Repairing Regressors for Fair Binary Classification at Any Decision Threshold

Kweku Kwegyir-Aggrey,[1] **Jessica Dai,**[2] **A. Feder Cooper,**[3] **John P. Dickerson,**[4] **Keegan Hines**[4]

[1]Brown University, [2]University of California – Berkeley, [3]Cornell University, [4]Arthur

Abstract

We study the problem of post-processing a supervised machine-learned regressor to maximize fair binary classification at all decision thresholds. By decreasing the statistical distance between each group's score distributions, we show that we can increase fair performance across all thresholds at once, and that we can do so without a large 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. Our main result is to put forward a novel post-processing algorithm based on optimal transport, which provably maximizes Distributional Parity, thereby attaining common notions of group fairness like Equalized Odds or Equal Opportunity at all thresholds. We demonstrate on two fairness benchmarks that our technique works well empirically, while also outperforming and generalizing similar techniques from related work.

1 Introduction

A common approach to fair machine learning is to train a classifier with a chosen decision threshold in order to attain a certain degree of accuracy, and then to post-process 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 popularity of this approach, it suffers from two major limitations. First, it is well-known that the specific choice of decision threshold can influence both fairness and accuracy in practice (Barocas, Hardt, and Narayanan 2019) producing an undesirable trade-off between the two objectives. Second, when deploying a classifier in the real world, practitioners typically need to tinker with the threshold as they evaluate whether a model meets their domain-specific needs (Kallus and Zhou 2019; Chouldechova 2016).

One strategy to address these limitations, to is to develop a procedure that produces regressors that guarantee a selected fairness notion at *all* possible thresholds, while simultaneously preserving accuracy. If a regressor is fair at all thresholds, then a practitioner can freely perform application-specific threshold tuning without ever needing to retrain.

Some prior work has investigated this strategy, by using optimal-transport methods to achieve a single, often trivially satisfied, fairness notion – Demographic Parity – at all thresholds (Jiang et al. 2020; Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020).

However, Hardt, Price, and Srebro (2016) and related impossibility results (Kleinberg 2018; Chouldechova 2016) demonstrate that attaining fairness only at Demographic Parity does *not* capture the nuances in unfairness arising from examining true positive rates, false positive rates, and combinations thereof (Barocas, Hardt, and Narayanan 2019). We therefore ask:

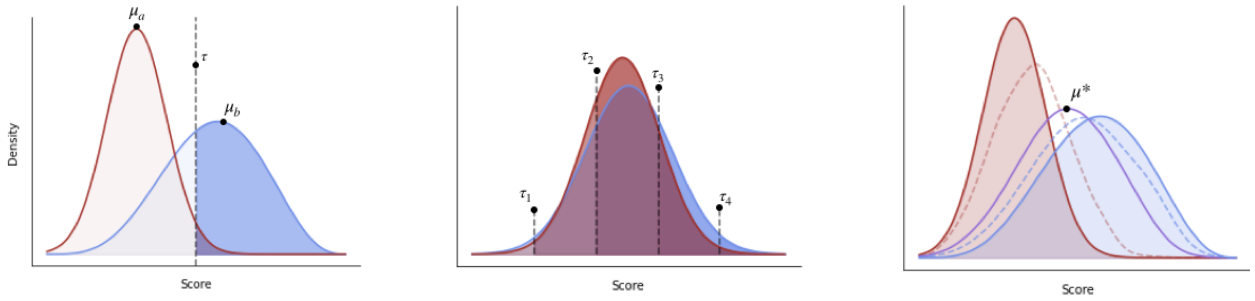
Can we train a regressor once and obtain fair binary classifiers at all thresholds for more flexible group fairness notions?

Our Work. We find that this is indeed possible. Our key insight is to observe that parity in the distributions of a regressor's output for each sensitive group, prior to the application of a threshold, can be harnessed to attain fairness at all thresholds simultaneously. This insight yields the following contributions: **(1)** We introduce a metric called *Distributional Parity* (Definition 3.1) based on the Wasserstein-1 Distance, which enables reasoning about fairness across all thresholds for a wide class of metrics. **(2)** We employ a technique called Geometric Repair (Feldman et al. 2015), which leverages an important connection between Wasserstein-2 barycenters to post-process a regressor under a Distributional Parity constraint, attaining all-threshold fairness. **(3)** We prove that that Distributional Parity is convex on the set of models produced by Geometric Repair, thereby enabling efficient computation of our proposed post processing. Additionally, we show that the models produced by geometric repair are Pareto optimal in the multi-objective optimization of accuracy (via an ℓ_1-type risk) and Distributional Parity. **(4)** Lastly, we synthesize these insights into a novel post-processing algorithm for a broad class of fairness metrics; our algorithm subsumes earlier work on all-threshold Demographic Parity, and we demonstrate its efficacy in experiments on common benchmarks.

2 Background

Let $X \subseteq \mathbb{R}^d$ be a feature space and $G = \{a, b\}$ be a set of binary protected attributes, for which a is the majority group and b is the minority group. We denote the label space as $Y = \{0, 1\}$, where 0 denotes the negative class and 1 the positive class. We assume elements in X, G, and Y are

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(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.

are drawn from some underlying distribution, with corresponding random variables $\boldsymbol{X}$, $\boldsymbol{G}$, and $\boldsymbol{Y}$. The proportion of each group is represented $\rho_g = \Pr[\boldsymbol{G} = g]$. Let $\mathcal{F}$ be a set of measureable *group-aware regressors* (from which binary classifiers are derived). Each regressor $f \in \mathcal{F}$ has signature $f : X \times G \to [0,1]$ and outputs a *score* $s \in [0,1]$ where $s = \Pr[\boldsymbol{Y} = 1 | \boldsymbol{X} = x, \boldsymbol{G} = g]$. For a fixed regressor $f \in \mathcal{F}$ and a decision threshold $\tau \in [0,1]$, we can derive a binary classifier from f by computing $\mathbb{1}\{f \geq \tau\}$ for any $\tau \in [0,1]$. For a group $g \in G$, the *group-conditional score distribution* is the distribution of scores produced by a regressor on that group. We denote this distribution $f(\boldsymbol{X}, \boldsymbol{G}) | \boldsymbol{G} = g$.

