal perturbation post-processing. However, the use of the Earth Mover distance might result in non-uniqueness issues. Our Theorem 3.3 gives a complete theoretical justification of the transport based fair classification algorithms. Theorem 4.3 in Section 4 further extends this connection to non-decomposable measures.

Besides, Jiang et al. (2020) introduced a notion of strong demographic parity, which amounts to taking classifiers $g : \mathcal{X} \times [K] \to \{0, 1\}$ for which there exists a score function $f : \mathcal{X} \times [K] \to [0, 1]$ such that $f(\boldsymbol{X}, S) \perp\!\!\!\perp S$ and $g(\boldsymbol{x}, s) = \mathbf{1}\left(f(\boldsymbol{x}, s) \geq 1/2\right)$. This notion was later used in (Chiappa et al., 2020; Chiappa and Pacchiano, 2021). Theorem 3.3 implies that the optimal classifier under the demographic parity constraint satisfies, *an a priori more restrictive* fairness notion—the strong demographic parity. Indeed, any classifier that satisfies strong demographic parity is demographic parity fair. Hence, we have deduced the equivalence between the two definitions at the optimum. The notion of strong demographic parity introduced by Jiang et al. (2020) can be seen in a downstream or post-hoc settings. That is, the learner first tries to fit a score function and only after a particular threshold is selected in a potentially non-stationary way. Strong demographic parity implies that *any* threshold selection made by the learner will yield a fair classifier. In that sense, our results show that building a score function via an optimal fair regression function is optimal for misclassification risk and, as we see in Section 4, for many other classification measures. In appendix we provide a simple proof of Theorem 3.3. The proof itself is rather instructive and gives rise to the following interpretation.

Remark 3.4. *The proof of Theorem 3.3 reveals that the optimal fair classifier can be written as $g^*(\boldsymbol{x}, s) = \mathbf{1}\left(F_{\mu_s(\eta)}\left(\eta(\boldsymbol{x}, s)\right) \geq \gamma^*\right)$, where γ^* is given by Eq. (19) of the proof. Recall that $q \mapsto \sum_{s=1}^K p_s F^{-1}_{\mu_s(\eta)}(q)$ is the quantile function of the Wasserstein-2 barycenter of measures $(\mu_s(\eta))_{s \in [K]}$, weighted by $(p_s)_{s \in [K]}$ (see, e.g., Agueh and Carlier, 2011, Section 6.1). Thus, denoting this barycenter*

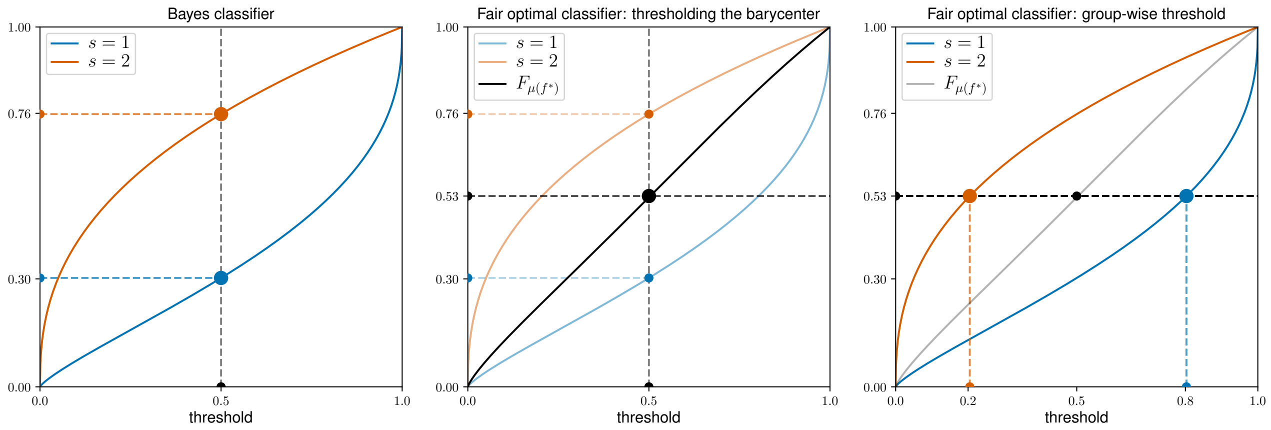

Figure 1: Illustration of Bayes and fair optimal classifiers. `Left`: group-wise cumulative distributions of $\eta(\boldsymbol{X}, S)$—the threshold is .5; `Middle`: Illustration of Theorem 3.3—black solid line corresponds to $F_{\mu(f^*)}$; `Right`: illustration of group-wise thresholds that correspond to the fair optimal classifier.

by $\bar{\mu}(\eta)$, g^* can be alternatively expressed as

$$g^*(\boldsymbol{x}, s) = \mathbf{1}\left(F_{\mu_s(\eta)}\left(\eta(\boldsymbol{x}, s)\right) \geq F_{\bar{\mu}(\eta)}(1/2)\right) \quad .$$

The last display shows that while the thresholds of η differ across groups (as per Theorem 3.2), this threshold sensitive-group independent if viewed from the perspective of group-wise ranking. Putting it simply, if $F_{\bar{\mu}(\eta)}(1/2) = p \in (0, 1)$, then the $(1-p) \times 100\%$ best *individuals from each group get classified positively. This property reflects the notion of rational ordering (see Lipton et al., 2018, Section 4) that follows from order preservation property of f^* (see Chzhen and Schreuder, 2022, Section 4). Figure 1 provides a graphical illustration of the above observations.*

We note that as in other works explaining a given fairness constraint, we do not argue for or against the policy itself.

Price of Fairness. From Theorem 3.3, we can derive an exact expression for the Price of Fairness (PoF) as well as an easy-to-estimate upper-bound. We recall that PoF is typically defined as the difference between the risk of fair optimal classifier and the Bayes optimal one, *i.e.*,

$$\mathrm{PoF} \triangleq \mathbb{P}(Y \neq g^*(\boldsymbol{X}, S)) - \min_{g \in \mathcal{G}} \mathbb{P}(Y \neq g(\boldsymbol{X}, S)) \quad ,$$

where $g^* : \mathcal{X} \times [K] \to \{0, 1\}$ is defined in (2).

Proposition 3.5. *Let Assumption 3.1 be satisfied. Then,*

$$\mathrm{PoF} = \mathbb{E}|\eta(\boldsymbol{X}, S) - 1/2| - \mathbb{E}|f^*(\boldsymbol{X}, S) - 1/2| \quad ,$$

with f^* defined in (1).

Corollary 3.6. *In the context of regression under the Demographic Parity constraint, Chzhen and Schreuder (2022) introduce a measure of unfairness for score functions f : $\mathcal{X} \times [K] \to \mathbb{R}$ as*

$$\mathcal{U}(f) = \min_{\nu \in \mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{K} p_s \mathsf{W}_2^2(\mu_s(f), \nu) \quad ,$$

where $\mathcal{P}_2(\mathbb{R})$ stands for univariate probability measures with finite second moment. In view of Proposition 3.5, we deduce that

$$\begin{aligned} \mathrm{PoF} &\leq \mathbb{E}|\eta(\boldsymbol{X}, S) - f^*(\boldsymbol{X}, S)| \\ &\leq \sqrt{\mathbb{E}|\eta(\boldsymbol{X}, S) - f^*(\boldsymbol{X}, S)|^2} = \sqrt{\mathcal{U}(\eta)} \quad , \end{aligned}$$

where the last equality is due to (Chzhen et al., 2020b, Theorem 2.3). The above inequality is rather intuitive and instructive—the price of fairness is controlled by the level of unfairness of the Bayes optimal prediction function.

This bound is a significant improvement to (Gordaliza et al., 2019, Theorem 3.3), who only derived this result for $K = 2$. Furthermore, their result required Lipschitz continuity of $\eta(\boldsymbol{x}, s)$ for $s \in \{1, 2\}$ and gave worse leading constants.

4 NON-DECOMPOSABLE PERFORMANCE MEASURES

In this part we extend the analysis of the previous section to a broader class of performance measures, which includes the F-score, the AM-mean, and the misclassification risk among others. We follow the framework put forward by Koyejo et al. (2014), who introduced the so-called *linear fractional performance measures*. Formally, given coefficients $(\mathsf{n}_0, \mathsf{n}_1, \mathsf{n}_2) \in \mathbb{R}^3$ and $(\mathsf{d}_0, \mathsf{d}_1, \mathsf{d}_2) \in \mathbb{R}^3$, the performance of a classifier $g : \mathcal{X} \times [K] \to \{0, 1\}$ is measured by its utility $\mathrm{U}_{(\mathsf{n}, \mathsf{d})}(g)$ defined as

$$\frac{\mathsf{n}_0 + \mathsf{n}_1 \mathbb{P}(g(\boldsymbol{X}, S){=}1, Y{=}1) + \mathsf{n}_2 \mathbb{P}(g(\boldsymbol{X}, S){=}1)}{\mathsf{d}_0 + \mathsf{d}_1 \mathbb{P}(g(\boldsymbol{X}, S){=}1, Y{=}1) + \mathsf{d}_2 \mathbb{P}(g(\boldsymbol{X}, S){=}1)} \quad . \tag{3}$$

	Expression	$(\mathsf{n}_0, \mathsf{n}_1, \mathsf{n}_2)$	$(\mathsf{d}_0, \mathsf{d}_1, \mathsf{d}_2)$
Accuracy	$\mathbb{P}(Y = g(\boldsymbol{X}, S))$	$(1 - p^{y=1},\ 2,\ -1)$	$(1,\ 0,\ 0)$
F$_b$-score	$\frac{(1+b^2)\mathbb{P}(Y=1,\, g(\boldsymbol{X},S)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g(\boldsymbol{X},S)=1)}$	$(0,\ 1 + b^2,\ 0)$	$(b^2 p^{y=1},\ 0,\ 1)$
Jaccard	$\frac{\mathbb{P}(Y=1,\, g(\boldsymbol{X},S)=1)}{\mathbb{P}(Y=1,\, g(\boldsymbol{X},S)=0)+\mathbb{P}(g(\boldsymbol{X},S)=1)}$	$(0,\ 1,\ 0)$	$(p^{y=1},\ -1, 1)$
AM-measure	$\frac{1}{2}\big\{\mathbb{P}(g(\boldsymbol{X}, S)=0 \mid Y=0)+\mathbb{P}(g(\boldsymbol{X}, S)=1 \mid Y=1)\big\}$	$\big(\frac{1}{2},\ \frac{1}{2p^{y=1}}+\frac{1}{2p^{y=0}},\ -\frac{1}{2p^{y=0}}\big)$	$(1,\ 0,\ 0)$

Table 1: Some examples of measure that can be represented by Eq. (3). For more examples see Choi et al. (2010). We set for this table $p^{y=1} \triangleq \mathbb{P}(Y = 1)$ and $p^{y=0} \triangleq \mathbb{P}(Y = 0)$.

We denote by $\mathrm{dom}(\mathrm{U}_{(\mathsf{n},\mathsf{d})}) \subset \mathcal{G}$ the set of all classifiers $g : \mathcal{X} \times [K] \to \{0, 1\}$ for which the denominator of $\mathrm{U}_{(\mathsf{n},\mathsf{d})}$ is non-zero. It is important to emphasize that both n and d are *allowed* to depend on the unknown distribution of the data $\mathbb{P}$ but *not* on the classifier g. For instance, the F_1-score (Van Rijsbergen, 1974) corresponds to the choice $(\mathsf{n}_0, \mathsf{n}_1, \mathsf{n}_2) = (0, 2, 0)$ and $(\mathsf{d}_0, \mathsf{d}_1, \mathsf{d}_2) = (\mathbb{P}(Y{=}1), 0, 1)$. We refer to (Choi et al., 2010) for additional examples of different choices of (n, d) corresponding to different classification performance measures. Recently, Yang et al. (2020) studied linear performance measures in the context of fairness, which essentially corresponds to the special case of the above linear fractional formulation with $(\mathsf{d}_0, \mathsf{d}_1, \mathsf{d}_2) = (1, 0, 0)$—which, for instance, does not encompass the F_1-score. In another direction, Celis et al. (2019) considered linear fractional formulation of *fairness* constraints while optimizing the misclassification risk. However, given the structure of the constraints, this problem can essentially be re-formulated as misclassification risk minimization under linear fairness constraints.

As it is common in the literature on generalized performance measures, we view $\mathrm{U}_{(\mathsf{n},\mathsf{d})}$ as a utility to be *maximized*, contrary to the minimization of the risk viewpoint from the previous section. Thus, our goal is to study

$$g^*_{(\mathsf{n},\mathsf{d})} \in \operatorname*{arg\,max}_{g \in \mathrm{dom}(\mathrm{U}_{(\mathsf{n},\mathsf{d})})} \left\{ \mathrm{U}_{(\mathsf{n},\mathsf{d})}(g) \,:\, g(\boldsymbol{X}, S) \perp\!\!\!\perp S \right\} \ . \quad (4)$$

A remarkable property of linear fractional measures is that the unconstrained maximizer can still be obtained by thresholding the regression function η. Yet, the threshold in this case might depend on the unknown distribution $\mathbb{P}$ and ought to be estimated. Let us provide couple of standard examples.

Example 4.1. *Consider the problem of maximizing the accuracy:* $\max_{g \in \mathcal{G}} \mathbb{P}(Y = g(\boldsymbol{X}, S))$*. Setting* $(\mathsf{n}_0, \mathsf{n}_1, \mathsf{n}_2) = (1 - \mathbb{P}(Y = 1), 2, -1)$ *and* $(\mathsf{d}_0, \mathsf{d}_1, \mathsf{d}_2) = (1, 0, 0)$*, we see that the above formulation falls within the considered framework.*

Example 4.2. *Consider the problem:*

$$\max_{g \in \mathcal{G}} \frac{2\mathbb{P}(g(\boldsymbol{X}, S) = 1, Y = 1)}{\mathbb{P}(Y = 1) + \mathbb{P}(g(\boldsymbol{X}, S) = 1)} \ ,$$

of maximizing the F_1*-score. Zhao et al. (2013) showed that the solution* g^* *of the above can be written as*

$$g^*(\boldsymbol{x}, s) = \mathbf{1}\left(\eta(\boldsymbol{x}, s) \geq \theta^*\right) \quad \textit{where}$$

θ^* *is the unique solution of* $\theta\mathbb{P}(Y = 1) = \mathbb{E}(\eta(\boldsymbol{X}, S) - \theta)_+$.

