A Tale of Two Measures: Fair Classification at Any Decision Threshold

Kweku Kwegyir-Aggrey,[*†] **Jessica Dai,**[*‡] **A. Feder Cooper,**[✠]
John P. Dickerson,[*] **Keegan Hines**[*]

[*]Arthur [†]Brown University [‡]University of California, Berkeley [✠]Cornell University

Abstract

We study the problem of post-processing a supervised machine learning regressor to maximize fair classification at all decision thresholds. Specifically, we show that by decreasing the statistical distance between each group's score distributions, we can increase fair performance across all thresholds at once, and that we can do so without a significant decrease in accuracy. To this end, we introduce a formal measure of *distributional parity*, which captures the degree of similarity in the distributions of classifications for different protected groups. In contrast to prior work, which formalizes a measure of Strong Demographic Parity by examining positive rates, our measure applies to a large class of fairness metrics. Our main result is to put forward a novel post-processing algorithm based on optimal transport, which provably maximizes distributional parity. We support this result with experiments on several fairness benchmarks.

1 Introduction

There is a growing gap between algorithmic fairness research and the empirical realities of deploying fair models in practice. In fair ML scholarship, a common paradigm involves training a classifier with a chosen decision threshold to attain a certain degree of accuracy, and then post-processing the classifier to correct for unfairness, according to a chosen fairness definition (Calders, Kamiran, and Pechenizkiy 2009; Hardt, Price, and Srebro 2016; Pleiss et al. 2017). Despite the preeminence of this approach, it is well-known that the specific choice of decision threshold can influence both fairness and accuracy in practice (Barocas, Hardt, and Narayanan 2019) (Figure 1). When deploying a classifier, practitioners typically need to tinker with the threshold and evaluate if the resulting model meets their domain-specific needs. Moreover, even once a threshold is selected, needs may change over time, requiring changes in the threshold (Kallus and Zhou 2019; Chouldechova 2016).

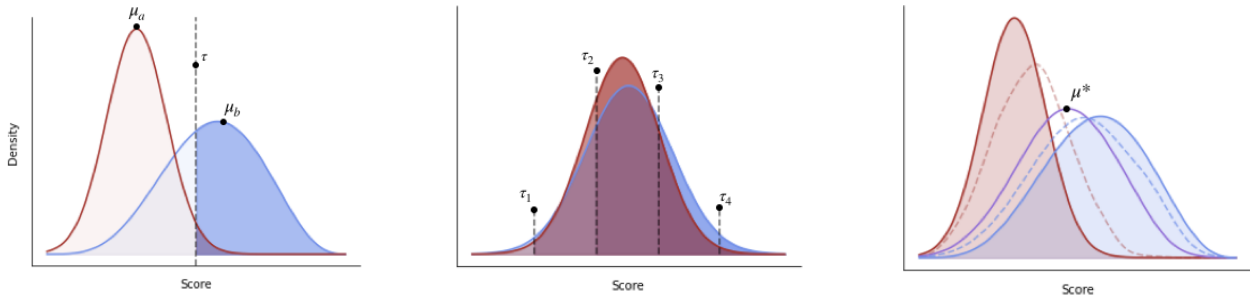
This presents two problems. First, it may be prohibitively expensive to retrain a model every time practitioners want to test a different threshold (Alla and Adari 2021; Shankar et al. 2022). Second, practitioners may not even have access to the training pipeline. It is now common for well-resourced companies and institutions to train and publicly release "general-purpose" models trained on proprietary datasets (Cooper et al. 2022; Kroll 2021); practitioners who use these models cannot retrain them, even if they had the resources to do so.

Given these tensions, an important research question is to see if it is possible to design tools that cohere better with ML practice. One natural strategy is to develop a procedure that produces regressors that guarantee classification fairness at all possible thresholds, while simultaneously not having too-large an effect on accuracy. If a regressor is fair at all thresholds, then a practitioner can do application-specific threshold tuning without ever needing to retrain.

Prior work has researched such all-threshold fairness guarantees for Demographic Parity (Jiang et al. 2020; Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020; Gordaliza et al. 2019), using optimal transport to transform group-conditional score distributions in order to equalize positive rates (PR) at all decision thresholds (Figure 1c). While a promising contribution in the study of all-threshold fairness, this prior work suffers from the limitation that it only works for Strong Demographic Parity (i.e., all-threshold Demographic Parity, SDP). As has been well-known since the landmark paper by Hardt, Price, and Srebro (2016), and subsequent impossibility results (Kleinberg 2018; Chouldechova 2016), looking only at Demographic Parity does not capture the nuances in unfairness made clear from looking at true positive rates, false positive rates, and combinations thereof.

The main contribution of our paper is to close this gap. Similar to how Hardt, Price, and Srebro (2016) expanded fairness metrics beyond Demographic Parity, we expand the set of all-threshold fairness interventions beyond Strong Demographic Parity (Jiang et al. 2020); whereas Hardt, Price, and Srebro (2016) introduced methods for satisfying Equalized Opportunity/Odds, we introduce and optimize *Distributional Parity* (Definition 3.1), an all-threshold fairness framework that applies to multiple, common fairness metrics. In summary, we:

- Introduce *distributional parity* to reason about the similarity of group-conditional score distributions with respect to a broad class of fairness metrics (Section 3).
- Prove that distributional parity can be maximized. Our key insight is to use *geometric repair* (Feldman et al. 2015), for which we prove properties that enable us to extend beyond SDP (Section 4).



(a) (b) (c)

Figure 1: (a) applies a threshold τ to groups with score distributions that differ, exhibiting classification disparity. (b) shows how similar such score distributions exhibit little-to-no decision disparity at *any* τ. (c) visualizes a technique for interpolating between group-conditional score distributions to find an intermediate distribution that achieves parity at all τ.

- Provide an efficient post-processing algorithm that follows directly from our theoretical results and avoids the need to re-train after changing the decision threshold (Section 5).
- Demonstrate empirically that our method both encompasses the prior work on SDP and outperforms related methods in practice (Section 6).

2 Fairness Preliminaries

Let $X \subseteq \mathbb{R}^d$ be some feature space and $G = \{a, b\}$ to be a set of binary protected attributes, for which a is the majority group and b is the minority group. We define the label space to be the set $Y = \{0, 1\}$, where 1 denotes the positive class and 0 the negative class. We also assume elements in X, G, and Y are are drawn from some underlying distribution, which has corresponding random variables $\boldsymbol{X}$, $\boldsymbol{G}$, and $\boldsymbol{Y}$. To model predictions made over covariates $x \in X$ and $g \in G$, we use a *regressor* $f : X \times G \to \Omega$, where $\Omega \subseteq [0, 1]$ is a closed subset denoting probabilities. That is, f estimates the likelihood that some (x, g) yields the positive outcome, i.e., $f(x, g) = \Pr(\boldsymbol{Y} = 1 | \boldsymbol{X} = x, \boldsymbol{G} = g)$. We refer to the probabilities output by $f(x, g) = s$ as *scores*, given that they "rank" the covariates (x, g) based on their predicted likelihood of attaining the $\boldsymbol{Y} = 1$ outcome. Predictions are computed by applying a threshold $\tau \in [0, 1]$ to scores s. A *classifier* is the combination of a regressor f and a threshold τ; classifications are made by applying the decision rule $\mathbb{1}_{s \geq \tau}$. Additionally, we refer to the distributions of scores produced by a regressor for each group as *group-conditional score distributions*. For a group g, the group-conditional score distribution has random variable $\boldsymbol{S}_g$, where $\boldsymbol{S}_g \sim f(\boldsymbol{X}, \boldsymbol{G}) | \boldsymbol{G} = g$. Our main results require describing the probability measures associated with these distributions. We define $\mathcal{P}_1(\Omega)$ as the set of probability measures on Ω with finite first-order moments. $\mu_g \in \mathcal{P}_1(\Omega)$ is the probability measure associated with $\boldsymbol{S}_g$, where $\mu_g = Law(\boldsymbol{S}_g)$, to which we apply the following standard assumption:

Assumption 2.1. *Any measure with finite first-order moments $\mu \in \mathcal{P}_1(\Omega)$ is non-atomic and absolutely continuous with respect to the Lebesgue measure.*

Fairness metrics. We measure fairness using familiar metrics: Positive Rate (PR), True Positive Rate (TPR), and False Positive Rate (FPR), from which popular fairness definitions, such as Demographic Parity (PR Parity) (Calders, Kamiran, and Pechenizkiy 2009), Equal Opportunity (TPR Parity), and Equalized Odds (TPR and FPR Parity) (Hardt, Price, and Srebro 2016) are computed. We formally define these rates for each group by applying a threshold to group-conditional score distributions, as shown in Table 1.

Table 1: Fairness metrics. $\boldsymbol{S}$ is the score distribution produced by a regressor f; $\tau \in [0, 1]$ is a decision threshold.

Metric	Formula
$\text{PR}_g^f(\tau)$	$\Pr[\boldsymbol{S} \geq \tau \mid \boldsymbol{G} = g]$
$\text{TPR}_g^f(\tau)$	$\Pr[\boldsymbol{S} \geq \tau \mid \boldsymbol{Y} = 1, \boldsymbol{G} = g]$
$\text{FPR}_g^f(\tau)$	$\Pr[\boldsymbol{S} \geq \tau \mid \boldsymbol{Y} = 0, \boldsymbol{G} = g]$

Prior work on fair classification. Much prior research either elides the choice of the decision threshold τ by focusing on already-thresholded decisions (Hardt, Price, and Srebro 2016), or aims to satisfy one of the above fairness definitions for a single, fixed threshold (Zafar et al. 2017). Nevertheless, different choices of τ can have direct effects on fairness and accuracy. Determining the right τ to achieve a desired level of performance often requires empirical investigation, and, of course, application performance requirements may change over time, necessitating changes in the threshold (Barocas, Hardt, and Narayanan 2019; Forde et al. 2021; Kallus and Zhou 2019). In tension with these empirical realities, practitioners may not have the resources to repeatedly re-train or may lack the ability to do so. For example, it is an increasingly common paradigm for model training and model deployment to involve separate sets of actors: One set of actors uses proprietary data and algorithms to train and release "general-purpose" models; practitioners then use these models in potentially-diverse deployment contexts (Cooper et al. 2022; Kroll 2021).

A potential solution to this problem is to produce a regressor f such that it is fair for *all* τ. This way, practitioners can change τ according to their specific fairness and accuracy needs, and they can do so without retraining f. Prior work has attempted to do this — providing a method for achieving Demographic Parity for all τ while preserving classifier

accuracy. Specifically, Jiang et al. (2020) use optimal transport and the Wasserstein-1 distance to show that making group-conditional score distributions equal is tantamount to achieving Demographic Parity across all decision thresholds. Chzhen et al. (2020); Le Gouic, Loubes, and Rigollet (2020) achieve similar results using the Wasserstein-2 distance. Unfortunately, this prior work does not extend to the other fairness metrics discussed above; it does not facilitate enforcing parity of group-conditional score distributions for $\text{TPR}_g(\tau)$ or $\text{FPR}_g(\tau)$. This is a major limitation, as prior algorithmic fairness scholarship has repeatedly illustrated that these rates are able to capture more-nuanced disparities in predictive outcomes across subgroups (Hardt, Price, and Srebro 2016; Corbett-Davies and Goel 2018).

Accounting for this richer set of metrics is our main contribution in the sections that follow. To do so, we first define a way to measure the degree of similarity between group-conditional score distributions, such that we can generalize a measurement of PR, TPR, and FPR across all thresholds (Section 3). Once we have defined this metric, which we call *distributional parity*, we can then rigorously characterize a procedure to maximize it (Section 4): We can develop a theoretically-backed approach that improves decision parity with respect to PR, TPR, or FPR at all decision thresholds τ, without having to retrain the underlying regressor f (Section 5).

3 Defining Distributional Parity

As in Jiang et al. (2020), we choose to measure the similarity of group-conditional score distributions using the Wasserstein-1 distance. To motivate this choice, we first explain what the Wasserstein-1 distance measures (Section 3). We then concretely show how it can be applied to capture our more-general, formal notion of *distributional parity*, which, unlike prior work, applies not just to PR and Demographic Parity, but also to TPR, FPR, and thus Equalized Odds and Equal Opportunity (Section 3).