For $p \geq 1$, we define $\mathcal{P}_p([0,1])$ be the space of probability measures on $[0,1]$ with finite p-order-moments. We use $\mu_g \in \mathcal{P}_p([0,1])$ to denote the probability measure associated with each group's score distribution. We also make the following (standard) assumption on these measures.

Assumption 2.1. *Any measure $\mu \in \mathcal{P}_p([0,1])$ with finite p-order moments is non-atomic and absolutely continuous with respect to the Lebesgue measure.*

This assumption provides two guarantees. First, it ensures the cumulative distribution function (CDF) of μ_g, denoted $F_{\mu_g}(\tau) = \mu_g([0,\tau))$, has a well defined inverse $F^{-1}_{\mu_g}$. Second, it ensures that certain optimal transport operations, upon which our contributions crucially rely, are well-defined.[1]

2.1 Wasserstein Distance and Wasserstein Barycenters

Before introducing our solution, we present some neccesary background on Wasserstein distance and Wasserstein barycenters. Readers familiar with optimal transport can skim this section.

Wasserstein Distance. Informally, the Wasserstein distance captures the difference between probability measures by measuring the *cost* of transforming one probability measure into the other. In the special case when distributions are univariate, the Wasserstein distance has a nice closed form.

Definition 2.1 (Wasserstein Distance)**.** *For two measures $\mu_1, \mu_2 \in \mathcal{P}_p([0,1])$*

$$\mathcal{W}_p^p(\mu_1, \mu_2) = \int_{[0,1]} |F^{-1}_{\mu_1}(q) - F^{-1}_{\mu_2}(q)|^p dq. \quad (1)$$

We can also define the Wasserstein distance using transport plans; this is commonly referred to as Monge's Formulation. A transport plan is a function $T \in \mathcal{T}$ where every function in $\mathcal{T}$ satisfies standard pushforward constraints, i.e, $T_{\#}\mu_1 = \mu_2$ such that $\mu_2(B) = \mu_1(T^{-1}(B))$ for all measurable $B \subseteq [0,1]$.

Definition 2.2 (Wasserstein Distance [Monge])**.**

$$\mathcal{W}_p^p(\mu_1, \mu_2) = \inf_{T \in \mathcal{T}} \int_{[0,1]} |q - T(q)|^p d\mu_1(q). \quad (2)$$

In our specific case where Assumption 2.1 is satisfied, we know that these transport plans which solve Monge's formulation exist and we can define them in closed form.

Remark 2.1. *The transport plan from $\mu_1 \to \mu_2$ which minimizes Eq. (2) is defined $T_1^2(x) = F^{-1}_{\mu_2}(F_{\mu_1}(x))$ for all $p \geq 1$ (Santambrogio 2015, Remark 2.6)*

Wasserstein Barycenter. The Wasserstein barycenter 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.

Definition 2.3 (Wasserstein Barycenter)**.** *For two measures $\mu_1, \mu_2 \in \mathcal{P}_p([0,1])$ their α-weighted Wasserstein barycenter[2] is denoted μ_α and is computed*

$$\mu_\alpha \leftarrow \operatorname*{arg\,min}_{\nu \in \mathcal{P}_p([0,1])} (1-\alpha)\mathcal{W}_p^p(\mu_1, \nu) + \alpha \mathcal{W}_p^p(\mu_2, \nu), \quad (3)$$

and in the special case when $\alpha = \rho_b$ we write μ^*.

To complete the weighted-average analogy, α behaves like a tunable knob: As $\alpha \to 0$ then μ_α will appear more like μ_1, and as $\alpha \to 1$ the more μ_α will appear like μ_2. As a consequence of this definition and Remark 2.1, we can express the transport plan to a barycenter in closed form, as well:

[1] In general we use $p = 2$. Under this assumption, we occasionally slightly abuse nomenclature, using "measure" and "distribution" interchangeably. In this well-behaved setting, the difference between a probability measure and its probability density function (as guaranteed by uniformity with respect to the Lebesgue measure) is a benign subtlety.

[2] Although defined by Agueh and Carlier (2011) for $p = 2$, this definition for all $p \geq 1$ is widely accepted.

Corollary 2.1. *Let μ_* be the ρ_b-weighted barycenter of μ_a, μ_b then the transport plan from $\mu_a \to \mu_*$ (wlog) is computed*

$$T_a^*(\omega) = (\rho_a F_{\mu_a}^{-1} + \rho_b F_{\mu_b}^{-1}) \circ F_{\mu_a}(\omega)$$

A Note on Our Use of $\mathcal{W}_1$ and $\mathcal{W}_2$. In this work, we make use of both $\mathcal{W}_1$ and $\mathcal{W}_2$. Our use of $\mathcal{W}_1$ is restricted to Distributional Parity computations (see Section 3). This choice is motivated by the fact when $\gamma = \mathcal{U}_{\text{PR}}$, the Wasserstein-1 distance recovers $\mathcal{U}_{\text{PR}}$. We use $\mathcal{W}_2$ to compute Wasserstein barycenters. Given that $\mathcal{W}_2$ is known to be strictly convex, and provided that some μ_g is non-atomic, for $p = 2$ the barycenter that minimizes Eq. 2.3 is unique (Agueh and Carlier 2011, Proposition 3.5).

3 A Distributional View of Fairness

Our goal is to post-process a regressor such that all binary classifiers derived from thresholding this regressor are group fair, i.e., attain fairness in the regressor at every threshold. To attain fairness at every threshold, we look to create parity in outcomes at the level of the regressor – before thresholds are applied – rather than at the level of the derived predictor. The intuition is simple: if a regressor outputs similar scores for two groups, then no matter what threshold is selected, the output derived predictor will be fair; this is illustrated in Figure 1. Specifically, we show that fairness can be attained at all thresholds by enforcing parity in the *distribution* of scores output by a regressor on some groups.

At the core of our new distributional definition of fairness are familiar metrics, namely: Positive Rate (PR), True Positive Rate (TPR), and False Positive Rate (FPR). From these metrics, we can obtain 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). These metrics are formally defined in Table 1.