Koyejo et al. (2014) pushed further these results demonstrating that the "thresholding principle" remains true for the whole family of linear fractional measures. In what follows, we will show that their result is still valid if one replaces η by f^*—the solution of the fair regression problem. This validity is established in a strong sense, meaning that even the equation (as in Example 4.2) determining the threshold is preserved.

Theorem 4.3 (Fair optimal classifier for non-decomposable measures)**.** *Let Assumption 3.1 be satisfied. Assume that* $\mathsf{d} \in \mathbb{R}^3$ *is such that:* $\mathsf{d}_0 + \min\big\{\min\{\mathsf{d}_1, 0\} + \mathsf{d}_2, 0\big\} \geq 0$. *Assume that the coefficients* $(\mathsf{n}, \mathsf{d}) \in \mathbb{R}^3 \times \mathbb{R}^3$ *satisfy one of the following mutually exclusive conditions:*

$$\begin{cases} \mathsf{d}_2\mathsf{n}_1 {>} \mathsf{n}_2\mathsf{d}_1 \textit{ and } \mathsf{d}_0\mathsf{n}_1 {-} \mathsf{n}_0\mathsf{d}_1 {\geq} (\mathsf{n}_0\mathsf{d}_2 {-} \mathsf{d}_0\mathsf{n}_2)_+ \\ \dfrac{\mathsf{n}_0\mathsf{d}_2 - \mathsf{d}_0\mathsf{n}_2}{\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1} \leq \mathbb{P}(Y = 1) \end{cases} , \quad (\mathtt{C1})$$

or

$$\begin{cases} \mathsf{d}_2\mathsf{n}_1 = \mathsf{n}_2\mathsf{d}_1 \textit{ and } \mathsf{n}_1\mathsf{d}_0 > \mathsf{d}_1\mathsf{n}_0 \\ \dfrac{\mathsf{d}_0\mathsf{n}_2 - \mathsf{n}_0\mathsf{d}_2}{\mathsf{n}_0\mathsf{d}_1 - \mathsf{d}_0\mathsf{n}_1} \in [0, 1] \end{cases} . \quad (\mathtt{C2})$$

Then, $g^*_{(\mathsf{n},\mathsf{d})}$ *defined in Eq. (4) can be expressed for all* $(\boldsymbol{x}, s) \in \mathcal{X} \times [K]$ *as*

$$g^*_{(\mathsf{n},\mathsf{d})}(\boldsymbol{x}, s) = \mathbf{1}\left(f^*(\boldsymbol{x}, s) \geq \theta^*_{(\mathsf{n},\mathsf{d})}\right) \ , \quad (5)$$

where $\theta^*_{(\mathsf{n},\mathsf{d})}$ *is either the unique solution of*

$$\begin{aligned} \mathbb{E}\left[(f^*(\boldsymbol{X}, S) - \theta)_+\right] = &\theta \cdot \left\{\frac{\mathsf{n}_0\mathsf{d}_1 - \mathsf{d}_0\mathsf{n}_1}{\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1}\right\} \\ &+ \left\{\frac{\mathsf{n}_0\mathsf{d}_2 - \mathsf{d}_0\mathsf{n}_2}{\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1}\right\} \ , \end{aligned} \quad (6)$$

if $\mathsf{n}_2\mathsf{d}_1 \neq \mathsf{d}_2\mathsf{n}_1$ *or* $\theta^*_{(\mathsf{n},\mathsf{d})} = \frac{\mathsf{d}_0\mathsf{n}_2 - \mathsf{n}_0\mathsf{d}_2}{\mathsf{n}_0\mathsf{d}_1 - \mathsf{d}_0\mathsf{n}_1}$ *otherwise.*

A few comments are in order. First of all, Theorem 4.3 states that the pre-cited "thresholding principle" still holds for optimizing linear-fractional performance measures under the demographic parity constraint: optimal fair classifiers can be obtained by thresholding the optimal fair regression function f^* at the right threshold level $\theta^*_{(\mathsf{n},\mathsf{d})}$. Moreover, in the case $\mathsf{n}_2\mathsf{d}_1 = \mathsf{d}_2\mathsf{n}_1$ an explicit expression is provided, while if $\mathsf{n}_2\mathsf{d}_1 \neq \mathsf{d}_2\mathsf{n}_1$ one needs to solve a fixed-point equation to find the optimal threshold. Given that the function defining the fixed-point equation is univariate, monotone and continuous, the bisection method (or any other univariate root-finding method) can be used to obtain an approximation of the optimal threshold up to arbitrary precision. Finally, since the conditions on the coefficients might seem opaque at first sight, let us argue why they are harmless and meaningful. Intuitively, these conditions specify only two requirements: **1)** The maximization of $\mathrm{U}_{(\mathsf{n},\mathsf{d})}(g)$ makes sense—the more the classifier aligns with Y, the better. In particular, they exclude $\mathbb{P}(Y \neq g(\boldsymbol{X}, S))$, whose maximization does not make sense. **2)** The denominator of $\mathrm{U}_{(\mathsf{n},\mathsf{d})}$ is non-negative. One can verify that all the measures presented in Table 1 do indeed satisfy these conditions as well as many other linear fractional performance measures from Choi et al. (2010). We would also like to point out that while the conditions of Theorem 4.3 are cumbersome, they are easy to check in practice, unlike those given in (Koyejo et al., 2014), who relied on $\operatorname{sign}(\mathsf{n}_1 - \mathrm{U}_{(\mathsf{n},\mathsf{d})}(g^*_{(\mathsf{n},\mathsf{d})})\mathsf{d}_1)$. Indeed, to check the latter, one needs to know or estimate the optimal value of $\mathrm{U}_{(\mathsf{n},\mathsf{d})}$ beforehand, which is not always feasible in practice. In contrast, conditions (C1) and (C2) only involve the known coefficients (n, d). Finally, let us remark that $\mathrm{U}_{(\mathsf{n},\mathsf{d})} = \mathrm{U}_{(-\mathsf{n},-\mathsf{d})}$ and both conditions (C1) and (C2) are invariant under the $(\mathsf{n}, \mathsf{d}) \mapsto (-\mathsf{n}, -\mathsf{d})$ transformation. Yet, to fix only one of them, we additionally require $\mathsf{d}_0 + \min\big\{\min\{\mathsf{d}_1, 0\} + \mathsf{d}_2, 0\big\} \geq 0$, which forces the user to fix the signs of d properly. Let us emphasize that, if $\mathsf{d}_0 + \min\big\{\min\{\mathsf{d}_1, 0\} + \mathsf{d}_2, 0\big\} > 0$, then $\mathrm{dom}(\mathsf{U}_{(\mathsf{n},\mathsf{d})}) = \mathcal{G}$—the denominator does not zero-out—which is a consequence of Lemma C.1.

Proof of Theorem 4.3 follows from the following two results. The first lemma is similar to the main result of (Koyejo et al., 2014), while the second one gives an explicit expression for the excess-score of *any* fair classifier. The actual proof technique shares some similarities with the analysis of F_1-score in (Chzhen, 2020) who provided an alternative proof to the result of Zhao et al. (2013) recalled in Example 4.2.

Lemma 4.4. *Let $g^*_{(\mathsf{n},\mathsf{d})}$ be defined in Theorem 4.3, assume that $\theta^*_{(\mathsf{n},\mathsf{d})}$ in (6) exists. Then, under Assumption 3.1,*

$$\mathrm{U}_{(\mathsf{n},\mathsf{d})}\left(g^*_{(\mathsf{n},\mathsf{d})}\right) = \frac{\mathsf{n}_2 + \theta^*_{(\mathsf{n},\mathsf{d})}\mathsf{n}_1}{\mathsf{d}_2 + \theta^*_{(\mathsf{n},\mathsf{d})}\mathsf{d}_1} \qquad \textit{if } \mathsf{n}_2\mathsf{d}_1 \neq \mathsf{d}_2\mathsf{n}_1 \textit{ or}$$

$$\mathrm{U}_{(\mathsf{n},\mathsf{d})}\left(g^*_{(\mathsf{n},\mathsf{d})}\right) = \frac{\mathsf{n}_0 + \mathsf{n}_1\mathbb{E}\big(f^*(\boldsymbol{X}, S) - \theta^*_{(\mathsf{n},\mathsf{d})}\big)_+}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}\big(f^*(\boldsymbol{X}, S) - \theta^*_{(\mathsf{n},\mathsf{d})}\big)_+} \enspace,$$

otherwise.

The next result provides an explicit expression for the excess score of any fair classifier g.

Lemma 4.5. *Let Assumption 3.1 be satisfied. Let $g^*_{(\mathsf{n},\mathsf{d})}$ be defined as in Theorem 4.3 and assume that $\theta^* \triangleq \theta^*_{(\mathsf{n},\mathsf{d})}$ defined in Eq. (6) exists. Let $\bar{\mu}(\eta)$ be the Wasserstein barycenter of measures $\mu_1(\eta), \ldots, \mu_K(\eta)$ weighted by $p_1, \ldots, p_K$, respectively. Define β^* as $\beta^* = F_{\bar{\mu}(\eta)}(\theta^*)$. Let $\mathcal{E}_{(\mathsf{n},\mathsf{d})}(g) \triangleq \mathrm{U}_{(\mathsf{n},\mathsf{d})}\big(g^*_{(\mathsf{n},\mathsf{d})}\big) - \mathrm{U}_{(\mathsf{n},\mathsf{d})}\big(g\big)$. Then, for any classifier $g \in \mathrm{dom}(\mathrm{U}_{(\mathsf{n},\mathsf{d})})$ such that $g(\boldsymbol{X}, S) \perp\!\!\!\perp S$, excess score $\mathcal{E}_{(\mathsf{n},\mathsf{d})}(g)$ equals to*

$$\mathsf{C}_{\mathbb{P},(\mathsf{n},\mathsf{d})} \cdot \frac{\mathbb{E}|\eta(\boldsymbol{X}, S) - F^{-1}_{\mu_S(\eta)}(\beta^*)|\mathbf{1}\left(g^*(\boldsymbol{X}, S) \neq g(\boldsymbol{X}, S)\right)}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{P}(Y{=}1,\ g(\boldsymbol{X}, S){=}1) + \mathsf{d}_2\mathbb{P}(g(\boldsymbol{X}, S){=}1)} \enspace,$$

where

$$\mathsf{C}_{\mathbb{P},(\mathsf{n},\mathsf{d})} = \begin{cases} \dfrac{\mathsf{d}_2\mathsf{n}_1 - \mathsf{n}_2\mathsf{d}_1}{\mathsf{d}_2 + \theta^*\mathsf{d}_1} & \mathsf{n}_2\mathsf{d}_1 \neq \mathsf{d}_2\mathsf{n}_1 \\ \dfrac{\mathsf{n}_1\mathsf{d}_0 - \mathsf{d}_1\mathsf{n}_0}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}(f^*(\boldsymbol{X}, S) - \theta^*)_+} & \mathsf{n}_2\mathsf{d}_1 = \mathsf{d}_2\mathsf{n}_1 \end{cases} .$$

Furthermore, under the conditions on (n, d) specified in Theorem 4.3; we have $\mathcal{E}_{(\mathsf{n},\mathsf{d})}(g) \geq 0$ for all classifiers $g : \mathcal{X} \times [K] \to \{0, 1\}$.

Remark 4.6. *Lemma 4.5, together with Lemma C.4, stated in appendix, implies that $\mathsf{C}_{\mathbb{P},(\mathsf{n},\mathsf{d})} = \Big(\mathsf{n}_1 - \mathsf{d}_1\mathrm{U}_{(\mathsf{n},\mathsf{d})}(g^*_{(\mathsf{n},\mathsf{d})})\Big)$. Hence, the inequality $\mathcal{E}_{(\mathsf{n},\mathsf{d})}(g) \geq 0$ for all g is implied from*

$$\begin{cases} \mathsf{d}_0 + \mathsf{d}_1\mathbb{P}(Y{=}1,\ g(\boldsymbol{X}, S){=}1) + \mathsf{d}_2\mathbb{P}(g(\boldsymbol{X}, S){=}1) > 0 \\ \mathsf{n}_1 - \mathsf{d}_1\mathrm{U}_{(\mathsf{n},\mathsf{d})}(g^*_{(\mathsf{n},\mathsf{d})}) \geq 0 \end{cases} ,$$

for all $g \in \mathrm{dom}(\mathrm{U}_{(\mathsf{n},\mathsf{d})})$. The first of the above conditions is ensured if $\mathsf{d}_0 + \min\big\{\min\{\mathsf{d}_1, 0\} + \mathsf{d}_2, 0\big\} \geq 0$ (Lemma C.1) assumed in Theorem 4.3 and the second one is ensured by (C1) or (C2), as proved in Lemma C.2.

Let us remark that the content of this section can be seen as a strict improvement over Koyejo et al. (2014) who only derived Lemma 4.4 in the absence of the fairness constraint. Indeed, assuming that $S \perp\!\!\!\perp \boldsymbol{X}$, ensures that *any* classifier g is demographic parity fair and that $f^* \equiv \eta$. In the absence of the demographic parity constraint, Assumption 3.1 is unnecessary and *exactly* the same strategy gives the characterization of the optimal unconstrained classifier.

Examples: accuracy and F_1-score. In this part, we give specific examples of the parameters $(\mathsf{n}_0, \mathsf{n}_1, \mathsf{n}_2)$ and $(\mathsf{d}_0, \mathsf{d}_1, \mathsf{d}_2)$ and instantiate Theorem 4.3 and Lemma 4.5. The first examples concerns the accuracy as a performance metric. It highlights the generality of the derived results.