Wasserstein-1 Distance

Informally, the Wasserstein-1 distance captures the difference between probability measures by measuring the *cost* of transforming one probability measure into the other. For two such measures $\mu_1, \mu_2 \in \mathcal{P}_1(\Omega)$ that satisfy Assumption 2.1, the Wasserstein-1 distance has a closed-form:

$$\mathcal{W}_1(\mu_1, \mu_2) = \int_\Omega |F^{-1}_{\mu_1}(\omega) - F^{-1}_{\mu_2}(\omega)| d\omega, \qquad (1)$$

where, recall from Section 2, the F^{-1}_μ are the inverse CDFs of the μ.

Additionally, as we suggest in Section 2, the Wasserstein-1 Distance (1) can be used to compute PR disparities across all thresholds Consider a single threshold $\tau \in [0, 1]$, where PR disparity is measured by taking the absolute difference in positive rates across two groups $a, b \in G$ at τ, i.e., $|\text{PR}_a(\tau) - \text{PR}_b(\tau)|$. A natural way to aggregate these single-threshold measurements into an all-threshold measurement is to take their sum across every possible τ, i.e.,

$$\mathcal{W}_1(\mu_a, \mu_b) = \int_0^1 |\text{PR}_a(\tau) - \text{PR}_b(\tau)| d\tau. \qquad (2)$$

Equivalently, we can view this sum as the average-case disparity taken uniformly over thresholds. If we let $U(\Omega)$ be a uniform distribution over Ω, this means that Equation (2) is equivalent to

$$\mathcal{W}_1(\mu_a, \mu_b) = \mathop{\mathbb{E}}_{\tau \sim U(\Omega)} |\text{PR}_a(\tau) - \text{PR}_b(\tau)|. \qquad (3)$$

This equivalence between (2) and (3) demonstrates that the Wasserstein-1 distance between group-conditional score distributions is equivalent to the average amount of PR disparity for groups a and b across all thresholds.

Distributional Parity

Based on the above, we can now present our first contribution: Extending (2) and (3) to account for TPR and FPR. We do so in the following definition:

Definition 3.1. For a fairness metric $\gamma \in \Gamma$, *distributional parity* for regressor f is satisfied when

$$\mathcal{U}_\gamma(f) \triangleq \mathop{\mathbb{E}}_{\tau \sim U(\Omega)} |\gamma_a(\tau) - \gamma_b(\tau)| = 0.$$

That is, the above definition generalizes the expression in Equation (3) to $\gamma_a, \gamma_b \in \Gamma$, rather than just considering PR_a and PR_b. If f attains parity at all thresholds for γ, i.e., $\mathcal{U}_\gamma(f) = 0$, then $\gamma_a(\tau) = \gamma_b(\tau)$ for all τ.

Note on prior work. Strong Demographic Parity (SDP), as originally presented in Jiang et al. (2020); Chzhen et al. (2020), can be understood as the special case of Definition 3.1, for which $\gamma = \text{PR}$.

Furthermore, it is important to note that our definition for distributional parity is closely related to $\mathcal{W}_1$ distance as described in Equation 3. This equivalence suggests that using techniques to reduce the Wasserstein-1 distance between group-conditional score distributions, e.g. μ_a, μ_b, will help us achieve our goal of developing a framework for enabling all-threshold parity for PR, TPR, or FPR. In the next section, we make this suggestion concrete by formally proving that geometric repair, based on optimal transport, can be used to maximize distributional parity via post-processing. Moreover, while distributional parity (Definition 3.1) considers a single γ and not *all* rates Γ at once, we will show that this post-processing method works for combinations of rates (Section 4). This allows practitioners to recover fairness metrics that consider multiple rates simultaneously, like Equalized Odds (Hardt, Price, and Srebro 2016).

4 Maximizing Distributional Parity

Relying on our definition for distributional parity (Definition 3.1), we can now present a theoretical characterization of how to maximize it. Like Chzhen et al. (2020), we also rely on *optimal transport* to accomplsih this (Section 4); additionally, we develop the key insight that using *geometric repair* enables us develop a method that also applies to TPR, FPR, and fairness definitions that use these rates (Section 4).

Optimal Transport and $\mathcal{W}_1$ Distance

For $\gamma = \text{PR}$ (Jiang et al. 2020), the best way to equate two measures μ_a and μ_b under $\mathcal{W}_1$— while also preserving accuracy — is to compute group-specific mappings T^*_g that

transform them onto some shared target representation μ_*, which ideally retains properties of the original μ_a and μ_b. More formally, if we transform $\mu_a \xrightarrow{T_a^*} \mu_*$ and $\mu_b \xrightarrow{T_b^*} \mu_*$, then clearly $T_a^*(\mu_a) = T_b^*(\mu_b) = \mu_*$. This means that both groups will share the same representation μ_*, and, by Definition 3.1, will satisfy distributional parity for fairness metric $\gamma \in \Gamma$ because, by Equation (1), $\mathcal{W}_1(T_a^*(\mu_a), T_b^*(\mu_b)) = \mathcal{W}_1(\mu_*, \mu_*) = 0$. Intuitively, we also want to make sure that μ_* is a good representation of the original measures μ_a and μ_b, as this will enable us to use μ_* to produce class predictions that satisfy γ across all thresholds, while also leaving the original class predictions under μ_a and μ_b mostly unchanged. This strategy for achieving distributional parity while retaining accuracy raises two questions: 1) how do we find the best target representation μ_*, and 2) how do we find the mappings T_g^* to produce it?

We employ two tools to answer these questions: *Wasserstein-1 barycenters* to reason about the ideal, shared target representation μ_*, and *optimal transport plans* to reason about the associated mappings T_g^*. We first introduce these two concepts generally, and then describe how prior work computes the specific Wasserstein-1 barycenter μ_* and optimal transport plans T_g^* to attain distributional parity for PR (i.e., Strong Demographic Parity). Then, we present our main proof result that extends to other metrics γ.

Wasserstein-1 barycenter. Informally, a Wasserstein barycenter (Agueh and Carlier 2011) is a weighted composition of two distributions, much like a weighted average or midpoint in the Euclidean sense; it provides a principled way to compose two measures. Consequently, we can use barycenters to compose μ_a and μ_b, making them our tool of choice for computing μ_*.

Definition 4.1. For two measures $\mu_a, \mu_b \in \mathcal{P}_1(\Omega)$ their λ-weighted *Wasserstein barycenter* is

$$\mu_\lambda \leftarrow \underset{\mu' \in \mathcal{P}_1(\Omega)}{\arg\min} (1-\lambda)\mathcal{W}_1(\mu_a, \mu') + \lambda \mathcal{W}_1(\mu_b, \mu');$$

and if following Assumption 2.1 is satisfied, admits a closed form (Santambrogio 2015, Thm 5.28)

$$\mu_\lambda = ((1-\lambda)\mathtt{id} + \lambda T_a^b)\#\mu_a. \tag{4}$$

To complete the weighted-average analogy, λ behaves like a tunable knob: As $\lambda \to 0$ then μ_λ will appear more like μ_a, and as $\lambda \to 1$ the more μ_λ will appear like μ_b. Given this definition, our first task now becomes finding the λ such that μ_λ is equivalent to our ideal target representation μ_*. In Section 4, we show concretely how to find the λ that satisfies this goal, thereby attaining distributional parity. Before going into these details, we outline how optimal transport plans help us solve the second task of finding the correct mappings T_g^*.

Optimal transport plans. The T_g^* that transform μ_a and μ_b into μ_* are called the *optimal* transport plans, where optimal refers to the least costly transformation from $\mu_g \to \mu_*$, with cost defined via the ℓ_1 norm on Ω (Section 3 and Appendix). For measures μ_a, μ_b, and their λ-weighted Wasserstein-1 barycenter μ_λ, (without loss of generality) the optimal transport plan from $\mu_a \to \mu_\lambda$ is denoted T_a^λ; it is defined as $T_a^\lambda(\omega) = F_{\mu_\lambda}^{-1}(F_{\mu_a}(\omega))$ where $\omega \in \Omega$ (Santambrogio 2015). Therefore, for the ideal target representation μ_* that yields distributional parity, we denote the associated optimal transport plans T_a^* and T_b^*.

Note about prior work. Jiang et al. (2020) uses λ-weighted Wasserstein-1 barycenters and optimal transport plans to find the μ_* and T_g^* that equalize PR for both groups at all thresholds. For the λ that yields μ_*, the authors set $\lambda = p_a$, where $p_a = \Pr(\boldsymbol{G} = a)$. They then show that the p_a-weighted Wasserstein-1 barycenter of μ_a and μ_b exactly computes our target μ_*, i.e.,

$$\mu_* \leftarrow \underset{\mu' \in \mathcal{P}_1(\Omega)}{\arg\min}\, p_a \mathcal{W}_1(\mu_a, \mu') + \underbrace{(1-p_a)}_{p_b} \mathcal{W}_1(\mu_b, \mu'), \tag{5}$$

which can be achieved with $T_g^*(\omega) = F_{\mu_*}^{-1}(F_{\mu_g}(\omega))$. The authors then define the Strong Demographic Parity regressor $f^*(x,s) \triangleq T_g^*(f(x,g))$; they show that, in addition to satisfying SDP, T_g^* minimizes changes to the predicted class, further corroborating our above described intuition.

Despite the success of this approach, it is easy to show that its restriction to PR, i.e., demographic parity, does not solve distributional parity for other metrics. Kleinberg (2018) and Chouldechova (2016) show that unless some very strict conditions are met (which do not hold in our setting), satisfying PR parity at any decision threshold guarantees *dis*parity in TPR or FPR at that same threshold; we cannot be fair for all metrics at once! This impossibility, paired with the myriad fair algorithmic scholarship advocating for better metrics, suggests the need for an all-threshold approach that accounts for other metrics.

Extending to other metrics using geometric repair

Similar to how Hardt, Price, and Srebro (2016) expanded fairness metrics beyond Demographic Parity, we expand the set of all-threshold fairness interventions beyond Strong Demographic Parity (Jiang et al. 2020); whereas Hardt, Price, and Srebro (2016) introduced methods for satisfying Equalized Opportunity/Odds, we introduce and optimize Distributional Parity (Definition 3.1), a general all-threshold framework that applies to multiple useful and common fairness metrics. In this section, we present our solution to produce all-threshold guarantees that apply to PR, TPR, or FPR. Our key insight is to use geometric repair — a tool that lets us prove all-threshold parity guarantees for all γ.

Geometric repair was initially proposed as a way to balance demographic parity against accuracy objectives at all thresholds; it interpolates between some regressor f and the SDP-corrected version of this same regressor, f^*.

Definition 4.2. Let f be a regressor and f^* be its SDP-corrected version. We call $\lambda \in [0,1]$ the *repair parameter* and define a *repaired* regressor f_λ as

$$f_\lambda(x,g) \triangleq (1-\lambda)f(x,g) + \lambda f^*(x,g)$$

Under this parametrization, $f_{\lambda=0}$ is the original f, and $f_{\lambda=1}$ recovers f^*. For this reason, we refer to f as the *unrepaired* regressor.

We prove properties of geometric repair that make it well-suited to address the all-threshold problem for other metrics. First, we show that geometric repair is not a destructive

transformation: We prove that it minimally impacts empirical risk while maximizing parity (i.e., *performance preservation*), and that it does not change the relative ordering scores (i.e., *rank preservation*). These results build on prior work (Feldman et al. 2015; Chzhen and Schreuder 2022). Our third result is a novel convexity theorem, for which we prove that, within the set of f_λ-repaired regressors, we can maximize distributional parity (Definition 3.1) (i.e., distributional parity is *convex* in λ).

Performance preservation. Suppose we want to bound the performance of f_λ under a risk minimization framework like Chzhen and Schreuder (2022); Le Gouic, Loubes, and Rigollet (2020). If we define risk as $\mathcal{R}_1(f_\lambda) \triangleq \|f - f_\lambda\|_1$ [1], then the following relationship between the risk of f_λ and f^* holds:

Proposition 1. For any f_λ, $\mathcal{R}_1(f_\lambda) \leq \mathcal{R}_1(f^*)$. More specifically, $\mathcal{R}_1(f_\lambda) = \lambda\mathcal{R}_1(f^*), \forall\lambda \in [0,1]$.

From this proposition, we can understand λ as a parameter that controls the trade-off between the risk of the Strong Demographic parity-attaining f^* and f. Moreover, Chzhen and Schreuder (2022) shows that geometric repair is *Pareto optimal* with regard to the fairness-accuracy trade-off: $\{f_\lambda\}_{\lambda\in[0,1]}$ forms a Pareto frontier in the multi-objective minimization of risk and maximization of distributional parity for PR.

Rank preservation. We also show that f_λ never changes the percentiles of scores induced by the original f. This entails a property called rational ordering, as introduced in Lipton, McAuley, and Chouldechova (2018, p. 6): "within each group, individuals with higher probability of belonging to the positive class are always assigned to the positive class ahead of those with lower probabilities."