Let the set of these metrics be $\Gamma = \{\text{PR}, \text{TPR}, \text{FPR}\}$ and any arbitrary metric be $\gamma \in \Gamma$. We write $\gamma_g(\tau; f)$ to denote the rate γ on group g at threshold τ for a score distribution produced by f. When obvious from context, we omit f from this γ notation, writing only $\gamma_g(\tau)$. Additionally, as we show via Corollary 5.1, we can combine these metrics additively, e.g., producing Equalized Odds which combines TPR and FPR.

At a single threshold, (un)fairness is commonly measured by taking the difference in some metric across groups — e.g., for the case of Demographic Parity where $\gamma = \text{PR}$, we can measure fairness by simply computing $|\text{PR}_a(\tau) - \text{PR}_b(\tau)|$.

A natural way to leverage these single-threshold measurements into an all-threshold measurement is to take their average across every possible τ. We formalize this idea in the following definition of *Distributional Parity*.

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

Table 1: The fairness metrics we consider; f is some regressor and $\tau \in [0, 1]$ is a decision threshold.

Definition 3.1 (Distributional parity)**.**

Let $U([0,1])$ be the uniform distribution on $[0,1]$. For a fairness metric $\gamma \in \Gamma$, a regressor f satisfies Distributional Parity *denoted $\mathcal{U}_\gamma(f) \triangleq \mathbb{E}_{\tau \sim U([0,1])} |\gamma_a(\tau) - \gamma_b(\tau)|$, when $\mathcal{U}_\gamma(f) = 0$.*[a]

[a] While ostensibly other priors can be considered, the Wasserstein Distance correspondence only holds for the uniform prior.

A useful property of this definition is that when $\gamma = \text{PR}$, this definition is closely related to the Wasserstein Distance, a distance which is frequently used to measure distance between probability distributions.

Proposition 3.1. *For $\mu_a, \mu_b \in \mathcal{P}_2([0,1])$ which are the groupwise score distributions of f, then $\mathcal{W}_2(\mu_a, \mu_b) = 0$ if and only if $\mathcal{U}_{\text{PR}}(f) = 0$.*

It is from this property that distributional parity is named. At its core, distributional parity is a way to quantify differences between *outcome* distributions – specifically the groupwise score distributions of f. This relationship between distributional parity – an all threshold fairness metric – and the Wasserstein distance – a measure of statistical distance – anchors our proposed shift in focus from thresholds to distributions. Next, we introduce our proposed post-processing objective for computing fair regressors under a distributional parity constraint. We will also begin to outline how we efficiently compute this post-processing, and how our solution elegantly addresses fairness-accuracy trade-off concerns.

3.1 Distributionally Fair Post-Processing

Our goal is to post-process a learned regressor f, such that it becomes (distributionally) fair while remaining accurate. The risk of some other regressor $\hat{f}$ (with respect to f) is computed

$$\mathcal{R}(\hat{f}) = \|\hat{f} - f\|_1 = \mathbb{E}\,|\hat{f}(\boldsymbol{X}, \boldsymbol{G}) - f(\boldsymbol{X}, \boldsymbol{G})|$$

Using this definition of risk, a simple fair post-processing objective can be written as follows,

$$\arg\inf_{\hat{f} \in \mathcal{F}} \mathcal{R}(\hat{f}) \quad \text{s.t.} \quad \mathcal{U}_\gamma(\hat{f}) \leq c, \tag{4}$$

where c is some small constant.

The special case of PR. In the special case where $\gamma = \text{PR}$ and $c = 0$, the solution to Eq. 4 can be computed using a solution based on optimal transport (Jiang et al. 2020). In

this solution, a learned regressor f is transformed into a new regressor we call f^* which provably minimizes risk (with respect to f) while attaining distributional parity for $\gamma = \text{PR}$, i.e, demographic parity at every threshold. This all threshold guarantee is attained with minimal impact to risk. It was shown in (Le Gouic, Loubes, and Rigollet 2020; Chzhen et al. 2020) that f^* is the regressor which increases risk the *least* amongst all regressors which satisfy all threshold demographic parity constraints.[3]

$$f^* \leftarrow \arg\min \mathcal{R}(\cdot) \quad \text{s.t.} \quad \mathcal{U}_{\text{PR}}(f^*) = 0. \tag{5}$$

This solution is strict – it enforces exact demographic parity, which may not always be desired (Chzhen and Schreuder 2022). We can address this concern by considering a relaxation of Eq. 4 which uses a parameter λ to balance the a trade-off between fairness and accuracy. Specifically, for every $\lambda \in [0, 1]$ there is some $f_\lambda \in \mathcal{F}$ which attains λ-increase in the fairness, in exchange for a λ-reduction in risk. We prove the existence of f_λ which satisfies this property in the following lemma .

Lemma 3.1. *Let f be some learned regressor. For all $\lambda \in [0, 1]$ the set of optimally fair regressors for λ-relaxations of f with respect to risk and distributional parity for $\gamma = PR$ are given by*

$$f_\lambda \leftarrow \arg\min_{\hat{f} \in \mathcal{F}} \lambda R(\hat{f}) \quad \textit{s.t.} \quad \mathcal{U}_{PR}(\hat{f}) = (1-\lambda)\mathcal{U}_{PR}(f) \tag{6}$$

The functions f_λ are Pareto-optimal: indeed, we show in Theorem 4.2 that all $\{f_\lambda\}_{\lambda \in [0,1]}$ are Pareto optimal in the multi-objective minimization of $\mathcal{R}$ and $\mathcal{U}_{\text{PR}}$. This means we can view these regressors as being optimally accuracy preserving, while also being fair. As a result, the above optimization (with $\gamma = \text{PR}$) can be rewritten simply as

$$\arg \min_{\lambda \in [0,1]} \mathcal{U}_\gamma(f_\lambda), \tag{7}$$

replacing the risk minimization objective with an objective that enforces distributional parity, given the aforementioned accuracy-preserving properties of f_λ.

Extending to other fairness metrics. The above approach to achieving all-threshold fairness has two steps: firstly, a characterization of a solution that achieves ideal fairness, and then the construction of a space of functions f_λ that allow for an optimal fairness-accuracy tradeoff. In order to generalize this to other fairness measures, we need version of both steps. The exact result for PR however relies heavily on optimal transport in a way that does not naturally generalize to other fairness measures $\gamma \neq \text{PR}$.