Example 4.7 (Accuracy under fairness constraint)**.** *Recalling the coefficients specified in Example 4.1, we see that in this case $\mathsf{n}_2\mathsf{d}_1 = \mathsf{d}_2\mathsf{n}_1$ and that condition (C2) is satisfied. Hence under Assumption 3.1, Theorem 4.3 states that*

$$g^*_{(\mathsf{n},\mathsf{d})}(\boldsymbol{x}, s) = \mathbf{1}\left(f^*(\boldsymbol{x}, s) \geq \theta^*_{(\mathsf{n},\mathsf{d})}\right) \enspace,$$

with $\theta^*_{(\mathsf{n},\mathsf{d})} = \frac{\mathsf{d}_0\mathsf{n}_2 - \mathsf{n}_0\mathsf{d}_2}{\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1} = \frac{1\cdot(-1)-(1-\mathbb{P}(Y=1))\cdot 0}{(1-\mathbb{P}(Y=1))\cdot 0 - 1\cdot 2} = \frac{1}{2}$ maximizes $\mathbb{P}(Y \neq g(\boldsymbol{X}, S))$ under the demographic parity constraint. Thus, it coincides with the result of Theorem 3.3. Furthermore, Lemma 4.5 states that for any classifier $g : \mathcal{X} \times [K] \to \{0, 1\}$ such that $g(\boldsymbol{X}, S) \perp\!\!\!\perp S$, it holds that $\mathbb{P}(Y{=}g^*_{(\mathsf{n},\mathsf{d})}(\boldsymbol{X}, S)) - \mathbb{P}(Y{=}g(\boldsymbol{X}, S))$ equals to

$$2\mathbb{E}|\eta(\boldsymbol{X}, S) - F^{-1}_{\mu_S(\eta)} \circ F_{\bar{\mu}(\eta)}(.5)|\mathbf{1}\left(g^*(\boldsymbol{X}, S) {\neq} g(\boldsymbol{X}, S)\right) \enspace .$$

We invite the reader to compare the above expression with its classical version (Devroye et al., 2013, Theorem 2.2).

The second example concerns the F_1-score that has been used in several empirical works on fairness as a performance measure (Wang and Singh, 2021; Dablain et al., 2022; Wick et al., 2019).

Example 4.8 (F_1-score under fairness constraint)**.** *Recall that the F_1-score is defined as*

$$\mathrm{F}_1(g) = \frac{2\mathbb{P}(g(\boldsymbol{X}, S) = 1,\, Y = 1)}{\mathbb{P}(Y = 1) + \mathbb{P}(g(\boldsymbol{X}, S) = 1)} \enspace .$$

Using the coefficients specified in Example 4.2, we see that for this case $\mathsf{n}_2\mathsf{d}_1 \neq \mathsf{d}_2\mathsf{n}_1$ and condition ($\mathtt{C1}$) is satisfied. Hence, under Assumption 3.1, Theorem 4.3 states that

$$g^*_{(\mathsf{n},\mathsf{d})}(\boldsymbol{x}, s) = \mathbf{1}\left(f^*(\boldsymbol{x}, s) \geq \theta^*_{(\mathsf{n},\mathsf{d})}\right) \enspace ,$$

with $\theta^*_{(\mathsf{n},\mathsf{d})}$ being a unique solution of

$$\mathbb{P}(Y = 1)\theta = \mathbb{E}(f^*(\boldsymbol{X}, S) - \theta)_+ \enspace ,$$

maximizes the F_1-score under the demographic parity constraint. Furthermore, Lemma 4.5 states that for any classifier $g : \mathcal{X} \times [K] \to \{0, 1\}$ such that $g(\boldsymbol{X}, S) \perp\!\!\!\perp S$, it holds that $\mathrm{F}_1\big(g^*_{(\mathsf{n},\mathsf{d})}\big) - \mathrm{F}_1\big(g\big)$ equals to

$$\frac{2\mathbb{E}\big|\eta(\boldsymbol{X}, S) - F^{-1}_{\mu_S(\eta)} \circ F_{\bar{\mu}(\eta)}\big(\theta^*_{(\mathsf{n},\mathsf{d})}\big)\big|\mathbf{1}\left(g^*_{(\mathsf{n},\mathsf{d})}(\boldsymbol{X}, S) {\neq} g(\boldsymbol{X}, S)\right)}{\mathbb{P}(Y = 1) \,+\, \mathbb{P}(g(\boldsymbol{X}, S) = 1)} \, .$$

We invite the reader to compare the above expression with its unconstrained version (Chzhen, 2020, Lemma 2).

5 THE UNAWARENESS CASE

All the previous parts were concerned with the awareness setup—we allowed ourselves to use the sensitive attribute explicitly. However, it can happen in practice that for legal or ethical reasons, the sensitive attribute cannot be used as an input at prediction time (Barocas and Selbst, 2016). Throughout this section we look at classifiers of the form $g : \mathcal{X} \to \{0, 1\}$. By abuse of notation, and as long as confusion cannot occur, we use the same notation $\mathcal{G}$ to denote the set of all classifiers in the unawareness setup. We also need to introduce the conditional distribution of the sensitive attribute S, given the nominally non-sensitive features $\boldsymbol{X}$. For all $s \in [K]$, we set $\tau_s(\boldsymbol{X}) = \mathbb{P}(S = s \mid \boldsymbol{X})$. With one more abuse of notation, we set $\eta(\boldsymbol{X}) \triangleq \mathbb{E}[Y \mid \boldsymbol{X}]$. In this section we look for

$$g^* \in \operatorname*{arg\,min}_{g \in \mathcal{G}} \left\{\mathbb{P}(g(\boldsymbol{X}) \neq Y) \,:\, g(\boldsymbol{X}) \perp\!\!\!\perp S\right\} \enspace . \quad (7)$$

Note that the only difference with the previous setup is the absence of the sensitive input S in the input of g. Lipton et al. (2018) investigated this framework empirically and provided evidence against its use in practice. In particular, they empirically showed that while not permitting using the sensitive attribute S, many algorithms still learn the link between S and $\boldsymbol{X}$ implicitly. Our first result gives a theoretical justification to this phenomenon.

As in the awareness case, we work under a continuity assumption, adapted to this scenario. Recall that Assumption 3.1 imposed continuity of the regression function distribution $\mathrm{Law}(\eta(\boldsymbol{X}, s))$ for each sensitive group $s \in \mathcal{S}$. Here we need a different assumption to account for the fact that S is not accessible anymore, namely the continuity of any linear combination of the regression functions distributions $\eta(\boldsymbol{X})$ and $(\tau_s(\boldsymbol{X}))_{s \in K}$.

Assumption 5.1. *For every $s \in [K]$ and for every vector $c_1, \ldots, c_K \in \mathbb{R}$ such that $c_1 + \ldots + c_K = 0$, the distribution $\mathrm{Law}(\eta(\boldsymbol{X}) + \sum_{\sigma=1}^K \frac{c_\sigma}{p_\sigma}\tau_\sigma(\boldsymbol{X}) \mid S = s)$ is continuous.*

Akin to Theorem 3.2, we derive the explicit form of an optimal fair classifier in the unawareness setting.

Theorem 5.2. *Under Assumption 5.1, a solution g^* defined in Eq. (7) can be expressed for $\boldsymbol{x} \in \mathcal{X}$ as*

$$g^*(\boldsymbol{x}) = \mathbf{1}\left(2\eta(\boldsymbol{x}) - 1 \geq \textstyle\sum_{\sigma=1}^K \lambda^*_\sigma \tau_\sigma(\boldsymbol{x})/p_\sigma\right) \enspace ,$$

where $\boldsymbol{\lambda}^* = (\lambda^*_1, \ldots, \lambda^*_K) \in \mathbb{R}^K$ is a solution of

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^K} \left\{\mathbb{E}\left|2\eta(\boldsymbol{X}) {-} 1 {-} \sum_{\sigma=1}^K \frac{\lambda_\sigma \tau_\sigma(\boldsymbol{X})}{p_\sigma}\right| \,:\, \mathbb{E}\left[\frac{\lambda_S}{p_S}\right] {=} 0\right\} \enspace . \quad (8)$$

We make two observations. First of all, the optimal fair classifier is no longer given by the group-wise threshold. Yet, one can think of the term $\theta(\boldsymbol{x}) \triangleq \sum_{\sigma=1}^K \frac{\lambda^*_\sigma \tau_\sigma(\boldsymbol{x})}{p_\sigma}$ as the $\boldsymbol{x}$-dependent threshold. The optimal classifier g^* tries to guess the value of the sensitive attribute from the features to properly set the threshold. Note that as in the awareness case, here we have $\mathbb{E}[\theta(\boldsymbol{X})] = 0$. Thus, in average, the "threshold" remains being equal to $1/2$ as in the standard classification setup. Secondly, we see that if S is measurable w.r.t. $\boldsymbol{X}$, we fall back to the awareness case. Otherwise each λ^*_s is weighted by the conditional distribution of $S \mid \boldsymbol{X}$.

Importantly, it is remains an open problem to give a connection of the above problem with the corresponding regression setup. The main reason for it is the current lack of an explicit solution to the optimal fair regression problem in the unawareness case. Some attempts were made in (Chzhen and Schreuder, 2020), yet they are unsatisfactory and do not give a complete picture. Intuitively, the difficulty of extending the optimal transport based approach to the unawareness setup lies in our inability to establish the source of a given $\boldsymbol{x}$. In other words, given $\boldsymbol{x}$, we have no idea which of $\mathbb{P}_{\boldsymbol{X}|S=1}, \ldots, \mathbb{P}_{\boldsymbol{X}|S=K}$ it was sampled from. Hence, we cannot build a transport map from $\operatorname{Law}(\eta(\boldsymbol{X}, S) \mid S = s)$ to their common barycenter since it requires the knowledge of S. Naively, one might think to use $\hat{S}(\boldsymbol{X})$—the best prediction of S given $\boldsymbol{X}$—instead of S. While intuitive, it is easy to see that simply replacing S by $\hat{S}(\boldsymbol{X})$ in Theorem 3.3 does not even satisfy the demographic parity constraint in general. As we show in the next paragraph, the connection between the fair classification and fair regression can be made explicit in the unawareness case if we consider the case of $K = 2$. The existence of such a connection is explained by the Hahn decomposition theorem for signed measure, whose generalization (even its formulation) to many measures is unclear.

Binary sensitive attribute: the $(\mathbb{P} \to \mathbb{P}^\star)$ reduction. In this section we describe a reduction of the fair unaware binary classification problem to the awareness case for $K = 2$. First of all, let us recall that the minimization of $\mathbb{P}(Y \neq g(\boldsymbol{X}, S))$ over g under any constraints is equivalent to the minimization of $\mathbb{E}[g(\boldsymbol{X}, S)(1-2\eta(\boldsymbol{X}, S))]$ under the same constraints. Furthermore, the same applies to the awareness case where we only need to replace $\eta(\boldsymbol{X}, S)$ by $\eta(\boldsymbol{X})$.

For our reduction, given a distribution $\mathbb{P}$ on $\mathcal{X} \times \{1, 2\} \times \{0, 1\}$, we build *another* distribution $\mathbb{P}^\star$ on $\mathcal{X} \times \{1, 2\}$ and a function $\tilde{\eta} : \mathcal{X} \times \{1, 2\} \to [0, +\infty)$ with the following property: there is a one-to-one correspondence between a solution $g^*_{\mathbb{P}}$ of

$$\min\left\{\mathbb{E}_{\mathbb{P}}[g(\boldsymbol{X})(1 - 2\eta(\boldsymbol{X}))] \,:\, g(\boldsymbol{X}) \perp\!\!\!\perp_{\mathbb{P}} S\right\} \enspace ,$$

and a solution $g^*_{\mathbb{P}^\star}$ of

$$\min\left\{\mathbb{E}_{\mathbb{P}^\star}[g(\boldsymbol{X}, S)(1 - 2\tilde{\eta}(\boldsymbol{X}, S))] \,:\, g(\boldsymbol{X}, S) \perp\!\!\!\perp_{\mathbb{P}^\star} S\right\} \enspace .$$

In other words, if $g^*_{\mathbb{P}^\star}$ is an optimal fair classifier for distribution $\mathbb{P}^\star$ under *awareness*, then $g^*_{\mathbb{P}^\star}$ can be transformed into an optimal fair classifier $g^*_{\mathbb{P}}$ for $\mathbb{P}$ under *unawareness*. In what follows, we present the reduction and, given the distribution $\mathbb{P}$, explain the procedure to build $\mathbb{P}^\star$.

Let $\mathsf{TV} \triangleq \frac{1}{2}\int \left|\mathrm{d}\mathbb{P}_{\boldsymbol{X}|S=1} - \mathrm{d}\mathbb{P}_{\boldsymbol{X}|S=2}\right|$. Note that if $\mathsf{TV} = 0$, then $\boldsymbol{X} \perp\!\!\!\perp S$ and any unaware classifier satisfies the demographic parity constraint. Hence, we assume that $\mathsf{TV} \in (0, 1]$. We define $\mathbb{P}^\star$ in three steps.
Step 1. The distribution of $\boldsymbol{X}$ given S under $\mathbb{P}^\star$ is

$$\mathbb{P}^\star_{\boldsymbol{X}|S=s} = (\mathbb{P}_{\boldsymbol{X}|S=s} - \mathbb{P}_{\boldsymbol{X}|S\neq s})_+/\mathsf{TV} \enspace ,$$

where $(\mathbb{P}_{\boldsymbol{X}|S=1} - \mathbb{P}_{\boldsymbol{X}|S=2})_+$ and $(\mathbb{P}_{\boldsymbol{X}|S=2} - \mathbb{P}_{\boldsymbol{X}|S=1})_+$ is the Hahn decomposition of the signed measure $\mathbb{P}_{\boldsymbol{X}|S=2} - \mathbb{P}_{\boldsymbol{X}|S=1}$ (see, e.g., Billingsley, 2008, Theorem 32.1);
Step 2. the distribution of S under $\mathbb{P}^\star$ is defined as: $\mathbb{P}^\star(S = 1) = \mathbb{P}^\star(S = 2) = 1/2$;
Step 3. the new pseudo-regression function $\tilde{\eta}$ is defined as

$$\tilde{\eta}(\boldsymbol{x}, s) = \frac{1}{2} + \frac{\mathsf{TV}}{2} \cdot \frac{2\eta(\boldsymbol{x}) - 1}{|(\tau_1(\boldsymbol{x})/p_1) - (\tau_2(\boldsymbol{x})/p_2)|}$$

for $\boldsymbol{x} \in \operatorname{supp}(\mathbb{P}^\star_{\boldsymbol{X}|S=1}) \cap \operatorname{supp}(\mathbb{P}^\star_{\boldsymbol{X}|S=2})$;
We note that under $\mathbb{P}^\star$, the sensitive attribute S is measurable w.r.t. $\boldsymbol{X}$ since the supports of $\mathbb{P}^\star_{\boldsymbol{X}|S=1}$ and $\mathbb{P}^\star_{\boldsymbol{X}|S=2}$ do not intersect. We refer $\tilde{\eta}$ as to the pseudo-regression function since it is not guaranteed that it takes values in $[0, 1]$ and, hence, is not necessary a valid regression function of $Y \mid \boldsymbol{X}$ under $\mathbb{P}^\star$ for $Y \in \{0, 1\}$.