Proposition 2. Any f_λ is rank preserving and therefore satisfies rational ordering. That is, for any $\lambda \in [0,1]$, $\tau \in \Omega$, and $(x_1, g), (x_2, g)$

$$\begin{aligned} \text{if} \quad & f(x_1, g) \leq f(x_2, g), \\ \text{then} \quad & f_\lambda(x_1, g) \leq f_\lambda(x_2, g). \end{aligned}$$

Convexity of distributional parity. Finally, we show that distributional parity is convex in the set of f_λ-repaired regressors, where convexity here implies the existence of some optimal f_{λ_*}. This means that we have finally arrived at our goal — we can find the repaired regressor that maximizes distributional parity (within the set). For readability, we include the details for this result in the Appendix. Here, we instead provide a proof sketch outlining our strategy.

Theorem 1. Recall $\mathcal{U}_\gamma$ denotes distributional parity (Definition 3.1). Fix $\gamma \in \Gamma$, and let f be a regressor and f_λ be this regressor under geometric repair for any $\lambda \in [0,1]$. The map $\lambda \mapsto \mathcal{U}_\gamma(f_\lambda)$ is convex in λ. That is, if λ_* satisfies

$$\lambda_* \leftarrow \arg\min_{\lambda\in[0,1]} \mathcal{U}_\gamma(f_\lambda),$$

then f_{λ_*} is the distributional-parity-maximizing regressor in the set of repaired regressors.

[1] This result appears in Chzhen and Schreuder (2022); Le Gouic, Loubes, and Rigollet (2020) works for the ℓ_2 norm.

Algorithm 1: Post-Processing for Distributional Parity

Input: Labeled training dataset $D = \{(x_i, g_i, y_i)\}_{i=1}^N$, regressor f, $\gamma_1 ... \gamma_m \subseteq \Gamma$

procedure FINDOPTIMAL($D, f, \gamma_1 ... \gamma_m$)

1. Compute the empirical barycenter $\hat{\mu}_*$ of $\hat{\mu}_a, \hat{\mu}_b$ from D (following Cuturi and Doucet (2014))
2. Use $\hat{\mu}_*$ to approximate the SDP regressor f^* (following Chzhen et al. (2020)), such that
$$f_\lambda(x, g) = (1 - \lambda)f(x, g) + \lambda f^*(x, g)$$
is well-defined.
3. Locate the optimal λ_*, i.e.,
$$\lambda_* \leftarrow \arg\min_{\lambda\in[0,1]} \sum_{k=1}^{m} \mathcal{U}_{\gamma_k}(f_\lambda).$$
4. **Output:** λ_* and f^*

end procedure

procedure REPAIREDINFER(x, g, f, f^*, λ_*)

5. **Output:** Repaired score s_{λ_*} for (x, g):
$$f_\lambda(x, g) = \underbrace{(1 - \lambda_*)f(x, g) + \lambda_* f^*(x, g)}_{s_{\lambda_*}}$$

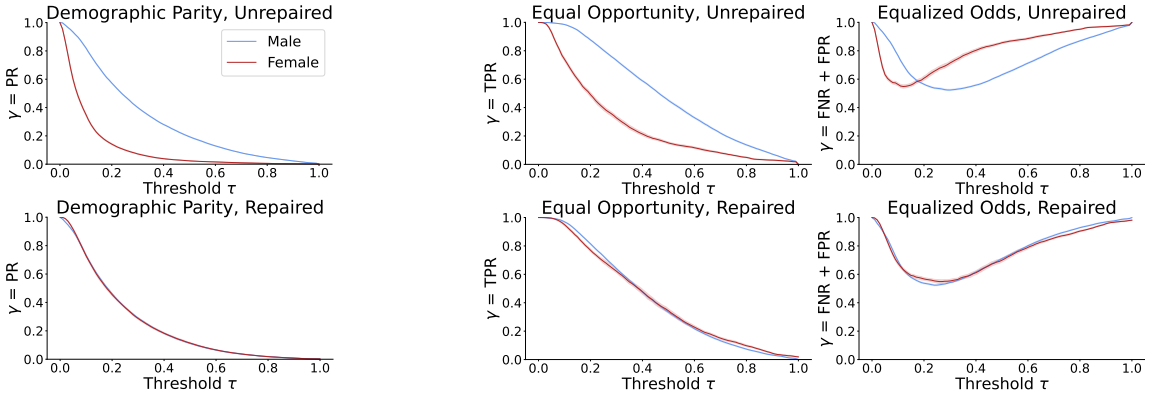
end procedure

Corollary 4.1. Since convex functions are closed under addition, the above theorem also applies to additive combinations of $\mathcal{U}_{\gamma_1}(f_\lambda) + \mathcal{U}_{\gamma_2}(f_\lambda) + ... + \mathcal{U}_{\gamma_m}(f_\lambda)$

Essentially, Theorem 1 shows that finding the distributional-parity-maximizing regressor simply reduces to a univariate optimization problem: Locating the optimal λ_* such that f_{λ_*} satisfies the desired fairness metric, with the following caveat. Our results for *maximizing* distributional parity do not claim that we achieve *perfect* distributional parity; however, we are able to show experimentally that our Algorithm 1 is remarkably successful in achieving parity in almost all cases (Section 6).

5 Post-Processing for Distributional Parity

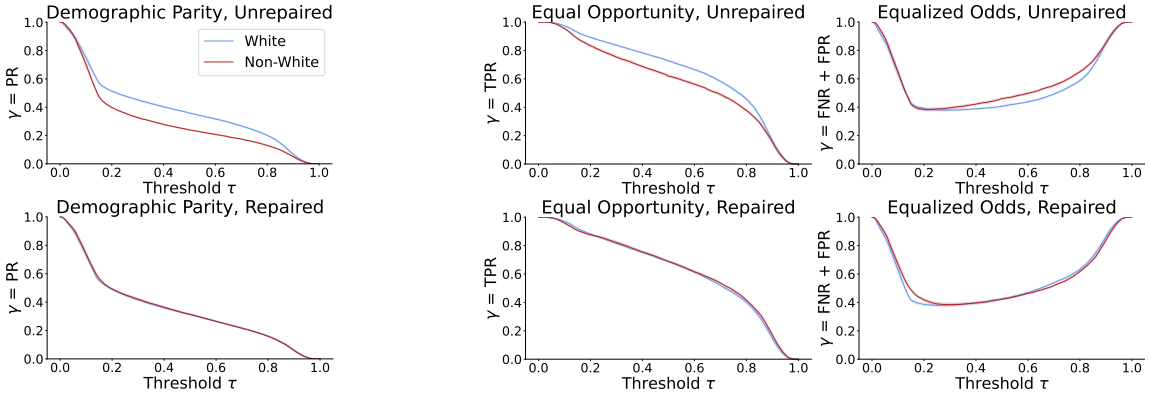
Our result in Theorem 1 naturally lends itself to a post-processing algorithm (Algorithm 1), which we describe in this section and test empirically in Section 6. At a high level, our algorithm approximates group-conditional score distributions for two groups a and b, finds their Wasserstein-1 barycenter, and then determines the amount of repair (i.e., λ_*) toward this barycenter that maximizes distributional parity for a given γ. Recall from Section 4 that, because distributional parity is convex in the space of repaired regressors, we are guaranteed to locate the optimal λ_*. In practice, at Step 3 we approximate $\mathcal{U}_\gamma(f_\lambda)$ (Definition 3.1) with a finite number of thresholds τ, e.g., $\tau \in \{0.01, ...0.99\}$. We then use an off-the-shelf univariate solver to locate the optimal λ_*, e.g., `scipy.minimize_scalar`. Also note that, although $\mathcal{U}$ is convex, we opt out of using differentiable convex optimization tools, since we cannot compute the derivative of $\mathcal{U}$ without a closed form for the probability density functions of μ_a and μ_b (Appendix).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

Figure 2: Comparing unrepaired and repaired Logistic Regression on `Adult` Income-Sex.



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

Figure 3: Comparing unrepaired and repaired SVMs on `Adult` Income-Race.

Takeaways. As stated in Sections 1 and 2, the need to retrain can severely limit the utility of a fairness intervention in practice. Practitioners may lack the resources to repeatedly retrain after adjusting the decision threshold, or may not have access to the training pipeline for logistical or proprietary reasons (Cooper et al. 2022; Kroll 2021; Alla and Adari 2021; Shankar et al. 2022). We highlight that our approach entirely sidesteps these issues because it *does not require model re-training*: In our algorithm, we only need to find the optimal λ_* once.

Note on prior work. The SDP solution from prior work (Jiang et al. 2020) or (Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020) in the $\mathcal{W}_2$ case corresponds to Steps 1-2 of Algorithm 1. Our approach encompasses SDP; without much additional cost (i.e., just the optimization at Step 3), we are able to produce repaired regressors for any weighted combination of $\gamma \in \Gamma$.

6 Experiments

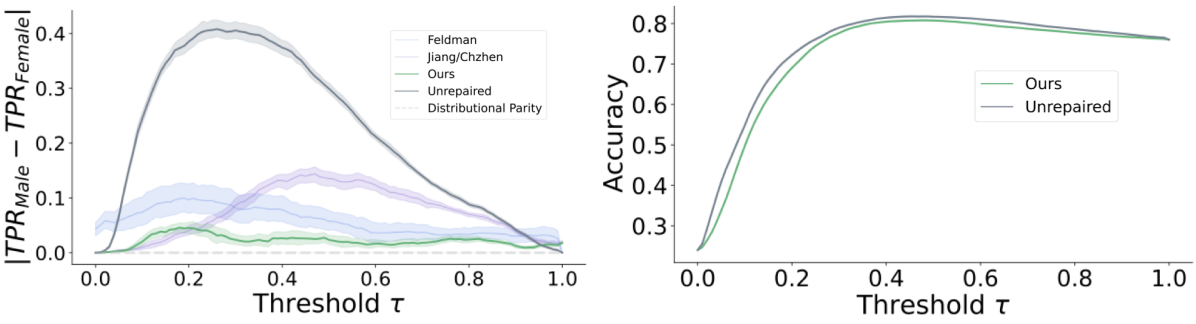
We test Algorithm 1 on several common algorithmic fairness datasets and models. We discuss only a small, representative subset of these results in the main paper, but include more comprehensive results in the Appendix.

Datasets and tasks. We highlight results for two datasets: `Adult` Income-Sex from the the UCI repository (Dua and Graff 2017), and `Adult` Income-Race from the new datasets produced in Ding et al. (2021). For both datasets, the task is to predict whether (1) or not (0) an individual's income exceeds $50,000. In `Adult` Income-Sex and `Adult` Income-Race, the protected attributes are sex and race, respectively, with these attribute names and values drawn from US census data.

Procedure. For each experiment, we split each dataset into three equal parts for 1) training the original regressor f, 2) finding the optimal λ_* and repaired regressor f^* (i.e., running FINDOPTIMAL in Algorithm 1), and 3) testing the fairness and accuracy of f^* to see how well it maximizes distributional parity (i.e., running REPAIREDINFER in Algorithm 1). We run this three-part procedure 10 times with different random seeds, and present statistics computed over these 10 trials.

In Step 1, we train f (either Logistic Regression (LR) or an SVM), using scikit-learn with its default model parameters and optimizers (Pedregosa et al. 2011). In Step 5, where we test f^* we produce binary classifications by thresholding f^* at a finite set of decision thresholds $\tau \in \{0, 0.01, 0.02, ..., 0.99, 1\}$. We compute $\gamma = \text{PR}$ and $\gamma = \text{TPR}$, as described in Section 2. We compute equalized odds by summing FPR and FNR, i.e., the misclassification rate.

We provide two sets of results: 1) validating that Algo-



(a) (b)

Figure 4: For LR on `Adult` Income-Sex and $\gamma = \text{TPR}$ (Equality of Opportunity), we show how Algorithm 1 (a) outperforms other methods in terms of distributional parity and (b) simultaneously preserves accuracy.

rithm 1 achieves almost-exact distributional parity for both SDP and other metrics; 2) showing that related methods under-perform Algorithm 1 in terms of distributional parity. Moreover, we show that Algorithm 1 also generally preserves accuracy.

Validating distributional parity. Figures 2 and 3 evaluate how well the repaired regressor f^* performs in comparison to the unrepaired regressor f, with respect to different fairness metrics γ for LR on `Adult` Income-Sex and SVMs on `Adult` Income-Race. Each plot shows the group-conditional scores at different thresholds. For each figure, the columns show different metrics γ; the top rows show how the unrepaired f performs for γ, while the bottom rows show f^*. Each f^* uses a different λ_* computed by Algorithm 5 (Appendix).