To address this, we provide two key insights. Firstly, that the f_λ can be expressed explicitly using optimal transport ideas in terms of score distributions that are independent of the choice of fairness measure, and secondly, that the optimization described in equation (7) describes a convex function of λ *independent of the choice of* γ.

[3]These results were proven for an ℓ_2 risk; we elide this distinction for ease of exposition.

This means that for any choice of γ, we can find the optimal value of λ to minimize $\mathcal{U}_\gamma(f_\lambda)$. And we will show empirically that this choice yields an almost perfect minimization of distributional parity, thus achieving an all-threshold fairness result as desired.

4 Maximizing Distributional Parity with Geometric Repair

We now introduce the method and main theorem that we use to compute distributionally fair post processing. The method, defined in Section 4.1 is called Geometric Repair (Feldman et al. 2015; Chzhen and Schreuder 2022), and is how we efficiently compute solutions to the objective stated in Equation 7. Our main theoretical result is stated in Theorem 4.1. Subsequently, we show in subsection 4.2 thatwe can make use of an elegant transformation to an optimal transport problem in order to achieve approximate distributional parity for all γ.

4.1 Defining Geometric Repair

Geometric repair is a technique for constructing a regressor that interpolates between the output of some learned regressor f (assumed to be accurate), and the output of a certain fair function f^*. Note, f^* must be specifically chosen in order to prove our results, but for ease of exposition, we defer formal definition of f^* to the following subsection.

Definition 4.1 (Geometric Repair)**.** *We call $\lambda \in [0, 1]$ the* repair parameter *and define a geometrically* repaired *regressor f_λ as*

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

Geometric repair enumerates a well structured set of regressors which achieve λ-relaxations of R and $\mathcal{U}_{\text{PR}}$ as described in Section 3.

Proposition 4.1. *For any $\lambda \in [0, 1]$, a repaired regressor f_λ satisfies the following*

$$R(f_\lambda) = \lambda R(f^*) \quad \textit{and} \quad \mathcal{U}_{PR}(f_\lambda) = (1-\lambda)\mathcal{U}_{PR}(f)$$

This is the set of regressors used to maximize distributional parity. The key to computing such a maximization lies in the following theorem, which shows that distributional parity is convex, on the set of repaired regressors. This convexity guarantee certifies our ability to locate the f_λ, amongst the set of repaired regressors, which *best* minimizes Distributional Parity for *any* γ.

Theorem 4.1. *Fix $\gamma \in \Gamma$. Let $f : X \times G \to [0, 1]$ be a regressor, and f_λ be the geometrically repaired regressor for any $\lambda \in [0, 1]$. The map $\lambda \mapsto \mathcal{U}_\gamma(f_\lambda)$ is convex in λ.*

The proof of this theorem crucially depends on the connection between f^* and Wasserstein barycenters. In the next section we, leverage this connection to analytically compute the distributions of f_λ, which is a crucial piece needed in proving the convexity of $\mathcal{U}_\gamma(f_\lambda)$.

4.2 How f^* Enables Geometric Repair

Here, we formalize the earlier definition of f^* from Section 4.2 and its connection between to Wasserstein barycenters in the context of geometric repair.

Definition 4.2 (Fully Repaired Regressor)**.** *The regressor f^* which satisfies distributional parity for $\gamma = PR$ while minimizing risk (with respect to f) is the computed*

$$f^* \leftarrow \arg\min_{f \in \mathcal{F}} \mathcal{R}(\cdot) \quad s.t. \quad \mathcal{U}_{PR}(f) = 0. \tag{8}$$

We call this regressor fully repaired *in that $f_{\lambda=1}$ is equivalent to f^*.*

The aforementioned property which relates f^* to $\mathcal{W}_2$ barycenters is the the fairness constraint in Eq. (8). To make this clear, recall Proposition 3.1 which states that removing the $\mathcal{W}_2$ distance between distributions is sufficient to satisfy distributional parity for $\gamma = \text{PR}$. The tool we will use to remove this distance is, indeed, Wasserstein barycenters. Prior work (Le Gouic and Loubes 2017; Chzhen et al. 2020) show that mapping μ_a, μ_b onto their ρ_b-weighted barycenter distribution, which we denote μ_*, removes the Wasserstein distance between μ_a, μ_b under this mapping, thereby satisfying $\mathcal{U}_{\text{PR}}(f^*) = 0$ and establishing that f^* is distributed like μ_*.

We can use this fact to rewrite the score distributions of each group under geometric repair. The following proposition formalizes this claim, by showing that the groupwise distributions of output by any f_λ can be computed as barycenters of μ_g and μ_*.

Proposition 4.2. *Let $\lambda \in [0,1]$. Let $\mu_{g,\lambda}$ be the λ-weighted barycenter between μ_g and μ_*, i.e.,*

$$\mu_{g,\lambda} \leftarrow \arg\min_{\nu \in \mathcal{P}_2([0,1])} (1-\lambda)\mathcal{W}_2^2(\mu_g, \nu) + \lambda \mathcal{W}_2^2(\mu_*, \nu),$$

then $\mu_{g,\lambda} = Law(f_\lambda(\boldsymbol{X}, \boldsymbol{G}) | \boldsymbol{G} = g)$.

This proposition shows us that the interpolation between f and f^* as parametrized by λ in geometric repair is replicated at the distributional level, i.e., λ also controls the interpolation from $\mu_{g,\lambda} \to \mu_*$; more importantly, the intermediate distributions of this interpolation have a special structure – they are barycenters. Note that under Assumption 2.1 and (Agueh and Carlier 2011, Proposition 3.5), these $\mu_{g,\lambda}$ are unique and guaranteed to exist. For clarity, we visualize this interpolation (over distributions) in Figure 2.

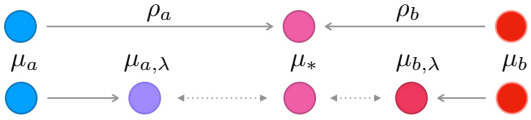


Figure 2: Let μ_a, μ_g be groupwise score distributions. We illustrate of the repaired score distributions $\mu_{g,\lambda}$ under geometric repair, where μ_* is the ρ_b-weighted barycenter.