Proposition 5.3 (Unawareness to awareness reduction)**.** *Let $\mathbb{P}$ be any distribution on $\mathcal{X} \times \{1, 2\} \times \{0, 1\}$. Let $\mathbb{P}^\star$ and $\tilde{\eta}$ be defined using the three steps procedure described above and $g^*_{\mathbb{P}^\star}$ be any solution of*

$$\min\left\{\mathbb{E}_{\mathbb{P}^\star}[g(\boldsymbol{X}, S)(1 - 2\tilde{\eta}(\boldsymbol{X}, S))] \,:\, g(\boldsymbol{X}, S) \perp\!\!\!\perp_{\mathbb{P}^\star} S\right\} \enspace .$$

Then, $g^*_{\mathbb{P}} : \mathcal{X} \to \{0, 1\}$ defined point-wise as

$$g^*_{\mathbb{P}}(\boldsymbol{x}) = \begin{cases} g^*_{\mathbb{P}^\star}(\boldsymbol{x}, 1) & \boldsymbol{x} \in \operatorname{supp}(\mathbb{P}^\star_{\boldsymbol{X}|S=1}) \\ g^*_{\mathbb{P}^\star}(\boldsymbol{x}, 2) & \boldsymbol{x} \in \operatorname{supp}(\mathbb{P}^\star_{\boldsymbol{X}|S=2}) \\ \mathbf{1}\,(\eta(\boldsymbol{x}) \geq 1/2) & \textit{otherwise} \end{cases} \enspace ,$$

is a solution of $\min\left\{\mathbb{E}_{\mathbb{P}}[g(\boldsymbol{X})(1 - 2\eta(\boldsymbol{X}))] \,:\, g(\boldsymbol{X}) \perp\!\!\!\perp_{\mathbb{P}} S\right\}$.

The above result provide a theoretical justification to the empirical observations made by Lipton et al. (2018). Indeed, they have empirically shown that in the unawareness setting, many classification algorithms tailored for the demographic parity constraint, are forced to "guess" the sensitive attribute S. Theoretically, this is reflected by the construction of the distribution $\boldsymbol{X} \mid S$ under $\mathbb{P}^\star$. Furthermore, since the reduction is performed to the awareness setup, the results of previous sections on the connection between fair regression and fair classification still applies. Yet, we emphasize that the above argument is only valid for $K = 2$ and its extension to $K > 2$ remains an open problem. The main difficulty comes from the absence of a version of the Hahn decomposition for more than two measures.

6 FAIR LEARNING: FROM INFINITE TO FINITE SAMPLE

All the previous sections were concerned with the "infinite sample" regime—the case of known distribution $\mathbb{P}$. While not being the main focus of the paper, given the established connection with the problem of fair regression, one can easily pass from the infinite to the finite-sample regime. Indeed, there are many algorithms that allow to consistently estimate the optimal fair score function f^*. For instance, Agarwal et al. (2019) give an in-processing algorithm with provable finite sample generalization bounds; Le Gouic et al. (2020) propose a consistent estimator of f^*; Chzhen et al. (2020b)

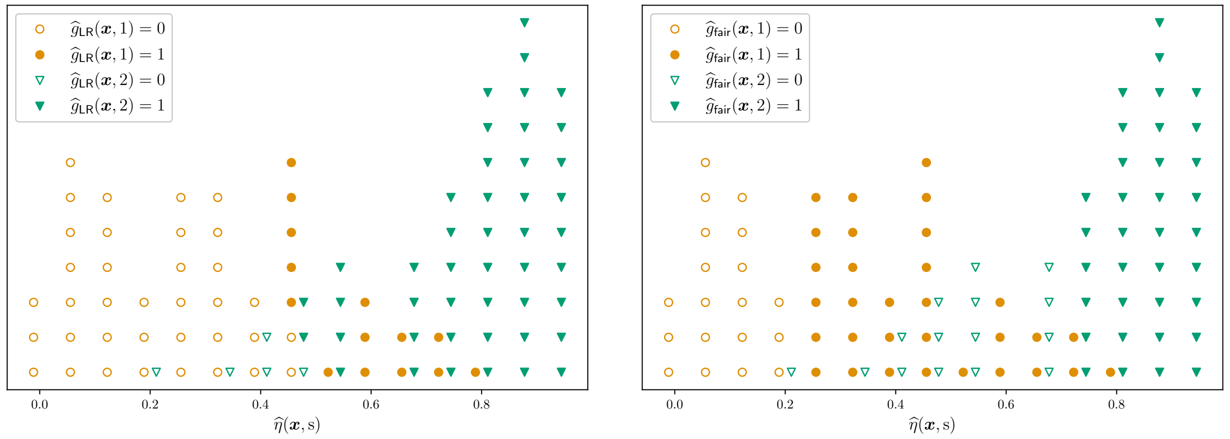

Figure 2: Empirical group-wise outcomes for logistic regression (`left`) and the proposed post-processed fair plug-in classifier (`right`) for the misclassification risk. The data was generated as follows. We drew N samples for each sensitive class as $\boldsymbol{X}|S{=}1 \sim \mathcal{N}(-1,1)$, $\boldsymbol{X}|S{=}2 \sim \mathcal{N}(1,1)$ and $Y|\boldsymbol{X}{=}\boldsymbol{x} \sim \text{Bernoulli}(1/(1{+}e^{-\boldsymbol{x}}))$. We fitted a Logistic Regression classifier $\widehat{g}_{\mathsf{LR}}$ using `scikit-learn` (Pedregosa et al., 2011). For the fair plug-in classifier, we obtained an estimator $\hat{f}$ of the optimal fair regression function f^* following Chzhen et al. (2020b) and we considered $\hat{g}_{\mathsf{fair}}(\boldsymbol{x},s) \triangleq \mathbf{1}(\hat{f}(\boldsymbol{x},s) \geq 1/2)$.

provide an algorithm with finite sample fairness and risk guarantees; Chzhen and Schreuder (2022) exhibit a modification of the two aforementioned estimators that enjoys stronger fairness and risk guarantees.

Once an estimator $\hat{f}$ of f^* is constructed, one only needs to estimate the threshold θ^* specified in Theorem 4.3. Recall that there are two cases considered in Theorem 4.3, the first one requires finding a root of a specific function and the second one gives an explicit expression for θ^*. For the first case one can use the *unsupervised* approach recycling $\hat{f}$ and only estimating $\mathbb{E}_{\boldsymbol{X}|S}[\cdot]$ and the, potentially distribution dependent coefficients, $(\mathsf{n}_0, \mathsf{n}_1, \mathsf{n}_2)$, $(\mathsf{d}_0, \mathsf{d}_1, \mathsf{d}_2)$. For the second case one only needs to estimate or substitute the values of $(\mathsf{n}_0, \mathsf{n}_1, \mathsf{n}_2)$, $(\mathsf{d}_0, \mathsf{d}_1, \mathsf{d}_2)$. Alternatively, for the threshold estimation, one can deploy the grid-search technique proposed by Koyejo et al. (2014) by again recycling the base estimator $\hat{f}$ of f^*. In either case one ends up with a flexible and rather direct approach for building data-driven algorithms. We note however that the second approach requires additional labeled data, while the first one is only based on the unlabeled data. The final classification algorithm eventually takes the form of $\mathbf{1}(\hat{f}(\boldsymbol{x},s) \geq \hat{\theta})$.

As a proof of concept, we have implemented the described approach on a toy example for the misclassification risk. A description of the considered toy problem as well as empirical results are provided in Figure 4. Figure 4 (`left`) displays predictions without fairness constraints for both groups; Figure 4 (`right`) displays predictions of the procedure described above. Appendix D contains additional experiments and figures. In particular, we display an empirical counter-part to Figure 1 and provide empirical results for other risk measures such as the F_b-score and Jaccard.

7 CONCLUSION

We have derived an explicit connection between the regression and classification under the demographic parity constraint problems. Leveraging the optimal transport interpretation of the optimal fair regressor, we have shown that the regression-classification link is akin to the classical unconstrained setup. As a by-product of this result, we have derived an exact expression for the Price of Fairness. This connection is extended to non-decomposable performance measures and, remarkably, amounts to replacing the standard regression function by its fair counterpart. Finally, we have provided a reduction scheme to pass from the unawareness setup to the awareness setup in the case of the binary sensitive attribute, hence giving the first explicit solution of the fair optimal unaware classifier. Our results are instructive and, relying on the previous studies, lead to wide spectrum of algorithms that can be used with non-decomposable measures. Future works will be focused on further clarification of other notions of fairness constraint by providing clean and interpretable theoretical studies.

Acknowledgements

We would like to thank the reviewers for their comments that helped to improve the paper. The work of E. Chzhen was supported by ANR PIA funding: ANR-20-IDEES-0002. N. Schreuder acknowledges the financial support of the European Research Council (grant SLING 819789). The work of S. Gaucher was supported by the French ANR Chair in Artificial Intelligence TopAI ANR-19-CHIA-001.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. *International Conference on Machine Learning*.

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

Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

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

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California law review*, pages 671–732.

Bascol, K., Emonet, R., Fromont, E., Habrard, A., Metzler, G., and Sebban, M. (2019). From cost-sensitive to tight f-measure bounds. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1245–1253. PMLR.

Bauschke, H. H. and Combettes, P. L. (2017). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer.

Bertsekas, D. P. (1973). Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9).

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Biswas, S. and Rajan, H. (2020). Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In Devanbu, P., Cohen, M. B., and Zimmermann, T., editors, *ESEC/FSE '20: 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Virtual Event, USA, November 8-13, 2020*, pages 642–653. ACM.

Biswas, S. and Rajan, H. (2021). Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In Spinellis, D., Gousios, G., Chechik, M., and Penta, M. D., editors, *ESEC/FSE '21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, August 23-28, 2021*, pages 981–993. ACM.

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6):e0177678.

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

Celis, E., Huang, L., Keswani, V., and Vishnoi, N. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328.

Chen, Z., Zhang, J. M., Sarro, F., and Harman, M. (2022). A comprehensive empirical study of bias mitigation methods for software fairness. *CoRR*, abs/2207.03277.

Chiappa, S., Jiang, R., Stepleton, T., Pacchiano, A., Jiang, H., and Aslanides, J. (2020). A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3633–3640.

Chiappa, S. and Pacchiano, A. (2021). Fairness with continuous optimal transport. *arXiv preprint arXiv:2101.02084*.

Chinchor, N. (1992). MUC-4 evaluation metrics. In *Proceedings of the 4th Conference on Message Understanding, MUC 1992, McLean, Virginia, USA, June 16-18, 1992*, pages 22–29. ACL.

Choi, S.-S., Cha, S.-H., and Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of systemics, cybernetics and informatics*, 8(1):43–48.

Chzhen, E. (2020). Optimal rates for nonparametric f-score binary classification via post-processing. *Mathematical Methods of Statistics*, 29(2):87–105.

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

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020b). Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*.

Chzhen, E. and Schreuder, N. (2020). An example of prediction which complies with demographic parity and equalizes group-wise risks in the context of regression. In *NeurIPS 2020 Workshop on Algorithmic Fairness through the Lens of Causality and Interpretability*.

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

Dablain, D., Krawczyk, B., and Chawla, N. (2022). Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning. *arXiv preprint arXiv:2207.06084*.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

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

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

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

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jasinska, K., Dembczynski, K., Busa-Fekete, R., Pfannschmidt, K., Klerx, T., and Hullermeier, E. (2016). Extreme f-measure maximization using sparse probability estimates. In *International conference on machine learning*, pages 1435–1444. PMLR.

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

Kotlowski, W. and Dembczyński, K. (2016). Surrogate regret bounds for generalized classification performance metrics. In *Asian Conference on Machine Learning*, pages 301–316. PMLR.

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

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., and Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. *Advances in neural information processing systems*, 27.

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

Lipton, Z., Chouldechova, A., and McAuley, J. (2018). Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8136–8146.

Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366.

Menon, A., Narasimhan, H., Agarwal, S., and Chawla, S. (2013). On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pages 603–611. PMLR.

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

Michele, D., Ben-David, S., Pontil, M., and Shawe-Taylor, J. (2017). An efficient method to impose fairness in linear models. In *NIPS Workshop on Prioritising Online Content*.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT Press.

Narasimhan, H. (2018). Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 1646–1654. PMLR.

Narasimhan, H., Kar, P., and Jain, P. (2015). Optimizing non-decomposable performance measures: A tale of two classes. In *International Conference on Machine Learning*, pages 199–208. PMLR.

Oneto, L., Donini, M., and Pontil, M. (2020). General fair empirical risk minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media.

Schreuder, N. and Chzhen, E. (2021). Classification with abstention but without disparities. In *Uncertainty in Artificial Intelligence*, pages 1227–1236. PMLR.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Van Rijsbergen, C. J. (1974). Foundation of evaluation. *Journal of documentation*, 30(4):365–373.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wang, Y. and Singh, L. (2021). Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 12(2):101–119.

Wick, M., Tristan, J.-B., et al. (2019). Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32.

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

Yang, F., Cisse, M., and Koyejo, S. (2020). Fairness with overlapping groups; a probabilistic perspective. *Advances in neural information processing systems*, 33:4067–4078.

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

Zeng, X., Dobriban, E., and Cheng, G. (2022). Bayes-optimal classifiers under group fairness. *arXiv preprint arXiv:2202.09724*.

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

Supplementary material for “Fair learning with Wasserstein barycenters for non-decomposable performance measures”

Additional notation For any probability measure μ on $\mathcal{X}$ and a function $f : \mathcal{X} \to \mathbb{R}$, we denote by $f\sharp\mu$, the image measure of μ. For any univariate measure μ, we denote by F_μ its cumulative distribution, and by F_μ^{-1} its quantile function, given by $F_\mu^{-1}(p) \triangleq \min\{x : \mu((-\infty, x]) \geq p\}$.