In both Figures 2 and 3, the overall takeaway is that Algorithm 1 effectively maximizes distributional parity for different datasets, models, and fairness metrics γ — achieving parity at almost every threshold. The top rows of both these figures demonstrate that f generally exhibits very different group-specific performance in terms of γ at different thresholds. In contrast, our repaired f^* does an excellent job of maximizing parity for γ across all τ, as is clear by the nearly-completely-overlapping group-specific performance curves. We also highlight that the plots in Figures 2a and 3a show results that can be achieved by both prior work (Jiang et al. 2020) and our algorithm, i.e., achieving distributional parity for $\gamma = \text{PR}$. In contrast, the plots shown in Figures 2b and 3b demonstrate results that are unique to our approach: Beyond SDP, Algorithm 1 can produce f^* that account for, e.g., Equal Opportunity or Equalized Odds.

Comparing distributional parity across methods. In Figure 4a, we provide an example illustrating how our method compares to others in achieving distributional parity for TPR. As discussed throughout this work, there are no existing methods that attempt to achieve fair classification at all thresholds for TPR and FPR, which makes direct comparisons challenging. Nevertheless, we pick two methods that we think are reasonable to examine: the SDP algorithm from Jiang et al. (2020); Chzhen et al. (2020) and the pre-processing geometric repair algorithm in Feldman et al. (2015). We believe that the former (in purple) is a natural baseline, given that our work generalizes SDP to other fairness metrics; since SDP can only perform full repair, we can interpret the results as similar to our algorithm, but with the repair parameter $\lambda = 1$. Since Feldman et al. (2015) (in blue) is a pre-processing method, unlike our algorithm, it repairs *inputs* rather than *outputs*. While Feldman et al. (2015) generally performs well (though worse than Algorithm 1), there are two additional observations worth emphasizing. First, Feldman et al. (2015) contains hyperparameters that require tuning to identify the *best possible* all-threshold performance that can be achieved for *any* repair-level λ. This means we have to train f multiple times before we can confidently produce the optimal f^*. In contrast, our method more-efficiently finds the optimal λ in the set of regressors in just one post-processing run. Second, as shown in the visualization of the standard error, our method achieves more-consistent performance than Feldman et al. (2015) across the 10 trials.

Altogether, the takeaway is that Algorithm 1 maximizes distributional parity — the green line in the figure is close to 0, indicating parity in TPR across thresholds. In contrast, while other methods improve upon the unrepaired scores, they are unable to match our method. Importantly, supporting our theory in Section 4, Figure 4b verifies that we are able to maximize distributional parity while achieving accuracy that is close to the original, unrepaired regressor.

7 Conclusion and Future Work

Our title, a play on words of the name Charles Dickens's popular work, succinctly summarizes our contribution: we show that by interpolating between group-conditional score distributions and their associated *measures*, we can generalize distributional approaches to all threshold fairness beyond demographic parity; namely, TPR and FPR, two common fairness *measures*. To this end, we introduce distributional disparity to *measure* decision parity at all thresholds, and provide a novel post-processing algorithm that 1) is theoretically-grounded by our convexity result, and 2) performs extremely well across a variety of benchmark datasets and tasks. In future work, we hope to position this work in context with other fairness metrics like calibration, and also in context with fair impossibility results.

Acknowledgements

Kweku Kwegyir-Aggrey and Jessica Dai are supported by Arthur. A. Feder Cooper is supported by the Artificial Intelligence Policy and Practice initiative at Cornell University and the John D. and Catherine T. MacArthur Foundation.

A Additional Background on Optimal Transport

In this section of the Appendix, we provide additional details for topics we could not fit into the main body of the paper. For readers familiar with Optimal Transport, much of the below material will not be new; nevertheless, we encourage all readers to use these sections as reference as they see fit.

Additional Notation. Let $O = \Omega \times Y \times G$. We define our probability measures w.r.t. a probability space $(O, \mathcal{F}, \mathbb{P})$ where $\mathcal{F}$ is a $\sigma-$algebra on O. $\mathbb{P}$ is the measure associated with the full joint distribution, and as defined above, μ are the measures for the underlying associated marginal distributions.

The $\texttt{id}$ function is shorthand for the identity function, i.e, $\texttt{id} : \Omega \to \Omega$ and $\texttt{id}(s) = s$ for $s \in \Omega$. Additionally we denote the push-forward operator as $f\#\mu$ (see Peyré and Cuturi (2018), Remark 2.5) where f is a measurable function pushing mass from a measure μ.

We also include the following remark, as we make use of it in the proof of Theorem 1 (Section B).

Remark A.1. If μ_a, μ_b are non-atomic, then T_a^b is continuous and non-decreasing (see Santambrogio (2015, p.55) where $T\#\mu_a = \mu_b$.

Barycenters

In general, and as stated above, Wasserstein-1 barycenters are defined as the solution to the following minimization problem:

Definition A.1. For two measures $\mu_a, \mu_b \in \mathcal{P}_1(\Omega)$, their λ-weighted *Wasserstein-1 barycenter* is

$$\mu_\lambda \leftarrow \underset{\mu' \in \mathcal{P}_1(\Omega)}{\arg\min} (1-\lambda)\mathcal{W}_1(\mu_a, \mu') + \lambda \mathcal{W}_1(\mu_b, \mu');$$

As shown in Santambrogio (2015, Thm 5.28), when Assumption 2.1 is satisfied, these measures satisfy a relatively simple closed form:

$$\mu_\lambda = ((1-\lambda)\texttt{id} + \lambda T_a^b)\#\mu_a. \tag{6}$$

In cases where we explicitly state that some barycenter-like-interpolation begins at a group-specific μ, we will include the group in the subscript, e.g., $\mu_{a,\lambda} = ((1-\lambda)\texttt{id} + \lambda T_a^b)\#\mu_a$.

Additionally, we include the following proposition which reveals a convenient property of barycenters that we make use of in the proof of Theorem 1 (Appendix B):

Proposition 3. (Proposition 1.3 from McCann (1997)) Let $\mu_a, \mu_b \in \mathcal{P}_1(\Omega)$ satisfy Assumption 2.1 and let μ_λ be the barycenter of μ_a and μ_b. Then $\mu_{a,\lambda} = \mu_{b,1-\lambda}$ and $\mu_{b,\lambda} = \mu_{a,1-\lambda}$.

Next, we introduce a result from prior work that show that the set of barycenters $\{\mu_\lambda\}_{\lambda \in [0,1]}$ admit a linear-like geometric structure in the space of probability measures, much like the linear interpolation performed in Equation (6). To clarify this result, we first introduce two new terms: Wasserstein Spaces, and (Wasserstein) Geodesics. First:

Definition A.2. The Wasserstein-1 Space is the set of distributions $\mathcal{P}_1(\Omega)$ endowed with the $\mathcal{W}_1$ metric.

In other words, the Wasserstein-1 Space is exactly a metric space over probability measures with finite first order moments, where the Wasserstein-1 distance is the chosen metric between them. To this end, it is well-known that the Wasserstein Distance satisfies the properties of a metric (Peyré and Cuturi 2018), e.g, the triangle inequality.

Second, we introduce geodesics. Informally, in a metric space, a geodesic is the shortest path between two points. Formally:

Definition A.3. Let $\lambda_1, \lambda_2 \in [0,1]$. A curve $\eta : [0,1] \to \mathcal{P}_1(\Omega)$ is a geodesic[2] in $\mathcal{W}_1(\Omega)$ if

$$\mathcal{W}_1(\eta(\lambda_1), \eta(\lambda_2)) = |\lambda_2 - \lambda_1| \mathcal{W}_1(\eta(0), \eta(1))$$

Geodesics are useful for making our informal, intuitive understanding of barycenters more precise. So far, we have described a barycenter as the least-expensive composition of two probability measures. We can alternatively understand this "least-expensive" intuition in precise geometric terms — i.e., as a geodesic.

Using Definitions A.2 and A.3, we state the following theorem:

Theorem 2. (Theorem 5.27 from Santambrogio (2015)) Suppose that Ω is a convex set. Let $\mu_a, \mu_b \in \mathcal{P}_1(\Omega)$ satisfy Assumption 2.1. If T_a^b is an optimal transport plan from μ_a to μ_b then the set of barycenters between these distributions is exactly the curve $\mu_\lambda = ((1-\lambda)\texttt{id} + \lambda T_a^b)\#\mu_a$, which is a geodesic in $\mathcal{W}_1(\Omega)$.

Informally, this theorem states that interpolation along λ-barycenters produces a geodesic in $\mathcal{W}_1(\Omega)$

It is important to note that, in general, the Wasserstein-1 space is not convex (Adve and Mészáros 2020); however, we can exploit the geometry of geodesics to make special-case convexity claims.

Connecting Barycenters and Geometric Repair

Recall from the definition of geometric repair that we can define a repaired regressor f_λ as

$$f_\lambda(x, g) \triangleq (1-\lambda) f(x, g) + \lambda f^*(x, g),$$

where $\lambda \in [0,1]$ is the repair parameter and $f^*(x,g) = T_g^*(f(x,g))$. Given that $f(x,g)$ is the original score, and $f^*(x,g) = T_g^*(f(x,g))$ is the transformed score, we can we equivalently define geometric repair as

$$f_\lambda(x, g) = ((1-\lambda)\texttt{id} + \lambda T_g^*) \circ f(x, g)$$

[2]Most precisely, this is the definition for a constant-speed geodesic, however we ignore this imprecision for clarity.

If we we use the geometric repair definition to push-forward μ_g, the measure by which the scores for group g are distributed, then we obtain the measure(s) which govern f_λ

$$((1-\lambda)\mathtt{id} + \lambda T_g^*)\#\mu_g = Law(f_\lambda(\boldsymbol{X}, g)), \quad (7)$$

for which we denote $Law(f_\lambda(\boldsymbol{X}, g))$ as $\mu_{g,\lambda}$. From this equality we can draw a connection between wasserstein barycenters and geometric repair – ultimately showing that as we vary the repair parameter, we are actually interpolating between each group's score distribution and the SDP barycenter distribution μ_*.

Which λ is Which? So far, we've introduced two styles of notation to discuss the interpolation of measures via barycenters, they are μ_λ, and $\mu_{g,\lambda}$. To disambiguate their difference, consider the following:

1. μ_λ denotes λ-weighted barycenters in a general context, and is not specific to geometric repair. Recall the formula for μ_λ, which is based on our general-case formula for barycenters (see Section A and Equation (6))

   $$\mu_\lambda = ((1-\lambda)\mathtt{id} + \lambda T_a^b)\#\mu_a$$

   In the above, we omit groups from the subscript because we tend to emphasize the transportation is happening *from μ_a to μ_b*.

2. $\mu_{g,\lambda}$ denotes the interpolation of some group's conditional distribution toward its p_a-barycenter, denoted μ_* (recall this is the SDP solution by Equation (5)) as a result of geometric repair, i.e.,

   $$\mu_{g,\lambda} = ((1-\lambda)\mathtt{id} + \lambda T_g^*)\#\mu_g. \quad (8)$$

   Under geometric repair, *each* group's distribution is moving toward the barycenter, so the difference between group a's λ-repaired distribution, and group b's λ-repaired distribution, is significant – hence the additional notation.

With this in mind, we revisit Proposition 3 in the context of geometric, and introduce the following Lemma, which we use in the proof of Theorem 1.

Lemma A.1. Let $\mu_a, \mu_b \in \mathcal{P}_1(\Omega)$ satisfy Assumption 2.1 and let $\mu_{a,\lambda}$ be the λ-barycenter of μ_a and μ_*, and let $\mu_{b,\lambda}$ be the λ-barycenter of μ_b and μ_* then

$$\mu_{a,\lambda} = \mu_{b,\frac{1-p_a\lambda}{1-p_a}}$$
$$\mu_{b,\lambda} = \mu_{a,\frac{1-\lambda}{p_a}+\lambda}$$

Proof. Recall by Equations (8) and (6) that $\mu_{a,\lambda} = \mu_{\lambda(1-p_a)}$. Similarly, $\mu_{b,t} = \mu_{1-tp_a}$. To prove the Lemma, we need some λ s.t. $\lambda = t$, such that λ becomes geometric repair parameter for both groups. Letting $\lambda(1-p_a) = 1-tp_a$ and solving for λ, yields the proposition, i.e., $\lambda = \frac{1-tp_a}{1-p_a}$ and therefore $\mu_{a,\lambda} = \mu_{b,\frac{1-p_a\lambda}{1-p_a}}$. The second equivalence follows symmetrically, by swapping t and λ.