Our proof of Theorem 4.1 computes distributional parity as a function of the score distributions of f_λ. With these established via Proposition 4.2, we are able to write a closed form expression for $\mathcal{U}_\gamma(f_\lambda)$. Using this expression, the proof proceeds computationally, showing that the second derivative of $\mathcal{U}_\gamma(f_\lambda)$ is non-negative to conclude convexity.

4.3 The Optimality of Geometric Repair in Balancing Fairness and Accuracy

Now, we will show that f_λ is optimal in the fairness-accuracy trade-off with respect to $\gamma = \text{PR}$.

Definition 4.3 (Pareto Optimality)**.** *For $f, f' \in \mathcal{F}$ we say f Pareto dominates f', denoted $f' \prec f$, if one of the following hold:*

$$\mathcal{R}(f) \le \mathcal{R}(f') \quad \mathcal{U}_{PR}(f) < \mathcal{U}_{PR}(f') \tag{9}$$

$$\mathcal{R}(f) < \mathcal{R}(f') \quad \mathcal{U}_{PR}(f) \le \mathcal{U}_{PR}(f') \tag{10}$$

A regressor f is Pareto optimal if there is no other regressor f' that has improved risk without also having strictly more unfairness, or vice-versa.

The proof of Pareto optimality of f_λ follows from Proposition 4.1. The main idea of this result is the following: f^* is the lowest risk classifier where $\mathcal{U}_{\text{PR}}(\cdot) = 0$ meaning that it is Pareto optimal by construction. Since f_λ is a λ-relaxation of f^* with regards to both risk and unfairness, f_λ preserves the Pareto optimality of f^*.

Theorem 4.2. *For all $\lambda \in [0,1]$, the repaired regressor f_λ is pareto optimal in the multi-objective minimization of $\mathcal{R}(\cdot)$ and $\mathcal{U}_{PR}(\cdot)$.*

5 Post-Processing Algorithms to Maximize Distributional Parity

Now that we've supported *why* we can use geometric repair to maximize distributional parity, we provide some practical algorithms showing *how* to do so. First, we'll show how to estimate f_λ from samples.

Plug-in Estimator for f_λ. Indeed, computation of f^*, and therefore μ_*, requires exact knowledge of μ_a, μ_b. In practice we only have sample access to both score distributions, and so we must approximate these distributions, and consequently their barycenter and f_λ. We show a plug-in estimator w/ the following convergence guarantee (Theorem 5.1) to approximate f_λ in Algorithm 1. Our approach to approximating f_λ only requires a input regressor f and access to some unlabeled dataset $D = (x_1, g_1) \dots (x_n, x_g)$. Let n_g denote the number of samples from a group g.

Theorem 5.1. *As $n_g \to \infty$ the empirical distribution of $\hat{f}_\lambda(x, g)$ converges to to $\mu_{g,\lambda}$ in $\mathcal{W}_2$ almost surely.*

Post-Processing to Maximize Distributional Parity. To actually compute the optimal λ_* for some metric, we propose the post-processing routine described in Algorithm 2. The algorithm consists of two main steps: approximating $\hat{f}_\lambda$ in Step 1, and finding the optimal λ_* in Step 2. Note that our objective $\hat{\mathcal{U}}_\gamma(f_\lambda)$ is parametrized by the scalar λ, and so we find its minima using a univariate solver; we found success using Brent's Method (Brent 2013). By the convexity of $\mathcal{U}_\gamma(\cdot)$ as proven in Theorem 4.1, we are guaranteed that the f_{λ_*} is optimal on the set of repaired regressors.

Corollary 5.1. *Since convex functions are closed under addition, Theorem 4.1 also applies to additive combinations of metrics, meaning that the objective in Step (2) of Alg 2 can be replaced by $\mathcal{U}_{\gamma_1}(f_\lambda) + \mathcal{U}_{\gamma_2}(f_\lambda) + ... + \mathcal{U}_{\gamma_m}(f_\lambda)$.*

Algorithm 1: An Estimator for f_λ

Input: A regressor f, and an unlabeled dataset $D = (x_1, g_1) \dots (x_n, g_n)$

1. Let $n_g = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{g_k=g}$. Use f to approximate the group-conditional distributions

$$\hat{\mu_g} = \frac{1}{n_g}\sum_{i=1} \delta_{f(x_i,g_i)}\mathbb{1}_{g_i=g}$$

2. Let $\hat{\rho}_g = \frac{n_g}{n}$ and compute the empirical optimal transport plans (see Remark 2.1)

$$\hat{T}_g^*(\omega) = (\hat{\rho}_a F_{\hat{\mu}_a}^{-1} + \hat{\rho}_b F_{\hat{\mu}_b}^{-1}) \circ F_{\hat{\mu}_a}(\omega)$$

3. For any $\lambda \in [0,1]$, compute $\hat{f}_\lambda$ where $\hat{f}_\lambda(x,g) = (1-\lambda)f(x,g) + \lambda\hat{T}_g^*(f(x,g))$

Algorithm 2: Post-Processing for Distributional Parity

Input: A metric $\gamma \in \Gamma$, learned regressor f, and labeled dataset $E = (x_1, g_1, y_1) \dots ...(x_k, g_k, y_k)$

1. Using Algorithm (1) to approximate f_λ by computing $\hat{T}_g$ such that for all $\lambda \in [0,1]$ geometric repair is well defined, i.e., $\hat{f}_\lambda(x,g) = (1-\lambda)f(x,g) + \lambda\hat{T}_g(f(x,g))$
2. Use Brent's algorithm to find the optimal λ which minimizes $\lambda_* \leftarrow Brent_{\lambda\in[0,1]}\hat{\mathcal{U}}_\gamma(f_\lambda)$ where $\hat{\mathcal{U}}(f_\lambda)$ is approximated for m randomly sampled $(\tau_1 ....\tau_m) \sim U([0,1])$ via

$$\hat{\mathcal{U}}(f_\lambda) = \frac{1}{m}\sum_{\ell=1}^{m} |\gamma_a(\tau_\ell; f_\lambda) - \gamma_b(\tau_\ell; f_\lambda)|.$$

3. **Output:** $f_{\lambda_*}(x,g)$ such that $\hat{\mathcal{U}}_\gamma(f_{\lambda_*})$ is minimized (distributional parity is maximized)

6 Experiments

In this section we present experiments that demonstrate the effectiveness of our proposed algorithms in Section 5. To that end, we provide two sets of results:

- Figure 3 validates that Algorithm 2 achieves almost-exact distributional parity for Demographic Parity, Equal Opportunity, and Equalized Odds.
- Table 2 shows that Algorithm 2 outperforms related methods in maximizing Distributional parity while preserving accuracy.