A A UNIFIED PROOF FOR DERIVING OPTIMAL FAIR CLASSIFIERS

In this section we state and prove a general result which implies both Theorem 3.2 and Theorem 5.2. On top of the problem setup presented in Section 2, let $\boldsymbol{W}$ be a random variable taking its values in some abstract space $\mathcal{W}$. Moreover, define the regression functions $\tau_s(\boldsymbol{w}) \triangleq \mathbb{P}\left(S = s \mid \boldsymbol{W} = \boldsymbol{w}\right), s \in [K]$. The random variable $\boldsymbol{W}$ should be thought as $(\boldsymbol{X}, S)$ for the awareness setting and $\boldsymbol{X}$ for the unawareness setting. Our goal is to find a solution

$$g^* \in \arg\min_{g \in \mathcal{G}} \left\{\mathbb{P}(g(\boldsymbol{W}) \neq Y) \,:\, g(\boldsymbol{W}) \perp\!\!\!\perp S\right\} \quad . \tag{9}$$

The general result will be stated under the following continuity assumption. It requires continuity of the distribution of any linear combination of the regression functions evaluated at $\boldsymbol{W}$.

Assumption A.1. *For every $s \in [K]$ and for every vector $\boldsymbol{c} = (c_1, \ldots, c_K)^\top \in \mathbb{R}^K$ such that $c_1 + \ldots + c_K = 0$, the distribution* $\mathrm{Law}(\eta(\boldsymbol{W}) + \sum_{\sigma=1}^{K} \frac{c_\sigma}{p_\sigma} \tau_\sigma(\boldsymbol{W}) \mid S = s)$ *is continuous.*

Akin to Assumptions 3.1 and 5.1, Assumption A.1 is not necessary to prove our result but it greatly simplifies its presentation and interpretation. Let us now state the general result which encompasses the two special cases presented in the main body of the paper.

Theorem A.1: Fair optimal classifier (unified version)

Let Assumption A.1 be satisfied. Then a solution g^* defined in Eq. (9) can be expressed for all $\boldsymbol{w} \in \mathcal{W}$ as

$$g^*(\boldsymbol{w}) = \mathbf{1}\left(2\eta(\boldsymbol{w}) - 1 \geq \sum_{\sigma=1}^{K} \frac{\lambda_\sigma^* \tau_\sigma(\boldsymbol{w})}{p_\sigma}\right) \quad ,$$

where $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_K^*) \in \mathbb{R}^K$ is a solution of

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^K} \left\{\mathbb{E}\left[\left|2\eta(\boldsymbol{W}) - 1 - \sum_{\sigma=1}^{K} \frac{\lambda_\sigma \tau_\sigma(\boldsymbol{W})}{p_\sigma}\right|\right] \,:\, \mathbb{E}\left[\frac{\lambda_S}{p_S}\right] = 0\right\} \quad . \tag{10}$$

Remark A.2 (Relating the above result to the main body)**.** *It is straightforward to derive Theorem 3.2 and Theorem 5.2 from Theorem A.1. Indeed, to prove Theorem 3.2, set $\boldsymbol{W} = (\boldsymbol{X}, S), \boldsymbol{w} = (\boldsymbol{x}, s)$ and notice that $\tau_\sigma(\boldsymbol{w}) = \mathbb{P}(S = \sigma \mid \boldsymbol{X} = \boldsymbol{x}, S = s) = \delta_s(\sigma)$. In particular, Assumption A.1 is weaker than Assumption 5.1 and one can check that the optimal fair classifiers coincide. Similarly, Theorem 5.2 can be derived from Theorem A.1 by setting $\boldsymbol{W} = \boldsymbol{X}, \boldsymbol{w} = \boldsymbol{x}$.*

Proof of Theorem A.1. One can verify that the minimization of $\mathbb{P}(g(\boldsymbol{W})) \neq Y)$ over g is equivalent to the minimization of $\mathbb{E}[g(\boldsymbol{W})(1 - 2\eta(\boldsymbol{W}))]$. Furthermore, the demographic parity constraint can be equivalently expressed as

$$\mathbb{E}[g(\boldsymbol{W}) \mid S = s] = \mathbb{E}[g(\boldsymbol{W})]\,,\ s \in [K] \quad .$$

Thus, we are interested in the solution of the optimization problem

$$\min_{g \in \mathcal{G}} \left\{\sum_{s \in [K]} p_s \mathbb{E}[g(\boldsymbol{W})(1 - 2\eta(\boldsymbol{W})) \mid S = s] \,:\, \mathbb{E}[g(\boldsymbol{W}) \mid S = s] = \mathbb{E}[g(\boldsymbol{W})]\,,\ s \in [K]\right\} \quad .$$

Recall that we defined the random variable $\tau_s(\boldsymbol{W}) = \mathbb{P}\left(S = s \mid \boldsymbol{W}\right), s \in [K]$. The Lagrangian for the above problem can be expressed as

$$\mathcal{L}(g, \boldsymbol{\lambda}) = \mathbb{E}\left[g(\boldsymbol{W})\left((1 - 2\eta(\boldsymbol{W})) - \sum_{\sigma=1}^{K} \lambda_\sigma (1 - p_\sigma^{-1}\tau_\sigma(\boldsymbol{W}))\right)\right] \enspace,$$

where $\boldsymbol{\lambda} \in \mathbb{R}^K$. Weak duality implies that

$$\min_g \max_{\boldsymbol{\lambda}} \mathcal{L}(g, \boldsymbol{\lambda}) \geq \max_{\boldsymbol{\lambda}} \min_g \mathcal{L}(g, \boldsymbol{\lambda}) \enspace . \tag{11}$$

Our approach to derive the optimal fair classifier can be decomposed in two classical steps: find optimal solutions to the dual problem $\max_{\boldsymbol{\lambda}} \min_g \mathcal{L}(g, \boldsymbol{\lambda})$; show that strong duality holds so that the optimal solutions to the dual problem are also optimal for the primal problem.

Solving the dual problem. In what follows we focus our attention on the dual $\max\min$ problem, which can be solved analytically. We first solve for any $\boldsymbol{\lambda}$ the inner minimization problem of the $\max\min$ formulation

$$\min_g \mathcal{L}(g, \boldsymbol{\lambda}) \enspace . \tag{12}$$

Since g can be any function from $\mathcal{W}$ to $\{0, 1\}$, the above problem can be solved point-wise. In particular, one can check that the solution is given by

$$g^*(\boldsymbol{w}) = \mathbf{1}\left(2\eta(\boldsymbol{w}) - 1 \geq \sum_{\sigma=1}^{K} \lambda_\sigma (p_\sigma^{-1}\tau_\sigma(\boldsymbol{w}) - 1)\right) \enspace .$$

Plugging the optimal solution g^* back in the dual problem, we obtain as solution of the outer maximization problem

$$\boldsymbol{\lambda}^* \in \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^K} \mathbb{E}\left[\left(2\eta(\boldsymbol{W}) - 1 + \sum_{\sigma=1}^{K} \lambda_\sigma (1 - p_\sigma^{-1}\tau_\sigma(\boldsymbol{W}))\right)_+\right] \enspace . \tag{13}$$

The objective of the above optimization problem is non-negative, continuous convex as a function of $\boldsymbol{\lambda}$. Lemma A.3 ensures that $\boldsymbol{\lambda}^*$ exists.

The objective function of problem in Eq. (13) is not smooth everywhere due to the presence of the positive part function. However, thanks to Assumption A.1, the set of points at which the objective function is not differentiable has zero Lebesgue measure and can thus be ignored (see, e.g., Bertsekas, 1973, Proposition 3). The First-Order Optimality Condition (FOOC) on the optimal Lagrange multiplier $\boldsymbol{\lambda}^*$ then reads as

$$\mathbb{E}[p_s^{-1}\tau_s(\boldsymbol{W})\mathbf{1}\left(g^*(\boldsymbol{W}) = 1\right)] = \mathbb{P}(g^*(\boldsymbol{W}) = 1)\,,\ \forall s \in [K] \enspace .$$

The LHS of the above inequality can be simplified into

$$\mathbb{E}[p_s^{-1}\tau_s(\boldsymbol{W})\mathbf{1}\left(g^*(\boldsymbol{W}) = 1\right)] = \sum_{s=1}^{K} \mathbb{E}[\tau_s(\boldsymbol{W})\mathbf{1}\left(g^*(\boldsymbol{W}) = 1\right) \mid S{=}s] = \mathbb{P}(g^*(\boldsymbol{W}) = 1 \mid S{=}s) \enspace ,$$

showing that the FOOC on $\boldsymbol{\lambda}^*$ is equivalent to g^* satisfying DP.

Strong duality. The above reasoning showed that g^* defined with the optimal Lagrange multiplier $\boldsymbol{\lambda}^*$ is feasible for the primal problem. Combining this property with Eq. (11) implies that g^* is also a solution of the primal problem.

A more convenient expression. Using the fact that $2(a)_+ = a + |a|$ and $\mathbb{E}\tau_s(\boldsymbol{W}) = p_s$, we can express the optimal Lagrange multiplier $\boldsymbol{\lambda}^*$ as

$$\boldsymbol{\lambda}^* \in \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^K} \mathbb{E}\left[\left|2\eta(\boldsymbol{W}) - 1 + \sum_{\sigma=1}^{K} \lambda_\sigma (1 - p_\sigma^{-1}\tau_\sigma(\boldsymbol{W}))\right|\right] \enspace .$$

Moreover, introducing $G(\boldsymbol{\lambda}) = \mathbb{E}\left[\left|2\eta(\boldsymbol{W}) - 1 + \sum_{\sigma=1}^{K}\lambda_\sigma(1 - p_\sigma^{-1}\tau_\sigma(\boldsymbol{W}))\right|\right]$, we observe that for any $c \in \mathbb{R}$ and $\boldsymbol{\lambda} \in \mathbb{R}^K$ it holds that $G(\boldsymbol{\lambda}) = G(\boldsymbol{\lambda} + c\boldsymbol{p})$, where $\boldsymbol{p} = (p_1, \ldots, p_K)^\top \in \mathbb{R}^K$. Hence, since we are interested in any solution of the above optimization problem, we can define $(g^*, \boldsymbol{\lambda}^*)$ as

$$g^*(\boldsymbol{w}) = \mathbf{1}\left(2\eta(\boldsymbol{w}) - 1 \geq \sum_{\sigma=1}^{K}\lambda_\sigma p_\sigma^{-1}\tau_\sigma(\boldsymbol{w})\right) \enspace,$$

$$\boldsymbol{\lambda}^* \in \mathop{\arg\min}_{\boldsymbol{\lambda}\in\mathbb{R}^K}\left\{\mathbb{E}\left[\left|2\eta(\boldsymbol{W}) - 1 - \sum_{\sigma=1}^{K}\lambda_\sigma p_\sigma^{-1}\tau_\sigma(\boldsymbol{W})\right|\right] \;:\; \bar{\boldsymbol{\lambda}} = 0\right\} \enspace. \qquad \square$$

Lemma A.3. *Let Assumption A.1 be satisfied, then the mapping*

$$\boldsymbol{\lambda} \mapsto \mathbb{E}\left[\left(2\eta(\boldsymbol{W}) - 1 + \sum_{\sigma=1}^{K}\lambda_\sigma(1 - p_\sigma^{-1}\tau_\sigma(\boldsymbol{W}))\right)_+\right] \tag{14}$$

attains its minimum.

Proof. In the end of the proof of Theorem A.1 we have show that minimization of (14) is equivalent to the minimization of

$$\boldsymbol{\lambda} \mapsto \mathbb{E}\left[\left|2\eta(\boldsymbol{W}) - 1 - \sum_{\sigma=1}^{K}\lambda_\sigma p_\sigma^{-1}\tau_\sigma(\boldsymbol{W})\right|\right]$$

on the hyperplane $\left\{\boldsymbol{\lambda} \in \mathbb{R}^K \;:\; \bar{\boldsymbol{\lambda}} = 0\right\}$. Thus, it is sufficient to show that

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^K}\left\{\mathbb{E}\left[\left|2\eta(\boldsymbol{W}) - 1 - \sum_{\sigma=1}^{K}\lambda_\sigma p_\sigma^{-1}\tau_\sigma(\boldsymbol{W})\right|\right] \;:\; \bar{\boldsymbol{\lambda}} = 0\right\}$$

is attained.

It is clear that the mapping in question is convex on $\mathbb{R}^K$. Hence, it is sufficient to show that it is coercive (see e.g. Bauschke and Combettes, 2017, Proposition 11.15). It holds that

$$\mathbb{E}\left|2\eta(\boldsymbol{W}) - 1 - \sum_{\sigma=1}^{K}\lambda_\sigma p_\sigma^{-1}\tau_\sigma(\boldsymbol{W})\right| = \mathbb{E}\left|\langle(\boldsymbol{\lambda}/\boldsymbol{p}, 1), (\boldsymbol{V}, H)\rangle\right| \enspace, \tag{15}$$

where we introduced the vector $\boldsymbol{V} \triangleq (\tau_1(\boldsymbol{W}), \ldots, \tau_K(\boldsymbol{W}))$, $H \triangleq 1 - 2\eta(\boldsymbol{W})$, and $(\boldsymbol{\lambda}/\boldsymbol{p}, 1) \triangleq (\lambda_1/p_1, \ldots, \lambda_K/p_K, 1) \in \mathbb{R}^{K+1}$. Thus, in view of (15), by Markov's inequality, for any $\kappa > 0$ it holds that

$$\mathbb{E}\left|2\eta(\boldsymbol{W}) - 1 - \sum_{\sigma=1}^{K}\frac{\lambda_\sigma}{p_\sigma}\tau_\sigma(\boldsymbol{W})\right| \geq \kappa\|(\boldsymbol{\lambda}/\boldsymbol{p}, 1)\|\mathbb{P}(|\langle(\boldsymbol{\lambda}/\boldsymbol{p}, 1), (\boldsymbol{V}, H)\rangle| > \kappa\|(\boldsymbol{\lambda}/\boldsymbol{p}, 1)\|) \enspace, \tag{16}$$

where $\|\cdot\|$ denotes the Euclidean norm. Note that if we are able to show that for some $\kappa_0 > 0$, the right hand side of the above inequality is bounded away from zero, the proof of coercivity is concluded since $\|(\boldsymbol{\lambda}/\boldsymbol{p}, 1)\| \geq \min_{s\in[K]}\{p_s^{-1}\}\|\boldsymbol{\lambda}\|$. To this end, let us introduce

$$F(\boldsymbol{u}, t) = \mathbb{P}(|\langle\boldsymbol{u}, (\boldsymbol{V}, H)\rangle| \leq t) \enspace,$$

for all $t \geq 0$ and $\boldsymbol{u} \in \mathcal{H}_0$ being defined as

$$\mathcal{H}_0 = \left\{\boldsymbol{u} \in \mathbb{R}^{K+1} \;:\; \|\boldsymbol{u}\| = 1,\; \boldsymbol{u} = (\lambda_1/p_1, \ldots, \lambda_K/p_k, 1) \text{ for some } \lambda_1 + \ldots + \lambda_K = 0\right\} \enspace.$$