□

Fairness Metrics as Functionals In general, we can view fairness metrics as a functionals over the space of probabilities. Specifically, the fairness metrics we consider in this work can be written as a "potential energy" functional and are defined by the integral of a given function taken over some measure.

Definition A.4 (Potential Energy)**.** The potential energy of a function $V : \Omega \to \mathbb{R}$ over some measure $\mu \in \mathcal{P}_1(\Omega)$ is defined

$$\mathcal{V}(\mu) = \int_\Omega V d\mu. \quad (9)$$

To express fairness metrics as a potential energy functional we need: the indicator function, a threshold, and the some group conditioned score measure. For example, we can re-write the positive rate for group g as

$$\text{PR}_g(\tau) = \mathop{\mathbb{E}}_{s\sim\boldsymbol{S}}[\mathbb{1}_{s\geq\tau}|\boldsymbol{G}=g] = \int_\Omega \mathbb{1}_{s\geq\tau} d\mu_g. \quad (10)$$

where the last equality follows by definition of conditional expectation. Similarly, we could express $\gamma \in \{\text{TPR}, \text{FPR}\}$

$$\gamma_g(\tau) = \mathop{\mathbb{E}}_{s\sim\boldsymbol{S}}[\mathbb{1}_{s\geq\tau}|\boldsymbol{G}=g, \boldsymbol{Y}] = \int_\Omega \mathbb{1}_{s\geq\tau} d\mu_{g|Y}. \quad (11)$$

Displacement Convexity Conceptually, displacement convexity describes the phenomena that some functional on probability measures is convex as one interpolates between two distributions, along their barycenters – recall that these barycenters form a geodesic (Theorem 2). Since the interpolation of score distributions via geometric repair produces a Wasserstein geodesic, displacement convexity is a tool well suited to help us characterize the convexity of fairness metrics under geometric repair.

Definition A.5. (Definition 7.2 from Agueh and Carlier (2011))A functional $\mathcal{V} : \mathcal{P}_1(\Omega) \to [0,1]$ is said to be displacement convex if the mapping $\lambda \mapsto \mathcal{V}(\mu_\lambda)$ is convex where $\mu_\lambda = ((1-\lambda)\mathtt{id} + \lambda T_a^b)\#\mu$ is a geodesic between any $\mu_a, \mu_b \in \mathcal{P}_1(\Omega)$ that satisfy Assumption 2.1.

The following theorem provides a necessary and sufficient condition to determine if potential energy functional is displacement convex. Unsurprisingly, the displacement convexity of the functional $\mathcal{V}$ rests entirely on the convexity of the function V that it integrates over.

Theorem 3. (Proposition 7.25 from Santambrogio (2015) and Proposition 7.7 from (Agueh and Carlier 2011)) The functional $\mathcal{V}$ is displacement convex iff V is convex.

In the case of fairness metrics, the function V is the indicator function over a score $s \in \Omega$ at some τ i.e. $\mathbb{1}_{[s\geq 1]}$. In general, the indicator function is convex, when the set it is indicating is also convex set (Frémond 2017). Clearly, $[\tau, 1]$ is a convex set, making the indicator function a convex function in our setting. From this, we can easily establish that $\gamma(\cdot)$ by Equation (11).

B Proofs

Proof of Proposition 1

Recall that Proposition 1 claims that geometric repair (Definition 4.2) is performance-preserving. In Section 4, we in-

formally define performance preservation to mean that parity is maximized while minimally impacting empirical risk.

Formally, we suppose that we want to bound the performance of f_λ under a risk minimization framework like Chzhen and Schreuder (2022); Le Gouic, Loubes, and Rigollet (2020). If we define risk as $\mathcal{R}_1(f_\lambda) \triangleq \|f - f_\lambda\|_1$, then the following relationship between the risk of f_λ and f^* holds:

Proposition 1, For any f_λ, $\mathcal{R}_1(f_\lambda) \leq \mathcal{R}_1(f^*)$. More specifically, $\mathcal{R}_1(f_\lambda) = \lambda\mathcal{R}_1(f^*), \forall\lambda \in [0,1]$.

Proof. By definition $\mathcal{R}_1(f_\lambda) = \|f - f_\lambda\| = \|f - (f + \lambda(f^* - f))\| = \lambda\|f - f^*\| = \lambda\mathcal{R}_1(f^*)$. The claimed inequality easily follows. □

We note that this result for the ℓ_1 norm is directly related to those that appear in Chzhen and Schreuder (2022); Le Gouic, Loubes, and Rigollet (2020) for the ℓ_2 norm.

Proof of Proposition 2

Recall that Proposition 2 claims that geometric repair (Definition 4.2) is rank-preserving. In Section 4, we informally define rank preservation to mean that the repaired regressor f_λ never changes the percentiles of scores induced by the original, unrepaired f. This entails a property called rational ordering, introduced in Lipton, McAuley, and Chouldechova (2018, p. 6), which informally means that "within each group, individuals with higher probability of belonging to the positive class are always assigned to the positive class ahead of those with lower probabilities." Formally,

Proposition 2. Any f_λ is rank preserving and therefore satisfies rational ordering. That is, for any $\lambda \in [0,1]$, $\tau \in \Omega$, and $(x_1, g_1), (x_2, g_2)$ then

$$\begin{aligned} \text{if} \quad & f(x_1, g_1) \leq f(x_2, g_2), \\ \text{then} \quad & f_\lambda(x_1, g_1) \leq f_\lambda(x_2, g_2). \end{aligned}$$

Proof. As shown in Section A we can write f_λ as $f_\lambda(x, g) = ((1-\lambda)\mathtt{id} + \lambda T_g^*) \circ f(x, g)$. However, the sum $(1-\lambda)\mathtt{id} + \lambda T_g^*$ is non-decreasing since it is the sum of non-decreasing T_g^* (Remark A.1) and $\mathtt{id}$ (by Definition). Since f_λ is non-decreasing, we can conclude the proof. □

Proof of Theorem 1

Theorem 1. Recall $\mathcal{U}_\gamma$ denotes distributional parity (Definition 3.1). Fix $\gamma \in \Gamma$, and let f be a regressor and f_λ be this regressor under geometric repair (Definition 4.2) for any $\lambda \in [0,1]$. The map $\lambda \mapsto \mathcal{U}_\gamma(f_\lambda)$ is convex in λ. That is, if λ_* satisfies

$$\lambda_* \leftarrow \arg\min_{\lambda \in [0,1]} \mathcal{U}_\gamma(f_\lambda),$$

then f_{λ_*} is the distributional-parity-maximizing regressor in the set of repaired regressors.

Proof. Let $\gamma \in \Gamma$. To prove convexity, we show that $\frac{d^2}{d\lambda^2}\mathcal{U}_\gamma(f_\lambda)$ is non-negative everywhere. First, we remind readers the definition of $\mathcal{U}_\gamma(f_\lambda)$ (distributional parity):

$$\mathcal{U}_\gamma(f_\lambda) \triangleq \mathop{\mathbb{E}}_{\tau \sim U(\Omega)} |\gamma_a(\tau) - \gamma_b(\tau)|.$$

where γ_g is a fairness metric on the score distributions of f_λ for group $g \in G$.

To compute this derivative, we will first compute the derivative of $\gamma_a(\tau) - \gamma_b(\tau)$, and then use this to compute the derivative of the $\mathbb{E}_{\tau \sim U(\Omega)} |\gamma_a(\tau) - \gamma_b(\tau)|$.

First, we analyze the derivative(s) of γ_a (wlog). As stated in Section A, we can express γ_a as a potential energy functional (see Definition A.4) of some repaired group-conditional score measure $\mu_{g,\lambda}$. Let $V_\tau(s) := \mathbb{1}_{s \geq \tau}$ where $s \in \Omega$. Then γ_a is defined (see Equation 11)

$$\gamma_a(\tau) = \int_\Omega V_\tau d\mu_{a|\boldsymbol{Y},\lambda}$$

where $\mu_{a|Y,\lambda}$ is $\mu_{a,\lambda}$ further conditioned on $\boldsymbol{Y}$. As we proceed, we suppress the $\boldsymbol{Y}$ from the notation, instead writing $\mu_{a,\lambda}$; however, the reader should keep in mind that this result is general, and that this suppression is for brevity and will not affect the computations that follow.

By Theorem 2, $\{\mu_{a,\lambda}\}_{\lambda \in [0,1]}$ is a geodesic. Therefore, we can rewrite any $\mu_{a,\lambda}$ as $\mu_{a,\lambda} := ((1-\lambda)id + \lambda T_a^*)\#\mu_a$. Recall T_a^* is the optimal transport plan from μ_a to the p_a-weighted barycenter μ_* As one final notational convenience, we use $\pi_{g,\lambda}$ to denote $(1-\lambda)id + \lambda T_g^*$. We'll first utilize this notation for the group a case. Using these substitutions, we have (wlog) that $\mu_{a,\lambda} := (\pi_{a,\lambda})\#\mu_a$, so γ_a can be equivalently written

$$\gamma_a(\tau) = \int_\Omega V_\tau d(\pi_{a,\lambda}\#\mu_a).$$

Next we'll apply several change of variables in order to write the above integral w.r.t the Lebesgue measure. The first change of variables follows by definition of the push-forward operator

$$\gamma_a(\tau) = \int_{\pi_{a,\lambda}^{-1}(\Omega)} V_\tau(\pi_{a,\lambda}) d\mu_a = \int_\Omega V_\tau(\pi_{a,\lambda}) d\mu_a.$$

We use the following argument to reason why the domain of integration is unchanged after the reader: by Remark, A.1 that T is continuous and non-decreasing. Over a closed domain $\mathtt{id}$ is also continuous and non-decreasing. Thus, their sum, π_λ must a bijective function (since it is continuous and monotone on a closed domain). Since $\pi_{g,\lambda}$ is a bijective mapping from $\Omega \to \Omega$, it follows that $\pi_{g,\lambda}^{-1}(\Omega) = \Omega$.

Next, we use absolutely continuity to re-write the above in terms of ℓ, i.e, by Assumption 2.1 μ_a is absolutely continuous with respect to ℓ meaning that by the Radon Nikodym-Theorem

$$\gamma_a(\tau) = \int_\Omega V_\tau(\pi_{g,\lambda}) d\mu_a = \int_\Omega \rho_{\mu_a} V_\tau(\pi_{g,\lambda}) d\ell$$

where ρ_{μ_a} is the Radon Nikodym Derivative, i.e., the probability density function associated with μ_a. This wraps up our definition γ_a. We'll also need to define this $b \in G$. To do this, we invoke Lemma A.1, i.e., $\mu_{b,\lambda} = \mu_{a, \frac{1-\lambda}{p_a} + \lambda}$. With this in mind, we can write the following, which symmetrically follows from the argument above for the μ_a case:

$$\gamma_b(\tau) = \int_\Omega \rho_{\mu_a} V_\tau(\pi_{b, \frac{1-\lambda}{p_a} + \lambda}) d\ell$$

Next, let $h_{a,\tau}(\lambda)$ be the mapping $\lambda \mapsto \mathcal{V}_\tau(\mu_{a,\lambda})$ and $h_{b,\tau}(\lambda)$ be defined similarly as $\lambda \mapsto \mathcal{V}_\tau(\mu_{a,\frac{1-\lambda}{p_a}+\lambda})$. Then, the difference $h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)$ is simply $\gamma_a(\tau) - \gamma_b(\tau)$ for f_λ. We take the derivative of this difference, i.e.,

$$\begin{aligned}\frac{d}{d\lambda}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)] &= \frac{d}{d\lambda}\int_\Omega \rho_{\mu_a} \\ &\quad \cdot [V_\tau(\pi_{a,\lambda}) - V_\tau(\pi_{b,\frac{1-\lambda}{p_a}+\lambda})]d\ell \\ &= \int_\Omega \rho_{\mu_a} \cdot \Bigg[\frac{d}{d\lambda}\Big(V_\tau(\pi_{a,\lambda}) \\ &\quad - V_\tau(\pi_{b,\frac{1-\lambda}{p_a}+\lambda})\Big)\Bigg] d\ell\end{aligned}$$

where the last equality in the above line follows from Leibniz Rule. Straightf-orward computation of the derivative of $V_\tau(\cdot)$ shows

$$\frac{d}{d\lambda}V_\tau(\pi_{a,\lambda}) = V_\tau^{'}(\pi_{a,\lambda})(T_a^* - \texttt{id})$$

and

$$\frac{d}{d\lambda}V_\tau(\pi_{b,\frac{1-\lambda}{p_a}+\lambda})$$

$$= \left(\frac{p_a - 1}{p_a}\right) V_\tau^{'}(\pi_{b,\frac{1-\lambda}{p_a}+\lambda})\texttt{id} - \left(\frac{1-p_a}{p_a}\right) V_\tau^{'}(\pi_{b,\frac{1-\lambda}{p_a}+\lambda})T_b^*,$$

indicating that the derivative we're after is nothing more than

$$\begin{aligned}\frac{d}{d\lambda}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)] = \int_\Omega \rho_{\mu_a} \cdot \Bigg[&T_a^*(V_\tau^{'}(\pi_{a,\lambda})) \\ &+ \left(\frac{1-p_a}{p_a}\right) V_\tau^{'}(\pi_{b,\frac{1-\lambda}{p_a}+\lambda})T_b^* \\ &- \left(\frac{p_a - 1}{p_a}\right) V_\tau^{'}(\pi_{b,\frac{1-\lambda}{p_a}+\lambda})\texttt{id} \\ &- V_\tau^{'}(\pi_{a,\lambda})\texttt{id})\Bigg] d\ell.\end{aligned}$$