Datasets. We use two datasets: `Adult` Income-Sex from the the UCI repository (Dua and Graff 2017), and `Adult` Income-Race from the 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. In the Income-Sex dataset, men comprise 66.9% of the dataset, 30.1% of whom have label 1; by contrast, 10.9% of the women have label 1. In Income-Race, 61.8% of individuals are identified as white, where 44.3% of whom have label 1; for individuals identified as non-white, 35.6% have label 1.

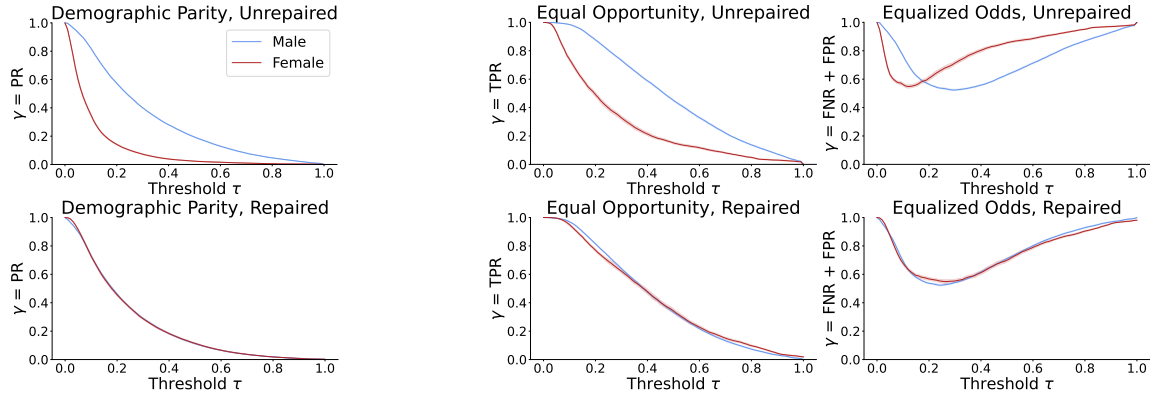
Model Training. To produce a model that we use in our experimentation, we implemented a Logistic Regression (**LR**) with ℓ_2 regularization, and an Support Vector Machine (**SVM**) with an Radial Basis Function kernel. Both were implemented using using scikit-learn with its default model parameters and optimizers (Pedregosa et al. 2011). We show results across 10 different training runs, each run using a different random seed for model initialization, and train/test/-validation splitting of the data.

Metrics. We use the following measurements of model performance: (1) We approximate **Distributional parity** $\mathcal{U}_\gamma(\cdot)$ as per Step 2 of Algorithm 2 using $m = 100$ randomly sampled thresholds. We denote the Equalized Odds metric EO = FNR + FPR, i.e., the misclassification rate(2) We measure accuracy using the **Area Under the Curve (AUC)** given that AUC averages model performance across all thresholds similar to $\mathcal{U}_\gamma$. (3) **Worst Case** refers to the worst disparity of the regressor at any threshold for the chosen γ, i.e., $\max_{\tau\in[0,1]} |\gamma_a(\tau) - \gamma_b(\tau)|$. In both Table 2 and Figure 3, each metric is averaged across 10 trials. We report the mean and standard deviation in the table.

Baselines. We use the following algorithms as baselines to compare our results: (**OG**) The output of the learned classifier with no additional processing. (**JIA**) The post-processing algorithm proposed by Jiang et al. (Jiang et al. 2020) which processes the output of a regressor such that continuous model output is independent of protected group (shown to be equal to satisfying $\mathcal{U}_{\text{PR}} = 0$, which is achieved by our method for $\lambda = 1$). (**FEL**) Pre-Processing of model inputs from Feldman et al. (Feldman et al. 2015) which seeks to reduce disparate impact across all thresholds. The "amount" of pre-processing is parametrized by a λ similar to ours (just over inputs) – we search for the optimal λ for each metric we compare against. We abbreviate geometric repair with (**GR**).

Procedure. For each experiment, we split each dataset into three equal parts: (1) training data; (2) validation data used for finding the optimal λ_* using Algorithm 2; (3) testing data for measuring the distributional parity and the other metrics we consider. We performed our experiment 10 times with different random seeds each trial. These seeds were used for model initialization and for the train/validation/testing splitting of the data. Binary classification were produced from model outputs using $\tau \in \{0, 0.01, 0.02, ..., 0.99, 1\}$.

Results. In Table 2 we show the effectiveness of our approach against several baselines. As denoted by the bolded



(a) Unrepaired vs. Full Repair

(b) Unrepaired vs. Optimal Repair

Figure 3: Performing geometric repair for $\gamma = \text{PR}$ (left), TPR(middle), EO (right) for Logistic Regression trained on `Adult` Income-Race. The top row depicts the rates for unrepaired regressors and the bottom row for the repaired regressor.

		TPR Optimized			EO Optimized		
		$\mathcal{U}_{\text{TPR}}$	Worst case	AUC	$\mathcal{U}_{\text{EO}}$	Worst case	AUC
Income-Race (LR)	GR	$\mathbf{0.025 \pm 0.013}$	$\mathbf{0.068 \pm 0.034}$	*0.816 ± 0.004*	$\mathbf{0.021 \pm 0.007}$	$\mathbf{0.055 \pm 0.02}$	*0.816 ± 0.005*
	JIA	0.028 ± 0.007	0.129 ± 0.021	0.8 ± 0.005	0.049 ± 0.006	0.094 ± 0.014	0.8 ± 0.005
	FEL	0.042 ± 0.039	0.106 ± 0.051	0.81 ± 0.015	0.103 ± 0.039	0.213 ± 0.081	0.811 ± 0.014
	OG	0.219 ± 0.01	0.430 ± 0.024	0.834 ± 0.005	0.142 ± 0.007	0.299 ± 0.016	0.834 ± 0.005
Income-Sex (SVM)	GR	$\mathbf{0.014 \pm 0.005}$	$\mathbf{0.042 \pm 0.015}$	*0.882 ± 0.004*	0.032 ± 0.006	0.079 ± 0.015	*0.882 ± 0.004*
	JIA	0.052 ± 0.009	0.104 ± 0.013	0.878 ± 0.004	0.047 ± 0.006	0.110 ± 0.011	0.878 ± 0.004
	FEL	0.014 ± 0.007	0.08 ± 0.015	0.769 ± 0.007	$\mathbf{0.026 \pm 0.006}$	0.081 ± 0.013	0.769 ± 0.007
	OG	0.065 ± 0.01	0.114 ± 0.015	0.884 ± 0.003	0.035 ± 0.008	$\mathbf{0.077 \pm 0.013}$	0.884 ± 0.003