By Assumption A.1, for any $\boldsymbol{u} \in \mathcal{H}_0$, the mapping $t \mapsto F(\boldsymbol{u}, t)$ is continuous on $(0, +\infty)$ with $F(\boldsymbol{u}, 0) = 0$ and $F(\boldsymbol{u}, +\infty) = 1$. Furthermore, for any $\boldsymbol{u} \in \mathcal{H}_0, \boldsymbol{h} \in \mathbb{R}^{K+1}$ such that $\boldsymbol{u} + \boldsymbol{h} \in \mathcal{H}_0$ and for any $\delta > 0, t > 0$, we have thanks to triangle's inequality and monotonicity of $F(\boldsymbol{u}, \cdot)$

$$F(\boldsymbol{u} + \boldsymbol{h}, t + \delta) \in \Big[F(\boldsymbol{u}, t + \delta - 2\|\boldsymbol{h}\|),\; F(\boldsymbol{u}, t + \delta + 2\|\boldsymbol{h}\|)\Big] \xrightarrow[]{\substack{\delta \longrightarrow 0 \\ \|\boldsymbol{h}\| \longrightarrow 0}} F(\boldsymbol{u}, t) \enspace,$$

where the convergence follows from the assumed continuity of $F(\boldsymbol{u}, \cdot)$. Thus, $(\boldsymbol{u}, t) \mapsto F(\boldsymbol{u}, t)$ is continuous. Since $\mathcal{H}_0$ is compact, we have that

$$G(t) \triangleq \sup_{\boldsymbol{u} \in \mathcal{H}_0} F(\boldsymbol{u}, t) \enspace ,$$

is continuous on $[0, +\infty)$. Hence, the intermediate value theorem guarantees that there exists $\kappa_0 > 0$ such that

$$G(\kappa_0) = 1 - \inf_{\lambda_1 + \ldots + \lambda_K = 0} \mathbb{P}(|\langle(\boldsymbol{\lambda}/\boldsymbol{p}, 1), (\boldsymbol{V}, H)\rangle| > \kappa_0 \|(\boldsymbol{\lambda}/\boldsymbol{p}, 1)\|) = \frac{1}{2} \enspace .$$

In view of Eq. (16), we conclude. □

B OMITTED PROOFS

Proof of Theorem 3.3. Theorem 3.2 implies that under Assumption 3.1 the optimal classifier is of the form $g^*(x, s) = \mathbf{1}\left(\eta(\boldsymbol{x}, s) \geq \beta^*_s\right)$ for some $\boldsymbol{\beta}^* = (\beta^*_s)_{s\in[K]} \in \mathbb{R}^K$. It follows from (Van der Vaart, 2000, Lemma 21.1(iv)) and Assumption 3.1 that $\eta(\boldsymbol{x}, s) = F^{-1}_{\mu_s(\eta)} \circ F_{\mu_s(\eta)}(\eta(\boldsymbol{x}, s))$ for almost all $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. $\mathbb{P}_{\boldsymbol{X}|S=s}$. Thus, it is sufficient to look at the classifiers of the form

$$g(\boldsymbol{x}, s) = \mathbf{1}\left(F^{-1}_{\mu_s(\eta)} \circ F_{\mu_s(\eta)}(\eta(\boldsymbol{x}, s)) \geq \beta_s\right) \enspace ,$$

or, equivalently, at $g(\boldsymbol{x}, s) = \mathbf{1}\left(F_{\mu_s(\eta)}(\eta(\boldsymbol{x}, s)) \geq F_{\mu_s(\eta)}(\beta_s)\right)$ (Van der Vaart, 2000, Lemma 21.1(i)). Now, the inverse transform theorem states that under Assumption 3.1, $F^{-1}_{\mu_s(\eta)}(U)$ has the same distribution as $\eta(\boldsymbol{X}, S)$ conditionally on $S = s$, for U uniformly distributed on $(0, 1)$. Then,

$$\mathbb{P}\left(g(\boldsymbol{X}, S) = 1 \mid S = s\right) = \mathbf{P}\left(F_{\mu_s(\eta)} \circ F^{-1}_{\mu_s(\eta)}(U) \geq F_{\mu_s(\eta)}(\beta_s)\right) = 1 - F_{\mu_s(\eta)}(\beta_s) \enspace ,$$

where we have used that $F_{\mu_s(\eta)} \circ F^{-1}_{\mu_s(\eta)}(u) = u$ for all $u \in (0, 1)$ (Van der Vaart, 2000, Lemma 21.1(ii)). Thus, g verifies the DP constraint if and only if $F_{\mu_s(\eta)}(\beta_s)$ does not depend on s. Denoting by γ this constant, we find that the optimal fair classifier must be of the form $g(\boldsymbol{x}, s) = \mathbf{1}\left(F_{\mu_s(\eta)}(\eta(\boldsymbol{x}, s)) \geq \gamma\right)$. The risk of any such classifier is given by

$$\mathcal{R}(g) = \mathbb{E}[Y] + \sum_{s\in[K]} p_s \mathbb{E}[\mathbf{1}\left(F_{\mu_s(\eta)}(\eta(\boldsymbol{x}, s)) \geq \gamma\right)(1 - 2\eta(\boldsymbol{X}, s)) \mid S = s] \enspace . \tag{17}$$

Using again inverse transform theorem, Eq. (17) can be further simplified to the following expression:

$$\mathcal{R}(g) = \mathbb{E}[Y] + \sum_{s\in[K]} p_s \int_0^1 \mathbf{1}\left(F_{\mu_s(\eta)} \circ F^{-1}_{\mu_s(\eta)}(u) \geq \gamma\right)(1 - 2F^{-1}_{\mu_s(\eta)}(u))\,\mathrm{d}u \enspace . \tag{18}$$

Under Assumption 3.1, $F_{\mu_s(\eta)} \circ F^{-1}_{\mu_s(\eta)}(u) = u$ for all $u \in (0, 1)$. Thus, Eq. (18) reduces to

$$\mathcal{R}(g) = \mathbb{E}[Y] + \int_\gamma^1 \sum_{s\in[K]} p_s(1 - 2F^{-1}_{\mu_s(\eta)}(u))\,\mathrm{d}u \enspace .$$

This function is minimized at γ^* which satisfies

$$\left(\sum_{s\in[K]} p_s F^{-1}_{\mu_s(\eta)}\right)(\gamma^*) = 1/2 \enspace , \tag{19}$$

and the optimal classifier under the demographic parity constraints is given by $g^*(\boldsymbol{x}, s) = \mathbf{1}\left(F_{\mu_s(\eta)}(\eta(\boldsymbol{x}, s)) \geq \gamma^*\right)$. Taking into account the condition satisfied by γ^*, we conclude. □

Proof of Proposition 3.5. It is a well known fact that $\min_{g\in\mathcal{G}} \mathbb{P}(Y \neq g(\boldsymbol{X}, S)) = \mathbb{E}\min\{\eta(\boldsymbol{X}, S), 1 - \eta(\boldsymbol{X}, S)\} = 1/2 - \mathbb{E}|\eta(\boldsymbol{X}, S) - 1/2|$, where the last equality follows from the fact that $\min\{a, b\} = \frac{1}{2}(a + b - |a - b|)$. Thus, we only need to show that $\mathbb{P}(Y \neq g^*(\boldsymbol{X}, S)) = 1/2 - \mathbb{E}|f^*(\boldsymbol{X}, S) - 1/2|$. For any classifier g, we have

$$\mathbb{P}(Y \neq g(\boldsymbol{X}, S)) = \mathbb{E}[\eta(\boldsymbol{X}, S)(1 - g(\boldsymbol{X}, S))] + \mathbb{E}[(1 - \eta(\boldsymbol{X}, S))g(\boldsymbol{X}, S)] \enspace .$$

We conclude the proof recalling the expression of g^* provided in Theorem 3.3 under Assumption 3.1 and using Lemma C.3. □

Proof of Theorem 4.3. Let us first show that $\theta^*_{(\mathsf{n},\mathsf{d})}$ exists and unique. Indeed, the mapping

$$\theta \mapsto \theta \cdot \left\{\frac{\mathsf{n}_0\mathsf{d}_1 - \mathsf{d}_0\mathsf{n}_1}{\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1}\right\} + \left\{\frac{\mathsf{n}_0\mathsf{d}_2 - \mathsf{d}_0\mathsf{n}_2}{\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1}\right\} - \mathbb{E}\left[(f^*(\boldsymbol{X}, S) - \theta)_+\right] \enspace ,$$

is continuous and monotone increasing on $[0, 1]$ under the specified conditions. On the one hand, for $\theta = 0$ we have $\mathbb{E}[f^*(\boldsymbol{X}, S)] = \mathbb{P}(Y = 1)$ (see Chzhen and Schreuder, 2022, Section 4, item 4 on average stability) the above mapping evaluates to $\left\{\frac{\mathsf{n}_0\mathsf{d}_2 - \mathsf{d}_0\mathsf{n}_2}{\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1}\right\} - \mathbb{P}(Y = 1) \leq 0$. On the other hand, for $\theta = 1$, it evaluates to $\left\{\frac{\mathsf{n}_0\mathsf{d}_1 - \mathsf{d}_0\mathsf{n}_1}{\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1}\right\} + \left\{\frac{\mathsf{n}_0\mathsf{d}_2 - \mathsf{d}_0\mathsf{n}_2}{\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1}\right\} \geq 0$. The existence follows from the intermediate value theorem and the uniqueness from monotonicity. The rest of the proof follows from Lemma 4.5 and Lemma 4.4. □

Proof of Lemma 4.4. For compactness we drop the subscripts (n, d) in this proof. Using Lemma C.3, we find that

$$\begin{aligned}\mathbb{P}\left(g^*(\boldsymbol{X}, S) = 1, Y = 1\right) &= \mathbb{E}\left[f^*(\boldsymbol{X}, S) g^*(\boldsymbol{X}, S)\right] \\ &= \mathbb{E}\left[\left(f^*(\boldsymbol{X}, S) - \theta^*\right)_+\right] + \theta^* \mathbb{E}\left[g^*(\boldsymbol{X}, S)\right] \ .\end{aligned}$$

Case 1: $\mathsf{n}_2\mathsf{d}_1 \neq \mathsf{d}_2\mathsf{n}_1$. Combining this result with (6), we obtain the following expression for $\mathrm{U}(g^*)$:

$$\frac{\mathsf{n}_0(\mathsf{n}_2\mathsf{d}_1 - \cancel{\mathsf{d}_2\mathsf{n}_1}) + \mathsf{n}_1\left(\theta^*(\mathsf{n}_0\mathsf{d}_1 - \mathsf{d}_0\mathsf{n}_1) + (\cancel{\mathsf{n}_0\mathsf{d}_2} - \mathsf{d}_0\mathsf{n}_2)\right) + (\mathsf{n}_2 + \theta^*\mathsf{n}_1)(\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1)\mathbb{E}[g^*(\boldsymbol{X}, S)]}{\mathsf{d}_0(\cancel{\mathsf{n}_2\mathsf{d}_1} - \mathsf{d}_2\mathsf{n}_1) + \mathsf{d}_1\left(\theta^*(\mathsf{n}_0\mathsf{d}_1 - \mathsf{d}_0\mathsf{n}_1) + (\mathsf{n}_0\mathsf{d}_2 - \cancel{\mathsf{d}_0\mathsf{n}_2})\right) + (\mathsf{d}_2 + \theta^*\mathsf{d}_1)(\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1)\mathbb{E}[g^*(\boldsymbol{X}, S)]} \ .$$

Factorizing the numerator and denominator by $(\mathsf{n}_2 + \theta^*\mathsf{n}_1)$ and $(\mathsf{d}_2 + \theta^*\mathsf{d}_1)$ respectively, the above can be written as

$$\mathrm{U}(g^*) = \frac{\mathsf{n}_2 + \theta^*\mathsf{n}_1}{\mathsf{d}_2 + \theta^*\mathsf{d}_1} \cdot \frac{(\mathsf{n}_0\mathsf{d}_1 - \mathsf{d}_0\mathsf{n}_1) + (\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1)\mathbb{E}[g^*(\boldsymbol{X}, S)]}{(\mathsf{n}_0\mathsf{d}_1 - \mathsf{d}_0\mathsf{n}_1) + (\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1)\mathbb{E}[g^*(\boldsymbol{X}, S)]} = \frac{\mathsf{n}_2 + \theta^*\mathsf{n}_1}{\mathsf{d}_2 + \theta^*\mathsf{d}_1} \ ,$$

concluding the proof for the first case.
Case 2: $\mathsf{n}_2\mathsf{d}_1 = \mathsf{d}_2\mathsf{n}_1$. In this case, notice that we have

$$\mathsf{n}_1\theta^* = \frac{\mathsf{n}_1\mathsf{n}_2\mathsf{d}_0 - \mathsf{n}_0\mathsf{n}_1\mathsf{d}_2}{\mathsf{n}_0\mathsf{d}_1 - \mathsf{n}_1\mathsf{d}_0} = \mathsf{n}_2\frac{\mathsf{n}_1\mathsf{d}_0 - \mathsf{n}_0\mathsf{d}_2}{\mathsf{n}_0\mathsf{d}_1 - \mathsf{n}_1\mathsf{d}_0} = -\mathsf{n}_2 \ ,$$

and, following the same computations, $\mathsf{d}_1\theta^* = -\mathsf{d}_2$. Plugging the above equalities in the definition of $\mathrm{U}(g^*)$ yields

$$\mathrm{U}(g^*) = \frac{\mathsf{n}_0 + \mathsf{n}_1\mathbb{E}(f^*(\boldsymbol{X}, S) - \theta^*)_+}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}\left(f^*(\boldsymbol{X}, S) - \theta^*\right)_+} \ .$$