By its definition, the derivative of $V_\tau^{'}(\pi_{g,\lambda})$ is exactly the Dirac delta function $\delta(\pi_{g,\lambda} - \tau)$. Making this substitution, the above yields

$$\begin{aligned}\frac{d}{d\lambda}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)] = \\ \int_\Omega \rho_{\mu_a} \cdot \Bigg[&T_a^*(\delta(\pi_{a,\lambda} - \tau)) \\ &+ \left(\frac{1-p_a}{p_a}\right) \delta(\pi_{b,\frac{1-\lambda}{p_a}+\lambda} \\ &- \tau)T_b^* \\ &- \left(\frac{p_a - 1}{p_a}\right) \delta(\pi_{b,\frac{1-\lambda}{p_a}+\lambda} \\ &- \tau)\texttt{id} - \delta(\pi_{a,\lambda} - \tau)\texttt{id})\Bigg].\end{aligned}$$

Then, by definition of δ, we at last obtain

$$\begin{aligned}\frac{d}{d\lambda}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)] = \Bigg[&T_a^* + \left(\frac{1-p_a}{p_a}\right) T_b^* \\ &- \left(\frac{p_a - 1}{p_a}\right) \texttt{id} - \texttt{id}\Bigg] \circ \tau.\end{aligned}$$

Since the above does not depend on λ, taking another derivative simply yields $\frac{d^2}{d\lambda^2}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)] = 0$. From this, we take the second derivative of the absolute value of this difference, i.e.,

$$\begin{aligned}&\frac{d}{d^2\lambda}|h_{a,\tau} - h_{b,\tau}| = \\ &\quad \text{sign}(h_{a,\tau} - h_{b,\tau}) \underbrace{\frac{d^2}{d\lambda^2}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)]}_{=0} \\ &\quad +2\underbrace{\delta(h_{a,\tau} - h_{b,\tau})}_{\simeq 0 \text{ or } 1}\underbrace{(\cdot)^2}_{\geq 0}.\end{aligned} \tag{12}$$

The first term on the r.h.s., we've already shown is zero, and the second term is also non-negative. Another application of Leibniz' Rule allows that

$$\frac{d}{d^2\lambda}\underbrace{\mathop{\mathbb{E}}_{\tau\sim U(\Omega)}|h_{a,\tau} - h_{b,\tau}|}_{\mathcal{U}_\gamma(f_\lambda)} = \mathop{\mathbb{E}}_{\tau\sim U(\Omega)}\left|\underbrace{\frac{d}{d^2\lambda}[h_{a,\tau} - h_{b,\tau}]}_{\geq 0 \text{ by (12)}}\right|.$$

This indicates that $\mathcal{U}_\gamma(f_\lambda)$ is convex (i.e., we have shown that the second derivative is non-negative). The existence of a a solution to $\lambda_* \leftarrow \arg\min_{\lambda\in[0,1]} \mathcal{U}_\gamma(f_\lambda)$ is guaranteed by this fact, therefore indicating that f_{λ_*} maximizes distributional parity in the set of repaired regressors. □

As noted in Section 4, the validity of Theorem 1 indicates that finding the distributional-parity-maximizing regressor reduces to a univariate optimization problem. Locating the optimal λ_*, such that f_{λ^*} satisfies the desired fairness metric γ, with the following caveat: Our results for *maximizing* distributional parity do not claim that we achieve *perfect* distributional parity; however, we are able to show experimentally that our Algorithm 1 is remarkably successful in achieving parity in almost all cases (Section 6 and Appendix C).

Corollary 4.1. Since convex functions are closed under addition, Theorem 1 also applies to additive combinations of $\mathcal{U}_{\gamma_1}(f_\lambda) + \mathcal{U}_{\gamma_2}(f_\lambda) + ... + \mathcal{U}_{\gamma_m}(f_\lambda)$

C Experimental Configuration and Additional Results

We provide details about our Algorithm, and experimental configurations in Section 6, as well as additional results on other datasets.

Algorithm 1 and Differentiation.

We show in the proof of Theorem B that computing the derivative of $\mathcal{U}_\gamma(f_\lambda)$ requires knowledge of the probability density function of either group's distribution of scores (denoted ρ_{μ_g}). Often, this distribution is not known, therefore preventing us from using more traditional, derivative based optimization techniques.

Experimental configuration details, and reproducibility for main paper results

Code for all experiments, as well as instructions for reproducing our specific results, can be found in the anonymous repository located https://anonymous.4open.science/r/distributional-fairness-436F/here.

For Figure 4 in the main paper, which compares our method for achieving distributional parity against other all-threshold fairness methods, we needed to do extensive hyperparameter tuning to find the best-performing repair parameter for the algorithm in Feldman et al. (2015). In more detail, the algorithm in Feldman et al. (2015) includes a `repairlevel` hyperparameter, which, similar to our λ (found through optimization, rather than a chosen hyperparameter), controls the extent to which inputs are adjusted during pre-processing. We performed grid search using `repairlevel` $\in \{0.05, 0.15, 0.2, ..., 0.95, 1.0\}$. When comparing our method to that in Feldman et al. (2015) for a given γ, we do so using the `repairlevel` that yielded the best results from grid search for that γ. For our implementation, we incorporated the open-source implementation of Feldman et al. (2015), available in the Fairlearn package (Bird et al. 2020).

Additional experiments

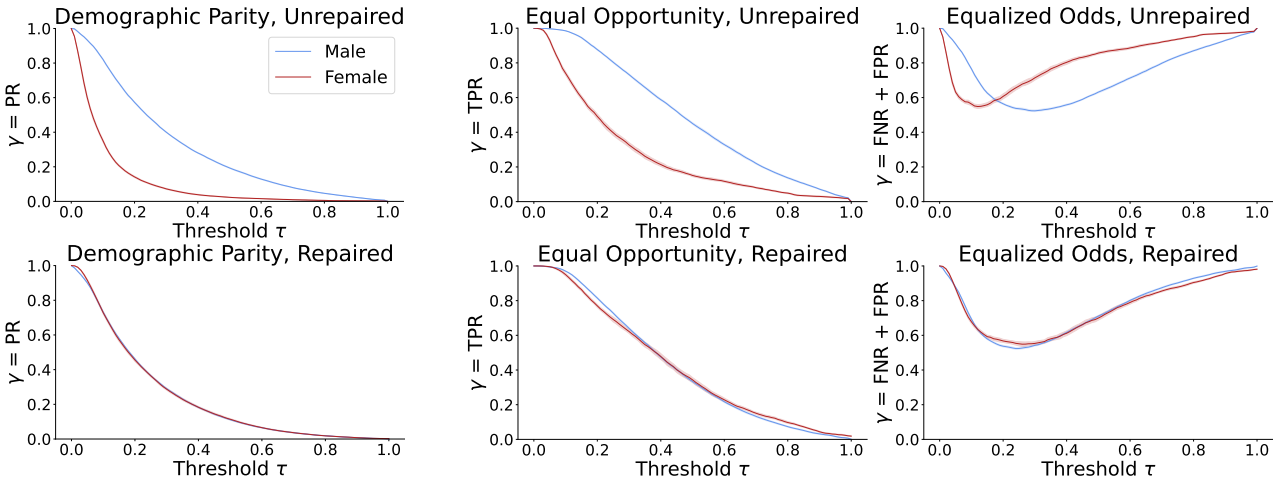
Additional dataset. In addition to the two datasets from the main paper, we also show results here for the `Taiwan Credit` dataset (Bay et al. 2000), where the sensitive attribution is education level and the target variable is good credit.

Additional models. In addition to Logistic Regression and SVM models, we show results below for scores generated by an underlying Random Forest and 2-layer neural network; again, we use `scikit-learn` implementations with default hyperparameters.

Additional Discussion. In our additional experiments we show that our method works very well across a combination of models and datasets however, there are two main minor limitations we'd like to discuss. The experiment in which we removed the least disparity, was SVMs on the `Adult` Income-Sex task, for Equalized Odds, see Figure 6. We have hypothesized that our method may be ineffective here due an abnormality in the classifier's score distributions. Specifically, we point out the SVM hardly assigns scores $< .15$, for the Female group, thereby inflating error rates for members of that group, with scores above 0.15. We look forward to investigating ways to improve our performance on this task in future work. Additionally, there are some cases in which we record very little disparity in the unrepaired regressor. For some of these cases, despite increasing parity at almost every threshold, our method *marginally* decreases parity at certain select threshold(s), see for example, in Figure 10 at $\tau = 0.2$. We believe that this effect is neither pervasive nor significant; in most cases, we do not see any parity decreases. With that said, in the scenarios where parity slightly decreases at a threshold, we note that our algorithm still increases total parity on average, almost achieving parity across all other thresholds.

The following figures represent additional results for models trained on:

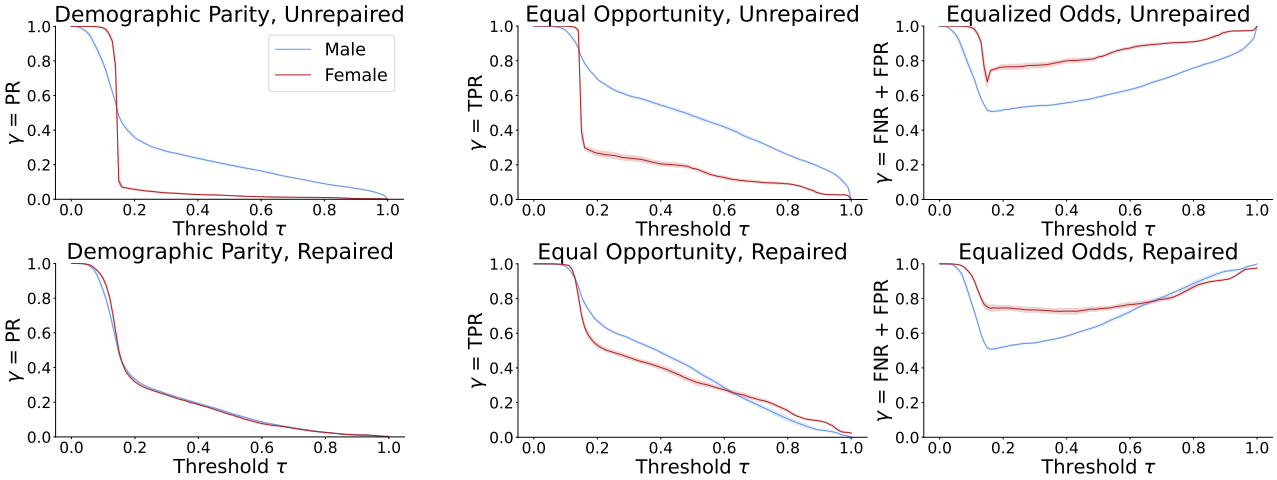
- `Adult` UCI (Dua and Graff 2017) (Figures 5, 6, 7, 8)
- New `Adult` (Ding et al. 2021) (Figures 9, 10, 11, 12)
- `Taiwan Credit`(Bay et al. 2000) (Figures 13, 14, 15, 16)