Table 2: Comparison of Geometric Repair (**GR**) against included baselines (abbreviations described under **baselines**). Results are averaged over ten trials, and the mean and standard deviation across all trials are reported for each metric.

cells in the $\mathcal{U}_\gamma$ and *Worst Case* columns, our method outperforms almost all baselines on both the `Adult` Income-Sex and `Adult` Income-Race tasks datasets, for both TPR and EO. The one exception is for $\gamma = \text{EO}$ on the Income-Sex task, however our method still attains a reduction in all-threshold disparity, and preserves significant accuracy. For the AUC column, we italicize the cell which has AUC closest to that of the original regressor; for both metrics and datasets, our method was superior to the baselines in this aspect. We show illustrate the effect of geometric repair at every threshold in Figure 3. For the For $\gamma = \text{PR}$ (left) we show the *full* correction $\lambda = 1$. For $\gamma = \text{TPR}$ (middle) we the computed optimal repair parameter $\lambda_* \approx 0.73 \pm 0.04$, and for $\gamma = \text{EO}$ (right) we computed $\lambda_* \approx 0.75 \pm 0.03$.

7 Related Work

A number of prior works have demonstrated how to achieve exact distributional parity in the special case when $\gamma = \text{PR}$. Our work is most closely related to (Jiang et al. 2020) who accomplish this using the $\mathcal{W}_1$ distance, in both in/post processing settings. Chzhen et al. (2020); Le Gouic, Loubes, and Rigollet (2020) report a similar post-processing result to ours, deriving an optimal fair predictor (also limited to $\gamma = \text{PR}$) in a regression setting and using $\mathcal{W}_2$ barycenters. We build on these approaches by extending them to a broader class of fairness metrics and definitions. Our technique is based on the *geometric repair* algorithm which was as originally introduced by (Feldman et al. 2015) as a way to navigate the fairness-accuracy trade-off. Geometric repair was also studied by (Gordaliza et al. 2019). In the post-processing setting, the effect of geometric repair on classifier accuracy and $\gamma = \text{PR}$ fairness was studied in (Chzhen and Schreuder 2022) – we extend these to all $\gamma \in \Gamma$ by showing convexity on the set of regressors enumerated by geometric repair.

8 Conclusion and Future Work

In this work, we show that by interpolating between group-conditional score distributions we can achieve all-threshold fairness on fairness metrics like Equal Opportunity and Equalized Odds. To this end, we introduce Distributional parity to measures parity in a fairness metric 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 benchmark datasets and tasks. In future work, we hope to position this work in context with other fairness metrics like calibration (Hébert-Johnson et al. 2018) or individual fairness, examine robustness of all-threshold fairness to variance in the underlying regressor (Cooper et al. 2023; Forde et al. 2021), explore the rate of convergence of our estimator of f_λ.

A Additional Background on Optimal Transport

In this section of the appendix, we present some additional background and theory from Optimal Transport. These results are necessary to prove some of the results in the main paper body.

A.1 Wasserstein Geodesics

One key property of Wasserstein Barycenters that we exploit in this work, which is not refereed to explicitly in the main paper body, is that Wasserstein Barycenters under Geometric Repair form a curve in the space of probability measures called a constant speed geodesic.

Definition A.1 (Santambrogio (2015), pg. 182)**.** *Let (X, d) be some metric space. A curve $\omega : [0, 1] \to X$ is a* constant speed geodesic *between $\omega(0)$ and $\omega(1)$ if it satisfies*

$$d(\omega(t), \omega(s)) = |t - s| d(\omega(0), \omega(1)) \quad \forall t, s \in [0, 1]$$

The following result from Santambrogio (2015, Theorem 5.27) proves that a specific interpolation of an optimal transport plan forms a geodesic in the space of probability measures metricized by the Wasserstein Distance.[4] We also remind the reader that in the following expression, $\#$ denotes the pushforward operator on measures and $\mathtt{id}$ denotes the identity function. [5]

Theorem A.1. *Suppose that Ω is convex, take $\mu, \nu \in \mathcal{P}_p(\Omega)$, and $\gamma \in \Gamma(\mu, \nu)$ an optimal transport plan for the cost $c(x, y) = |x - y|^p$ w/ $(p \geq 1)$. Define $\pi_t : \Omega \times \Omega \to \Omega$ through $\pi_t(x, y) = (1 - t)x + ty$. Then the curve $\mu_t(\pi_t)_{\#}\gamma$ is a constant speed geodesic connecting $\mu_0 = \mu$ to $\mu_1 = \nu$. In the particular case where μ is absolutely continuous then this curve is obtained as $((1 - t)id + tT)_{\#\mu}$*

For $p = 2$, this special form of the interpolation between measures given in the above theorem is actually the exact same interpolation that is carried out by Wasserstein Barycenters.[6]

Proposition A.1 (Agueh and Carlier (2011))**.** *Let $\mu, \nu \in \mathcal{P}_2([0, 1])$ satisfy Assumption 2.1 then α-weighted barycenters*

$$\mu_\alpha \leftarrow \underset{\in \mathcal{P}_p([0,1])}{\arg\min} (1 - \alpha)\mathcal{W}_2^2(\mu, \cdot) + \alpha\mathcal{W}_2^2(\nu, \cdot),$$

can be equivalently computed $((1 - t)id + tT)_{\#\mu}$ where T is the transport plan that solves transport from $\mu \to \nu$.