The proof is concluded. □

Proof of Lemma 4.5. Let $\bar{\mu}(\eta)$ be the Wasserstein barycenter of measures $\mu_1(\eta), \ldots, \mu_K(\eta)$, weighted by $p_1, \ldots, p_K$ respectively. Assumption 3.1 and the form of f^* ensures that the fair optimal classifier in Eq. (5) can be expressed as

$$g^*(\boldsymbol{x}, s) = \mathbf{1}\left(\eta(\boldsymbol{x}, s) \geq F^{-1}_{\mu_s(\eta)} \circ F_{\bar{\mu}(\eta)}(\theta^*)\right) = \mathbf{1}\left(\eta(\boldsymbol{x}, s) \geq F^{-1}_{\mu_s(\eta)}(\beta^*)\right) \ ,$$

where $\beta^* = F_{\bar{\mu}(\eta)}(\theta^*)$. Fix an arbitrary classifier g which satisfies the demographic parity constraint.
Our goal is to develop $\mathrm{U}(g^*) - \mathrm{U}(g)$, which we express as a sum of two terms $\mathsf{I} + \mathsf{II}$, with

$$\mathsf{I} \triangleq \frac{\mathsf{n}_1\left(\mathbb{E}[\eta(\boldsymbol{X}, S)(g^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S))]\right) + \mathsf{n}_2\mathbb{E}[g^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S)]}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)]} \ ,$$

and

$$\mathsf{II} \triangleq -\mathrm{U}(g)\frac{\mathsf{d}_1\left(\mathbb{E}[\eta(\boldsymbol{X}, S)(g^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S))]\right) + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S)]}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)]} \ .$$

One verifies that indeed $\mathrm{U}(g^*) - \mathrm{U}(g) = \mathsf{I} + \mathsf{II}$. Thanks to the alternative definition of g^* introduced in the beginning of this proof, for any $a, b \in \mathbb{R}$ we have

$$\begin{aligned}&a\mathbb{E}[\eta(\boldsymbol{X}, S)(g^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S))] + b\mathbb{E}[g^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S)] \\ &\qquad= a\mathbb{E}\left[\left|\eta(\boldsymbol{X}, S) - F^{-1}_{\mu_S(\eta)}(\beta^*)\right|\mathbf{1}\left(g^*(\boldsymbol{X}, S) \neq g(\boldsymbol{X}, S)\right)\right] \\ &\qquad\quad+ \mathbb{E}[(b + aF^{-1}_{\mu_S(\eta)}(\beta^*))(g^*(\boldsymbol{X}, S)) - g(\boldsymbol{X}, S)] \\ &\qquad= a\mathbb{E}\left[\left|\eta(\boldsymbol{X}, S) - F^{-1}_{\mu_S(\eta)}(\beta^*)\right|\mathbf{1}\left(g^*(\boldsymbol{X}, S) \neq g(\boldsymbol{X}, S)\right)\right] \\ &\qquad\quad+ (b + aF^{-1}_{\bar{\mu}(\eta)}(\beta^*))\mathbb{E}[g^*(\boldsymbol{X}, S)) - g(\boldsymbol{X}, S)] \ ,\end{aligned}$$

where the last equality is due to the fact that g satisfies the demographic parity constraint. Thus, setting $\Delta(g^*, g) \triangleq \mathbb{E}\left[|\eta(\boldsymbol{X}, S) - F^{-1}_{\mu_S(\eta)}(\beta^*)|\mathbf{1}\left(g^*(\boldsymbol{X}, S) \neq g(\boldsymbol{X}, S)\right)\right]$ and recalling that $\theta^* = F^{-1}_{\bar{\mu}(\eta)}(\beta^*)$ we can express I and II as

$$\begin{aligned}\mathsf{I} &= \frac{\mathsf{n}_1\Delta(g^*, g) + (\mathsf{n}_2 + \mathsf{n}_1\theta^*)\mathbb{E}[g^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S)]}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)]} \ , \\ \mathsf{II} &= -\mathrm{U}(g)\frac{\mathsf{d}_1\Delta(g^*, g) + (\mathsf{d}_2 + \mathsf{d}_1\theta^*)\mathbb{E}[g^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S)]}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)]} \ .\end{aligned}$$

Case 1: $\mathsf{n}_2\mathsf{d}_1 \neq \mathsf{d}_2\mathsf{n}_1$. Lemma 4.4 implies that

$$\mathsf{I} = \frac{\mathsf{n}_1\Delta(g^*, g) + \mathrm{U}(g^*)(\mathsf{d}_2 + \mathsf{d}_1\theta^*)\mathbb{E}[g^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S)]}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)]} \enspace .$$

Combining the above two expressions for I and II we obtain

$$\begin{aligned}\mathrm{U}(g^*) - \mathrm{U}(g) = (\mathrm{U}(g^*) - \mathrm{U}(g))\, &\frac{(\mathsf{d}_2 + \mathsf{d}_1\theta^*)\mathbb{E}[g^*(\boldsymbol{X}, S) - g(\boldsymbol{X}, S)]}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)]}\\ + (\mathsf{n}_1 - \mathrm{U}(g)\mathsf{d}_1)\, &\frac{\Delta(g^*, g)}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)]} \enspace .\end{aligned}$$

Simplifying the above and using Lemma C.3, we obtain

$$\mathrm{U}(g^*) - \mathrm{U}(g) = (\mathsf{n}_1 - \mathrm{U}(g)\mathsf{d}_1)\, \frac{\Delta(g^*, g)}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}\left[(f^*(\boldsymbol{X}, S) - \theta^*)_+\right] + (\mathsf{d}_2 + \theta^*\mathsf{d}_1)\mathbb{E}[g(\boldsymbol{X}, S)]} \enspace .$$

As in Lemma 4.4 (using the expression for the numerator), we deduce that

$$\begin{aligned}&\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}\left[(f^*(\boldsymbol{X}, S) - \theta^*)_+\right] + (\mathsf{d}_2 + \theta^*\mathsf{d}_1)\mathbb{E}[g(\boldsymbol{X}, S)]\\ &\qquad = \frac{(\mathsf{d}_2 + \theta^*\mathsf{d}_1)\left((\mathsf{n}_0\mathsf{d}_1 - \mathsf{d}_0\mathsf{n}_1) + (\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1)\mathbb{E}[g(\boldsymbol{X}, S)]\right)}{\mathsf{n}_2\mathsf{d}_1 - \mathsf{d}_2\mathsf{n}_1} \enspace ,\end{aligned}$$

and using the definition of $\mathrm{U}(g)$, we can write

$$\mathsf{n}_1 - \mathrm{U}(g)\mathsf{d}_1 = \frac{(\mathsf{n}_1\mathsf{d}_0 - \mathsf{d}_1\mathsf{n}_0) + (\mathsf{n}_1\mathsf{d}_2 - \mathsf{d}_1\mathsf{n}_2)\mathbb{E}[g(\boldsymbol{X}, S)]}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g(\boldsymbol{X}, S)]} \enspace . \tag{20}$$

Combining the last three displays, we arrive at the claimed equality

$$\mathrm{U}(g^*) - \mathrm{U}(g) = \frac{\mathsf{d}_2\mathsf{n}_1 - \mathsf{n}_2\mathsf{d}_1}{\mathsf{d}_2 + \theta^*\mathsf{d}_1} \cdot \frac{\mathbb{E}|\eta(\boldsymbol{X}, S) - F^{-1}_{\mu_S(\eta)}(\beta^*)|\mathbf{1}\left(g^*(\boldsymbol{X}, S) \neq g(\boldsymbol{X}, S)\right)}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g(\boldsymbol{X}, S)]} \enspace .$$

Case 2: $\mathsf{n}_2\mathsf{d}_1 = \mathsf{d}_2\mathsf{n}_1$. We have shown in the proof of Lemma 4.4 that in this particular case, $\mathsf{n}_1\theta^* + \mathsf{n}_2 = \mathsf{d}_1\theta^* + \mathsf{d}_2 = 0$. Hence I and II reduce to

$$\begin{aligned}\mathsf{I} &= \frac{\mathsf{n}_1\Delta(g^*, g)}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)]} \enspace ,\\ \mathsf{II} &= -\mathrm{U}(g)\frac{\mathsf{d}_1\Delta(g^*, g)}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)]} \enspace .\end{aligned}$$

Consequently, the difference of utilities is expressed as

$$\mathrm{U}(g^*) - \mathrm{U}(g) = (\mathsf{n}_1 - \mathrm{U}(g)\mathsf{d}_1)\, \frac{\Delta(g^*, g)}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)]} \enspace .$$

Again invoking the result of Lemma C.3, we deduce

$$\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g^*(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g^*(\boldsymbol{X}, S)] = \mathsf{d}_0 + \mathsf{d}_1\mathbb{E}\left[(f^*(\boldsymbol{X}, S) - \theta^*)_+\right] \enspace .$$

The above two displays combined with Eq. (20) and the condition $\mathsf{n}_2\mathsf{d}_1 = \mathsf{d}_2\mathsf{n}_1$ yield

$$\mathrm{U}(g^*) - \mathrm{U}(g) = \frac{\mathsf{n}_1\mathsf{d}_0 - \mathsf{d}_1\mathsf{n}_0}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}\left[(f^*(\boldsymbol{X}, S) - \theta^*)_+\right]} \cdot \frac{\Delta(g^*, g)}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}[\eta(\boldsymbol{X}, S)g(\boldsymbol{X}, S)] + \mathsf{d}_2\mathbb{E}[g(\boldsymbol{X}, S)]} \enspace .$$

The proof is concluded. □

Proof of Proposition 5.3. For any $g : \mathcal{X} \times \{1, 2\} \to \{0, 1\}$, define $\tilde{g} : \mathcal{X} \to \{0, 1\}$ as

$$\tilde{g}(\boldsymbol{x}) = \begin{cases} g(\boldsymbol{x}, 1) & \boldsymbol{x} \in \mathrm{supp}(\mathbb{P}^\star_{\boldsymbol{X}|S=1})\\ g(\boldsymbol{x}, 2) & \boldsymbol{x} \in \mathrm{supp}(\mathbb{P}^\star_{\boldsymbol{X}|S=2})\\ \mathbf{1}\left(\eta(\boldsymbol{x}) \geq 1/2\right) & \boldsymbol{x} \in \mathrm{supp}(\mathbb{P}_{\boldsymbol{X}}) \setminus \Big(\mathrm{supp}(\mathbb{P}^\star_{\boldsymbol{X}|S=1}) \cup \mathrm{supp}(\mathbb{P}^\star_{\boldsymbol{X}|S=2})\Big) \end{cases} \enspace .$$

Note that the above correspondence of g and $\tilde{g}$ is invertible since the supports of $\mathbb{P}^{\star}_{\boldsymbol{X}|S=1}$ and $\mathbb{P}^{\star}_{\boldsymbol{X}|S=2}$ do not intersect by construction. Observe that for any $g : \mathcal{X} \times \{1, 2\} \to \{0, 1\}$ it holds that

$$g(\boldsymbol{X}, S) \perp\!\!\!\perp_{\mathbb{P}^{\star}} S \iff g(\cdot, 1)\sharp\mathbb{P}^{\star}_{\boldsymbol{X}|S=1} = g(\cdot, 2)\sharp\mathbb{P}^{\star}_{\boldsymbol{X}|S=2} \iff \tilde{g}\sharp\mathbb{P}_{\boldsymbol{X}|S=1} = \tilde{g}\sharp\mathbb{P}_{\boldsymbol{X}|S=2} \enspace .$$

Thus, given any classifier g satisfying the demographic parity constraint under $\mathbb{P}^{\star}$, we can transform it to a classifier that satisfies the constraints under $\mathbb{P}$. Furthermore, since

$$\mathbb{E}_{\mathbb{P}^{\star}}[g(\boldsymbol{X}, S)(1 - 2\tilde{\eta}(\boldsymbol{X}, S))] = \mathbb{E}_{\mathbb{P}}[\tilde{g}(\boldsymbol{X})(1 - 2Y)\mathbf{1}\left(\boldsymbol{X} \in \mathrm{supp}(\mathbb{P}^{\star}_{\boldsymbol{X}|S=1}) \cap \mathrm{supp}(\mathbb{P}^{\star}_{\boldsymbol{X}|S=2})\right)] \enspace ,$$

taking any classifier $\bar{g} : \mathcal{X} \to \{0, 1\}$ we can write

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[\bar{g}(\boldsymbol{X})(1 - 2Y)] = \mathbb{E}_{\mathbb{P}^{\star}}[\bar{g}(\boldsymbol{X}, S)(1 - 2\tilde{\eta}(\boldsymbol{X}, S))\mathbf{1}\left(\boldsymbol{X} \in \mathrm{supp}(\mathbb{P}^{\star}_{\boldsymbol{X}|S=1}) \cap \mathrm{supp}(\mathbb{P}^{\star}_{\boldsymbol{X}|S=2})\right)]\\ + \mathbb{E}_{\mathbb{P}}[\bar{g}(\boldsymbol{X})(1 - 2Y)\mathbf{1}\left(\boldsymbol{X} \notin \mathrm{supp}(\mathbb{P}^{\star}_{\boldsymbol{X}|S=1}) \cap \mathrm{supp}(\mathbb{P}^{\star}_{\boldsymbol{X}|S=2})\right)] \enspace ,\end{aligned}$$

where in the first equality, we added the input S to $\bar{g}$ sue to the fact that S is $\boldsymbol{X}$ measurable under $\mathbb{P}^{\star}$. Note that the second term is minimized point-wise by the Bayes classifier, while the first term is minimized by $g^{*}_{\mathbb{P}^{\star}}$ thanks to the equivalence established for the demographic parity constraint. □

C AUXILIARY RESULTS

The first lemma ensures that under certain conditions, the denominator of the linear fractional performance measure is always positive.

Lemma C.1. *Assume that* $\mathsf{d}_0 + \min\big\{\min\{\mathsf{d}_1, 0\} + \mathsf{d}_2, 0\big\} \geq 0$*, then for any classifier* $g : \mathcal{X} \times [K] \to \{0, 1\}$

$$\mathsf{d}_0 + \mathsf{d}_1\mathbb{P}(Y = 1, g(\boldsymbol{X}, S) = 1) + \mathsf{d}_2\mathbb{P}(g(\boldsymbol{X}, S) = 1) \geq 0 \enspace .$$

Furthermore, if $\mathsf{d}_0 + \min\big\{\min\{\mathsf{d}_1, 0\} + \mathsf{d}_2, 0\big\} > 0$*, then the above inequality is strict.*

Proof. Observe that

$$\begin{aligned}\mathsf{d}_0 + \mathsf{d}_1\mathbb{P}(Y = 1, g(\boldsymbol{X}, S) = 1) + \mathsf{d}_2\mathbb{P}(g(\boldsymbol{X}, S) = 1) &= \mathsf{d}_0 + \mathbb{E}[(\mathsf{d}_1 Y + \mathsf{d}_2)g(\boldsymbol{X}, S)]\\ &\geq \mathsf{d}_0 + \mathbb{E}[(\min\{\mathsf{d}_1, 0\} + \mathsf{d}_2)g(\boldsymbol{X}, S)]\\ &\geq \mathsf{d}_0 + \min\big\{\min\{\mathsf{d}_1, 0\} + \mathsf{d}_2, 0\big\}\\ &\geq 0 \enspace .\end{aligned}$$

The second claim follows the same lines. □

The second result gives a sufficient condition for positivity of the leading coefficient in Remark 4.6.