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

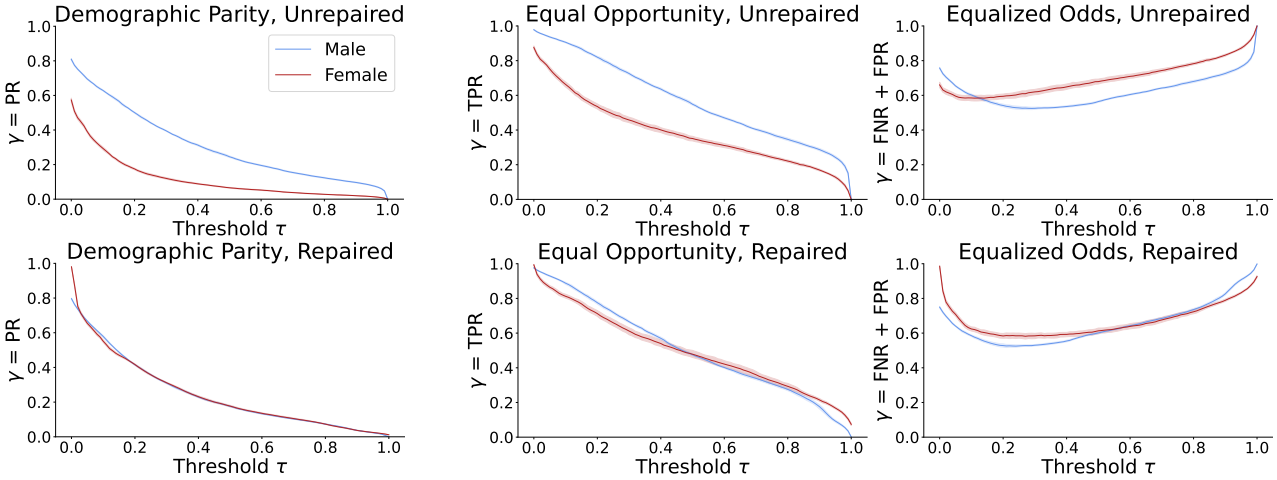
Figure 5: Comparing unrepaired and repaired logistic regression on `Adult` Income-Sex (Dua and Graff 2017).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

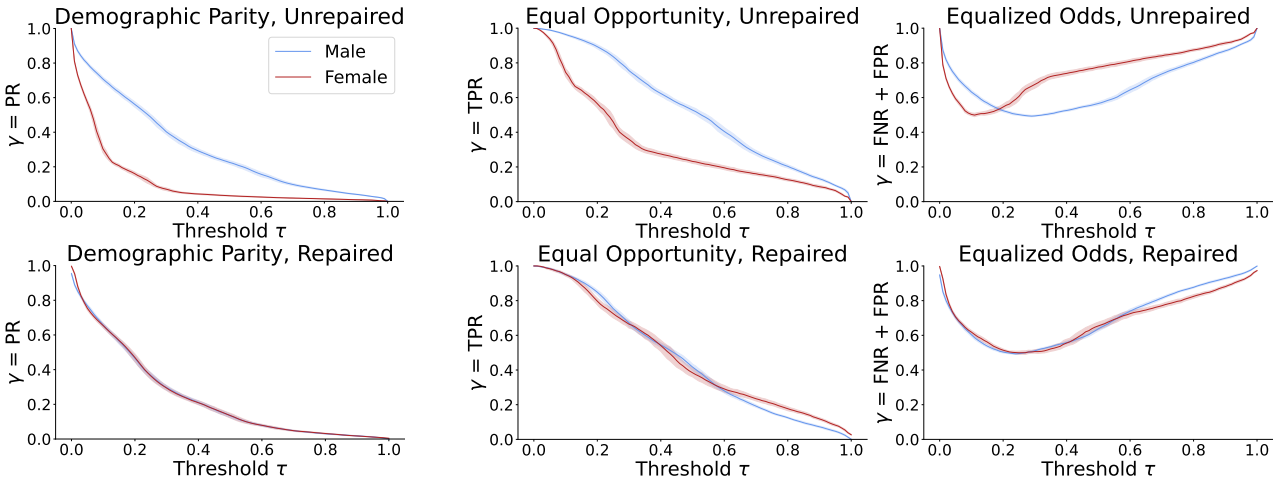
Figure 6: Comparing unrepaired and repaired SVMs on `Adult` Income-Sex (Dua and Graff 2017).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

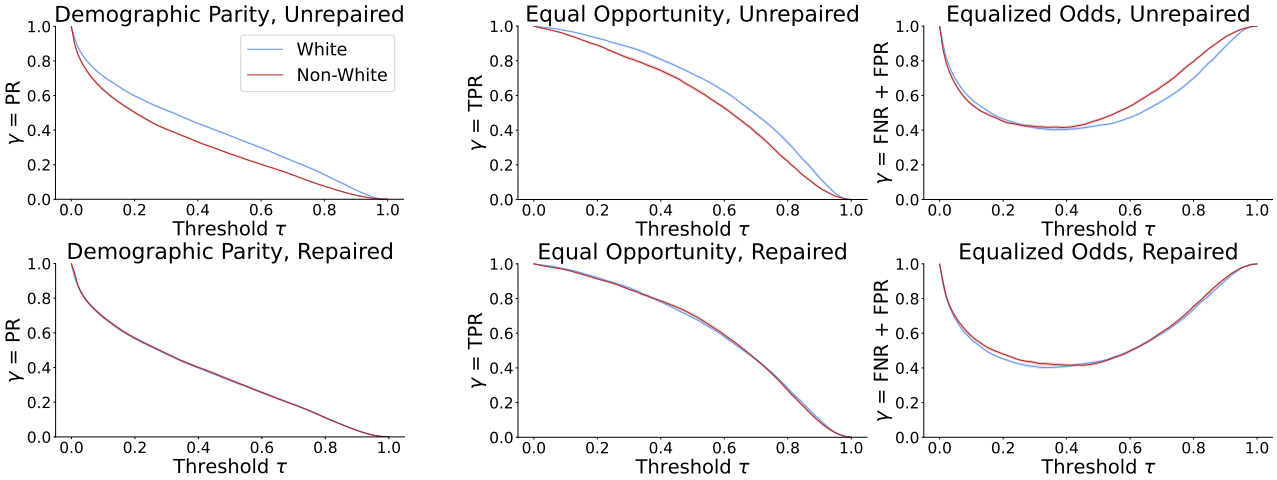
Figure 7: Comparing unrepaired and repaired Random Forests on `Adult` Income-Sex (Dua and Graff 2017).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

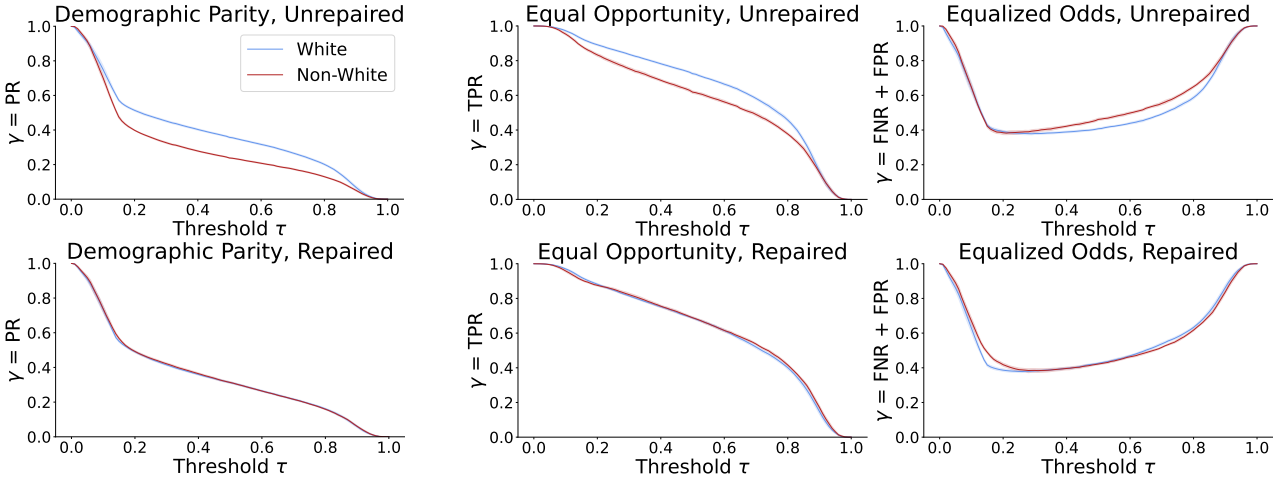
Figure 8: Comparing unrepaired and repaired MLPs on `Adult` Income-Race (Dua and Graff 2017).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

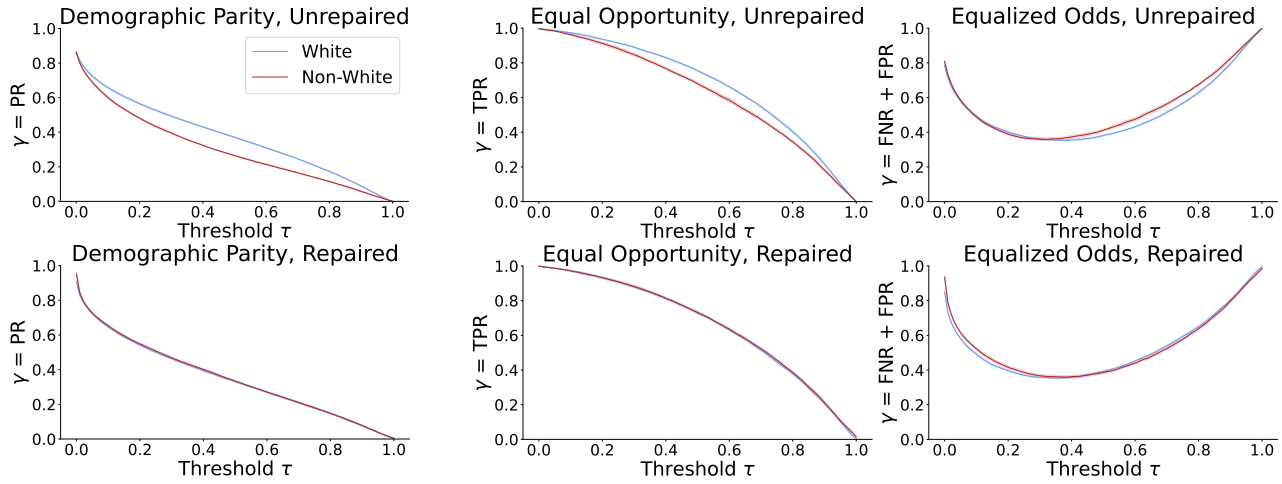
Figure 9: Comparing unrepaired and repaired logistic regression on New `Adult` Income-Race (Ding et al. 2021).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

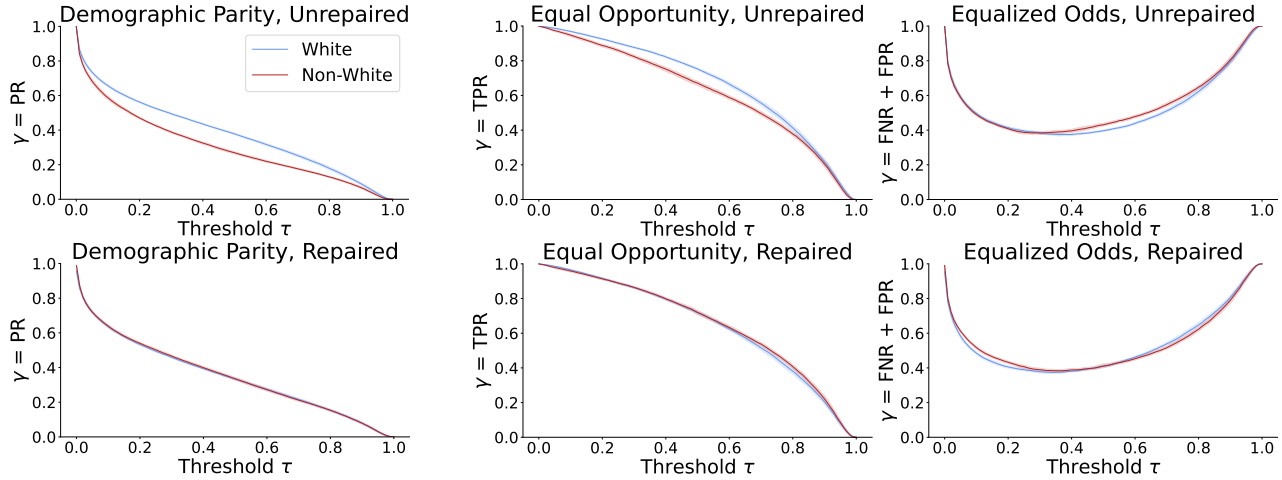
Figure 10: Comparing unrepaired and repaired SVMs on New `Adult` Income-Race (Ding et al. 2021).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