Lemma C.2. *Assume that* $\mathsf{d}_0 + \min\big\{\min\{\mathsf{d}_1, 0\} + \mathsf{d}_2, 0\big\} \geq 0$ *and either Eq.* (C1) *or Eq.* (C2) *is satisfied, then for any classifier* $g \in \mathrm{dom}(\mathrm{U}_{(\mathsf{n},\mathsf{d})})$

$$\mathsf{n}_1 - \mathsf{d}_1\mathrm{U}_{(\mathsf{n},\mathsf{d})}(g) \geq 0 \enspace .$$

Proof. Observe that in both cases, by Lemma C.1, we have

$$\mathrm{sign}(\mathsf{n}_1 - \mathsf{d}_1\mathrm{U}_{(\mathsf{n},\mathsf{d})}(g)) = \mathrm{sign}\Big((\mathsf{n}_1\mathsf{d}_0 - \mathsf{d}_1\mathsf{n}_0) + (\mathsf{n}_1\mathsf{d}_2 - \mathsf{d}_1\mathsf{n}_2)\mathbb{E}[g(\boldsymbol{X}, S)]\Big) \enspace . \tag{21}$$

Case 1: $\mathsf{n}_2\mathsf{d}_1 \neq \mathsf{d}_2\mathsf{n}_1$. In that case condition (C1) implies that $\mathsf{n}_1\mathsf{d}_2 - \mathsf{d}_1\mathsf{n}_2 > 0$ and $\frac{\mathsf{n}_1\mathsf{d}_0 - \mathsf{d}_1\mathsf{n}_0}{\mathsf{n}_1\mathsf{d}_2 - \mathsf{d}_1\mathsf{n}_2} \geq 0$. In view of (21) we conclude.
Case 2: $\mathsf{n}_2\mathsf{d}_1 = \mathsf{d}_2\mathsf{n}_1$. The proof is immediate from (21) and the first part of condition (C2). □

The next lemma establishes an extended average stability property from (Chzhen and Schreuder, 2022).

Lemma C.3. *Let Assumption 3.1 be satisfied, then*

$$\mathbb{E}[(f^*(\boldsymbol{X}, S) - \eta(\boldsymbol{X}, S))\mathbf{1}\left(f^*(\boldsymbol{X}, S) \geq \theta\right)] = 0 \enspace ,$$

for all $\theta \in [0, 1]$.

Proof. Fix some $\theta \in [0, 1]$. Introducing $T^*(\cdot) \triangleq \Big(\sum_{\sigma=1}^K p_\sigma F^{-1}_{\mu_\sigma(\eta)}\Big)(\cdot)$, we recall that

$$f^*(\boldsymbol{x}, s) = T^* \circ F_{\mu_s(\eta)}(\eta(\boldsymbol{x}, s)) \enspace .$$

Furthermore, since both $F_{\mu_S(\eta)}(\eta(\boldsymbol{X}, S))$ and $(F_{\mu_S(\eta)}(\eta(\boldsymbol{X}, S)) \mid S = s)$ are distributed uniformly on $(0, 1)$ under Assumption 3.1, we can write

$$\begin{aligned}&\mathbb{E}[(f^*(\boldsymbol{X}, S) - \eta(\boldsymbol{X}, S))g^*(\boldsymbol{X}, S)]\\ &\quad= \mathbb{E}[T^*(U)\mathbf{1}\left(T^*(U) \geq \theta\right)] - \sum_{s=1}^K p_s\mathbb{E}[F^{-1}_{\mu_s(\eta)}(U)\mathbf{1}\left(T^*(U) \geq \theta\right) \mid S = s] = 0 \enspace .\end{aligned}$$

□

Finally, the last result relates the excess risk obtained in Lemma 4.5 with the expression presented in Remark 4.6.

Lemma C.4. *Under the conditions of Lemma 4.4, we have*

$$\mathsf{n}_1 - \mathsf{d}_1 \mathrm{U}(g^*) = \begin{cases} \dfrac{\mathsf{d}_2\mathsf{n}_1 - \mathsf{n}_2\mathsf{d}_1}{\mathsf{d}_2 + \theta^*_{(\mathsf{n},\mathsf{d})}\mathsf{d}_1} & \textit{if } \mathsf{d}_2\mathsf{n}_1 \neq \mathsf{n}_2\mathsf{d}_1 \\ \dfrac{\mathsf{n}_1\mathsf{d}_0 - \mathsf{d}_1\mathsf{n}_0}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}(f^*(\boldsymbol{X}, S) - \theta^*_{(\mathsf{n},\mathsf{d})})_+} & \textit{if } \mathsf{d}_2\mathsf{n}_1 = \mathsf{n}_2\mathsf{d}_1 \end{cases} .$$

Proof. We drop the subscript (n, d) for compactness.
Case 1: $\mathsf{d}_2\mathsf{n}_1 \neq \mathsf{n}_2\mathsf{d}_1$. Using the corresponding case of Lemma 4.4 and solving it for θ^*, we deduce that

$$\theta^* = \frac{\mathsf{n}_2 - \mathsf{d}_2\mathrm{U}(g^*)}{\mathsf{d}_1\mathrm{U}(g^*) - \mathsf{n}_1} .$$

Hence, from the above we deduce that

$$\mathsf{d}_2 + \theta^*\mathsf{d}_1 = \frac{\mathsf{d}_1\mathsf{n}_2 - \mathsf{d}_2\mathsf{n}_1}{\mathsf{d}_1\mathrm{U}(g^*) - \mathsf{n}_1} \qquad \Longrightarrow \qquad \frac{\mathsf{d}_2\mathsf{n}_1 - \mathsf{n}_2\mathsf{d}_1}{\mathsf{d}_2 + \theta^*\mathsf{d}_1} = \mathsf{n}_1 - \mathsf{d}_1\mathrm{U}(g^*) .$$

Case 1: $\mathsf{d}_2\mathsf{n}_1 = \mathsf{n}_2\mathsf{d}_1$. Again using the corresponding case of Lemma 4.4 and solving it for $\mathbb{E}(f^*(\boldsymbol{X}, S) - \theta^*)_+$, we deduce that

$$\mathbb{E}(f^*(\boldsymbol{X}, S) - \theta^*_{(\mathsf{n},\mathsf{d})})_+ = \frac{\mathsf{d}_0\mathrm{U}(g^*) - \mathsf{n}_0}{\mathsf{n}_1 - \mathsf{d}_1\mathrm{U}(g^*)} .$$

Hence, from the above we deduce that

$$\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}(f^*(\boldsymbol{X}, S) - \theta^*)_+ = \frac{\mathsf{d}_0\mathsf{n}_1 - \mathsf{d}_1\mathsf{n}_0}{\mathsf{n}_1 - \mathsf{d}_1\mathrm{U}(g^*)} \Longrightarrow \frac{\mathsf{n}_1\mathsf{d}_0 - \mathsf{d}_1\mathsf{n}_0}{\mathsf{d}_0 + \mathsf{d}_1\mathbb{E}(f^*(\boldsymbol{X}, S) - \theta^*)_+} = \mathsf{n}_1 - \mathsf{d}_1\mathrm{U}(g^*) .$$

The proof is concluded. □

D ADDITIONAL PLOTS

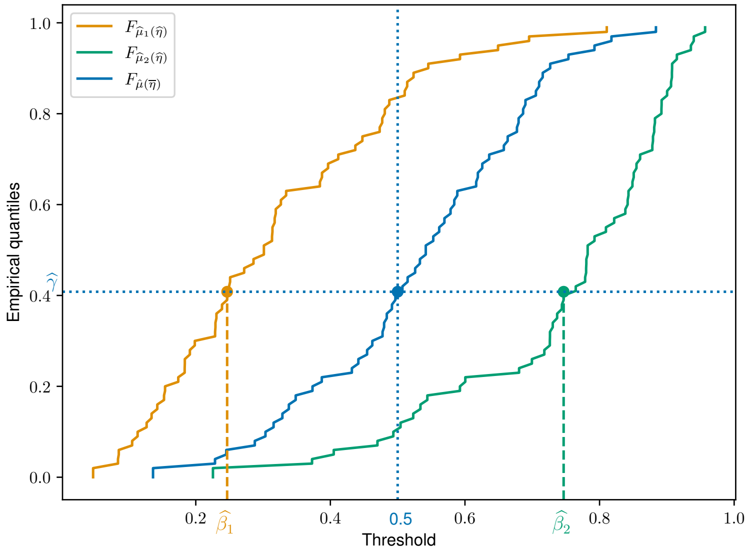

Figure 3: Empirical quantile functions of $\hat{\eta}(\cdot, 1)$, $\hat{\eta}(\cdot, 2)$ and of their Wasserstein-2 barycenter (orange, green, and blue curves, respectively).

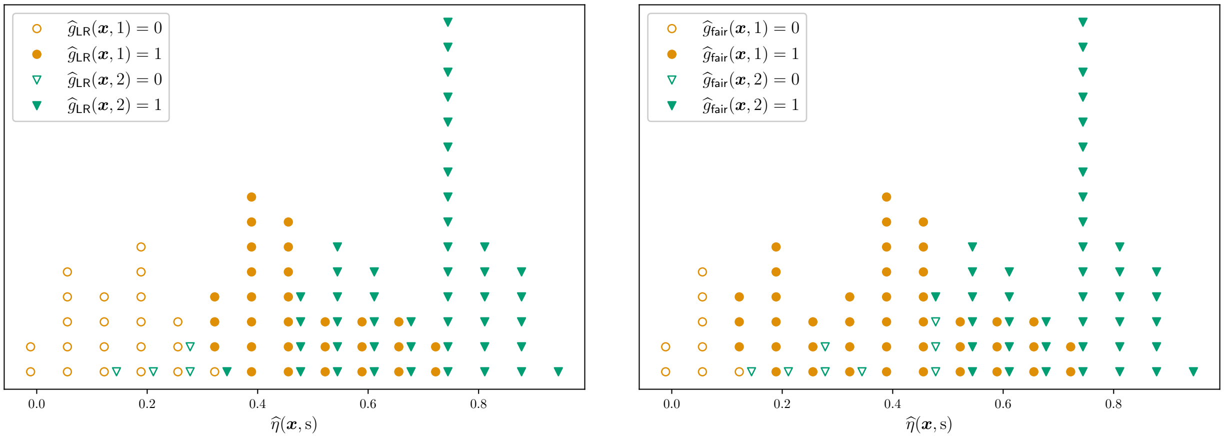

Figure 4: Empirical group-wise outcomes for logistic regression (`left`) and the proposed post-processed fair plug-in classifier (`right`) for the Jaccard risk measure. The experimental setting is the same as that of Figure 4. The optimal threshold given in Theorem 4.3 was estimated using Brent's method as implemented in `scipy` (Virtanen et al., 2020). For the fair plug-in classifier, we obtained an estimator $\hat{f}$ of the optimal fair regression function f^* following Chzhen et al. (2020b) and an estimator of the optimal threshold $\hat{\theta}$ given in Theorem 4.3 using Brent's method as implemented in `scipy` (Virtanen et al., 2020). We then considered $\hat{g}_{\mathsf{fair}}(\boldsymbol{x},s) \triangleq \mathbf{1}(\hat{f}(\boldsymbol{x},s) \geq \hat{\theta})$.

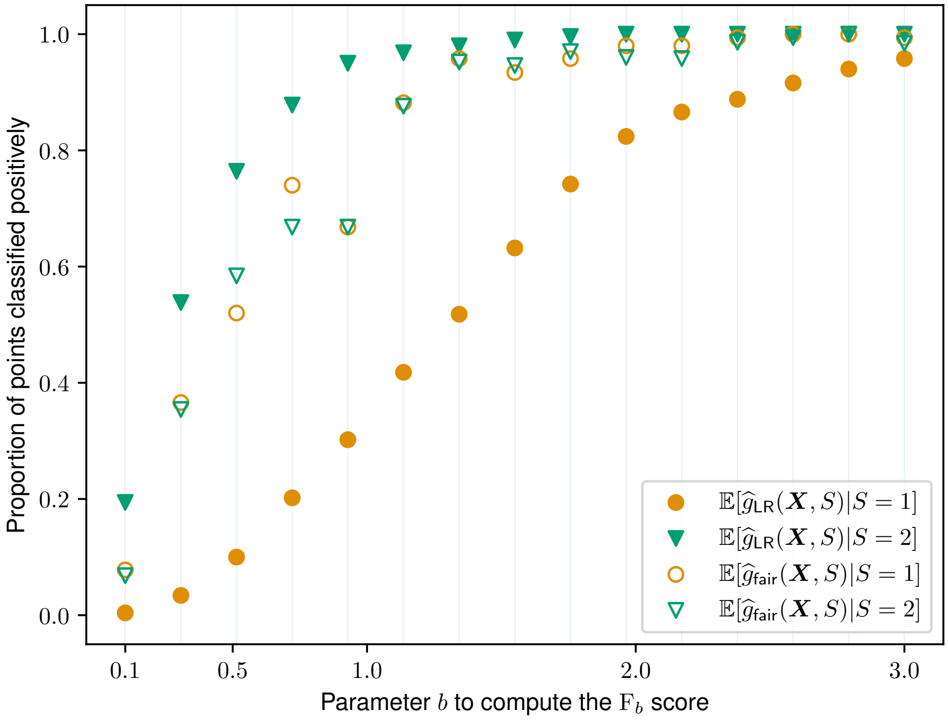

Figure 5: Proportion of positively classified observations as the F_b score precision-recall trade-off parameter b varies (see Table 1 for the definition of F_b score). The experimental setting is the same as that of Figure 4. The optimal threshold given in Theorem 4.3 was estimated using Brent's method as implemented in `scipy` (Virtanen et al., 2020).