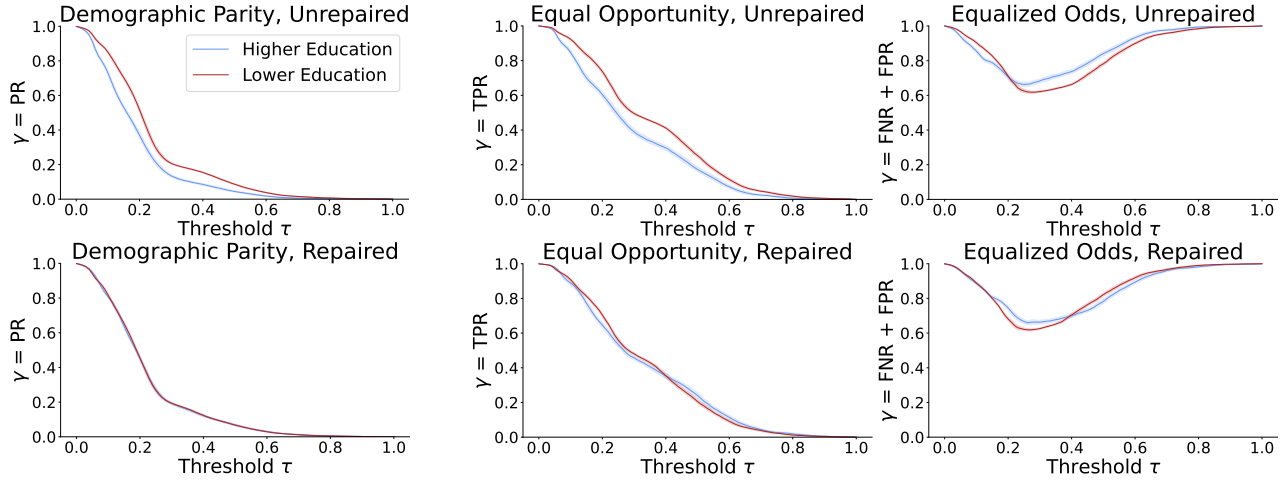
Figure 11: Comparing unrepaired and repaired Random Forests on New `Adult` Income-Race (Ding et al. 2021).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

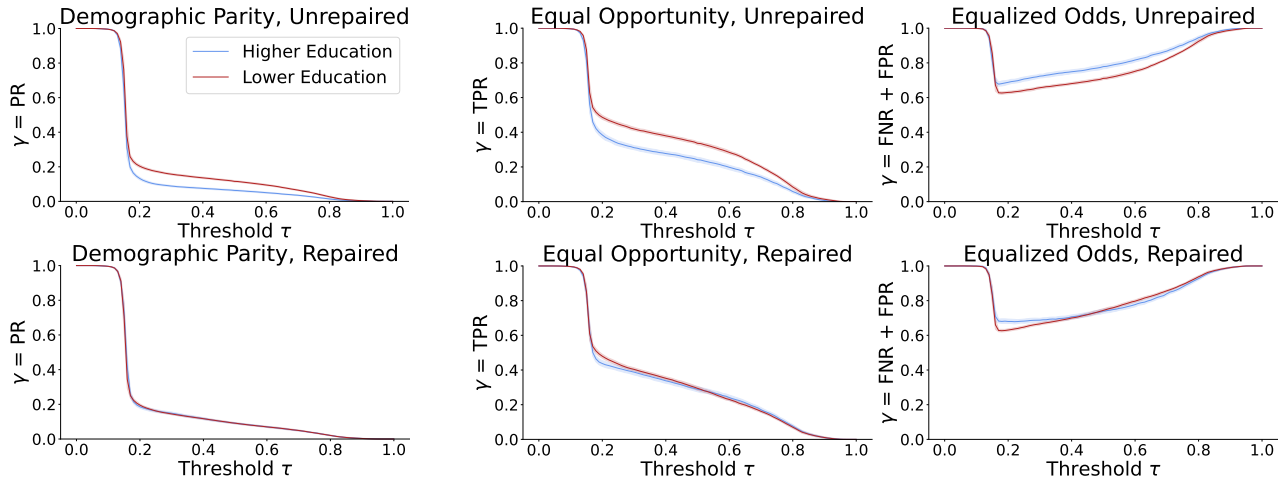
Figure 12: Comparing unrepaired and repaired MLPs on New `Adult` Income-Race (Ding et al. 2021).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

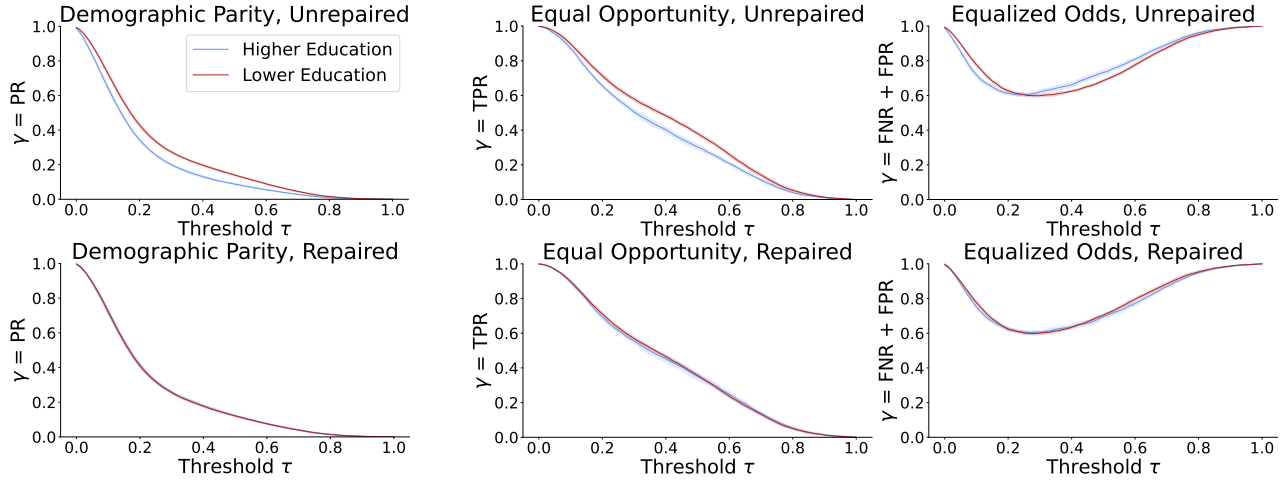
Figure 13: Comparing unrepaired and repaired logistic regression on `Taiwan Credit` (Bay et al. 2000).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

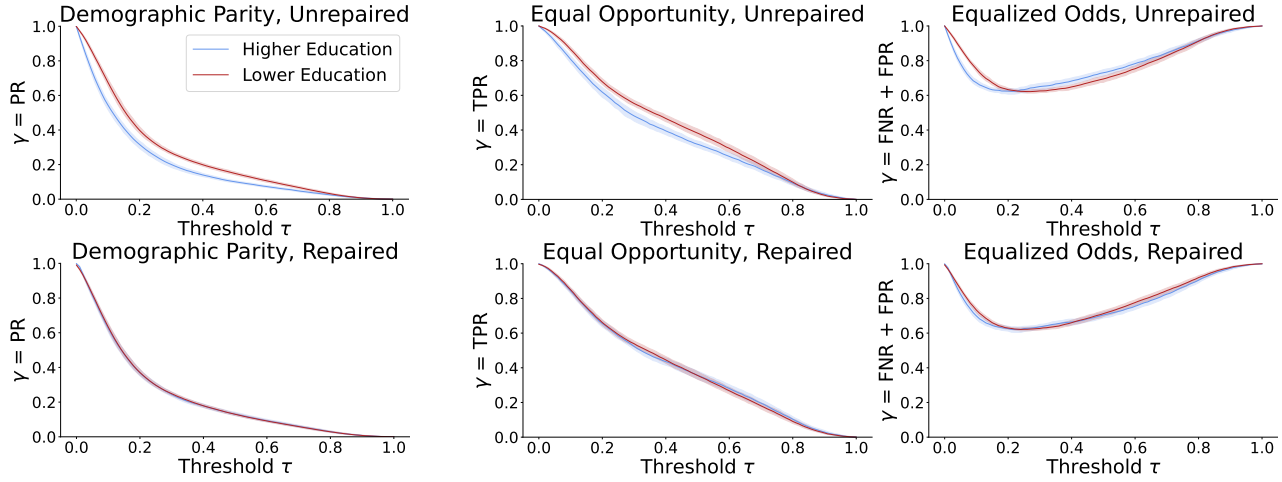
Figure 14: Comparing unrepaired and repaired SVMs on `Taiwan Credit` (Bay et al. 2000).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

Figure 15: Comparing unrepaired and repaired Random Forests on `Taiwan Credit` (Bay et al. 2000).



(a) Algorithm 1 results for SDP

(b) Algorithm 1 results for Eq. Odds and Eq. Opportunity

Figure 16: Comparing unrepaired and repaired MLPs on `Taiwan Credit` (Bay et al. 2000).

References

Adve, A.; and Mészáros, A. 2020. On nonexpansiveness of metric projection operators on Wasserstein spaces.

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

Alla, S.; and Adari, S. K. 2021. What Is MLOps? In *Beginning MLOps with MLFlow: Deploy Models in AWS SageMaker, Google Cloud, and Microsoft Azure*, 79–124. Springer. ISBN 978-1-4842-6549-9.

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

Bay, S. D.; Kibler, D.; Pazzani, M. J.; and Smyth, P. 2000. The UCI KDD archive of large data sets for data mining research and experimentation. *ACM SIGKDD explorations newsletter*, 2(2): 81–85.

Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 13–18.

Chouldechova, A. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:1610.07524.

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2020. Fair regression with Wasserstein barycenters. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 7321–7331. Curran Associates, Inc.

Chzhen, E.; and Schreuder, N. 2022. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*.

Cooper, A. F.; Moss, E.; Laufer, B.; and Nissenbaum, H. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 864–876. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Corbett-Davies, S.; and Goel, S. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. ArXiv preprint.

Cuturi, M.; and Doucet, A. 2014. Fast computation of Wasserstein barycenters. In *International conference on machine learning*, 685–693. PMLR.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *arXiv preprint arXiv:2108.04884*.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 259–268.

Forde, J. Z.; Cooper, A. F.; Kwegyir-Aggrey, K.; De Sa, C.; and Littman, M. 2021. Model Selection's Disparate Impact in Real-World Deep Learning Applications. *arXiv preprint arXiv:2104.00606*.

Frémond, M. 2017. *Collisions Engineering: Theory and Applications*. Springer.

Gordaliza, P.; Barrio, E. D.; Fabrice, G.; and Loubes, J.-M. 2019. Obtaining Fairness using Optimal Transport Theory. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2357–2365. PMLR.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NeurIPS '16, 3323–3331.

Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; and Chiappa, S. 2020. Wasserstein Fair Classification. In Adams, R. P.; and Gogate, V., eds., *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, 862–872. PMLR.

Kallus, N.; and Zhou, A. 2019. The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the xAUC Metric. arXiv:1902.05826.

Kleinberg, J. 2018. Inherent Trade-Offs in Algorithmic Fairness. *SIGMETRICS Perform. Eval. Rev.*, 46(1): 40.

Kroll, J. A. 2021. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 758–771. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Le Gouic, T.; Loubes, J.-M.; and Rigollet, P. 2020. Projection to Fairness in Statistical Learning. ArXiv preprint.

Lipton, Z.; McAuley, J.; and Chouldechova, A. 2018. Does mitigating ML's impact disparity require treatment disparity? In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

McCann, R. J. 1997. A Convexity Principle for Interacting Gases. *Advances in Mathematics*, 128(1): 153–179.

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. *J. Mach. Learn. Res.*, 12: 2825–2830.

Peyré, G.; and Cuturi, M. 2018. Computational Optimal Transport.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On Fairness and Calibration. *Advances in Neural Information Processing Systems*, 30: 5680–5689.

Santambrogio, F. 2015. Optimal Transport for Applied Mathematicians. *Birkäuser, NY*, 55(58-63): 94.

Shankar, S.; Garcia, R.; Hellerstein, J. M.; and Parameswaran, A. G. 2022. Operationalizing Machine Learning: An Interview Study.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; Gummadi, K. P.; and Weller, A. 2017. From Parity to Preference-based Notions of Fairness in Classification. arXiv:1707.00010.