This means, under our mild assumptions, that barycenters both (a) follow the special form in Theorem A.1 and (b) are constant speed geodesics. We use this fact to show that the distance between λ repaired measures $\mu_{a,\lambda}, \mu_{b,\lambda}$ can be written as a $1 - \lambda$ weighted fraction of the Wasserstein distance of the unrepaired measures μ_a, μ_b.

Proposition A.2. *Since $\mu_{g,\lambda}$ is a constant speed geodesic, the Wasserstein distance between repaired measures is proportional to the repair amount, i.e., $\mathcal{W}_1(\mu_{a,\lambda}, \mu_{b,\lambda}) = (1 - \lambda)\mathcal{W}_1(\mu_a, \mu_b)$.*

Proof. Let $\mu_a, \mu_b \in \mathcal{P}_2$ and T_a^b be the optimal transport plan from $\mu_a \to \mu_b$. Suppose we parametrize the interpolation from μ_a to μ_b with a function $w : [0, 1] \to \mathcal{P}_1([0, 1])$ where $w(\alpha) = ((1 - \alpha)\mathtt{id} + \alpha T_a^b)\#\mu_a$. By Theorem A.1, this curve is a constant speed geodesic. Now, consider the geometric repair score distributions $\mu_{a,\lambda}$ and $\mu_{b,\lambda}$. We see from Proposition A.2 that each distribution $\mu_{g,\lambda}$ is the result of the λ-weighted interpolation of μ_g to the barycenter μ_*. These barycenters can alternatively computed by interpolating from $\mu_a \to \mu_b$, i.e.,

$$\mu_{a,\lambda} = ((1 - \lambda\rho_b)\mathtt{id} + \lambda\rho_b T_a^b)\#\mu_a$$
$$\mu_{b,\lambda} = (\lambda\rho_a \mathtt{id} + (1 - \lambda\rho_a)T_a^b)\#\mu_a.$$

From this, a reparametrization of the above interpolation under geometric using $w(\cdot)$ yields,

$$\mu_{a,\lambda} = w(\lambda\rho_b) \quad \text{and} \quad \mu_{b,\lambda} = w(1 - \lambda\rho_a).$$

Then, the corollary follows from the definition of constant speed geodesics in Definition A.1, i.e.,

$$\begin{aligned}
\mathcal{W}_1(\mu_{a,\lambda}, \mu_{b,\lambda}) &= \mathcal{W}_1(w(\lambda\rho_b), w(1 - \lambda\rho_a)) && (11)\\
&= |\lambda\rho_b - (1 - \lambda\rho_a)|\mathcal{W}_1(\mu_a, \mu_b) && (12)\\
&= |\lambda(\underbrace{\rho_a + \rho_b}_{=1 \text{ by Def}}) - 1|\mathcal{W}_1(\mu_a, \mu_b) && (13)\\
&= (1 - \lambda)\mathcal{W}_1(\mu_a, \mu_b) && (14)
\end{aligned}$$

□

[4] We can metricize $\mathcal{P}_p$ with $\mathcal{W}_p$ under (Villani 2008, Theorem 6.9)

[5] In this section, we use a sub-scripted μ_t to denote a measure that is the result of some interpolation when clear from context; this subscript notation should not to be confused with the sub-scripts used on measures, e.g. μ_2, in other places in the paper.

[6] This result is stated in Agueh and Carlier (2011) as the conclusion of Section 4 (see eq. 4.10) and in Section 6.2 of the same work

The last additional result we'll need to aid our effort to prove Theorem 4.1, is the following Lemma. Please note this lemma differs from the above corollary due to the specific optimal transport problems being solved. In the above, we are consider a parametrization of $\mu_{g,\lambda}$ along the interpolation from $\mu_a \to \mu_b$. In the below result we consider the interpolation of repaired distributions to their barycenter, i.e., $\mu_{a,\lambda} \to \mu_*$.

Lemma A.1. *Let $\mu_a, \mu_b \in \mathcal{P}_2([0,1])$ 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*

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

Proof. Let μ_* be the ρ_b barycenter of μ_a, μ_b. It is easy to show from their definitions that $\mu_{a,\lambda_1} = \mu_{\lambda_1(1-\rho_a)}$ and $\mu_{b,\lambda_2} = \mu_{1-\lambda_2\rho_a}$ (Figure 2 provides a nice illustration of this fact). To prove the Lemma, we let $\lambda_1(1-\rho_a) = 1-\lambda_2\rho_a$. Solving for λ_1, yields the proposition, i.e., $\lambda_1 = \frac{1-\lambda_2\rho_a}{\rho_b}$ and therefore $\mu_{a,\lambda_1} = \mu_{b,\frac{1-\rho_a\lambda_2}{\rho_b}}$. Letting $\lambda_1 = \lambda_2$, such that both μ_a, μ_b are controlled by the same repair parameter yields the first equality. Solving for λ_2 and making the same substitution ($\lambda_2 = \lambda_1$) yields the second equality. □

A.2 The Relationship Between Fair Risk Minimization and Barycenters

In this subsection we give an additional result relating the lowest risk $\gamma =$ PR regressor to the distance of that regressors groupwise score distributions, to their barycenter.

Lemma A.2. *Let $\mathcal{F}_{PR} \subset \mathcal{F}$ be a subset of regressors where $\mathcal{F}_{PR} = \{f \in \mathcal{F} : \mathcal{U}_{PR}(f) = 0\}$. The minimum risk in $\mathcal{F}_{PR}$ is defined*

$$\min_{\hat{f}\in\mathcal{F}_{PR}} \mathcal{R}(\hat{f}) = \min_{\nu\in\mathcal{P}_1([0,1])} \sum_{g\in g} p_g \mathcal{W}_1(\mu_g, \nu)$$

Proof. Suppose h is the regressor which minimizes the l.h.s. and let $\mu_h = Law(h(\boldsymbol{X}, \boldsymbol{G}))$. We can re-write

$$\begin{aligned}\sum_{g\in g} p_g \mathcal{W}_1(\mu_g, \mu_h) &= \sum_{g\in g} p_g \min_{T\in\mathcal{T}_g^h} \int_{[0,1]} |x - T(x)| d\mu_g \\ &= \sum_{g\in g} p_g \min_{T\in\mathcal{T}_g^h} \int_X |f(\boldsymbol{X}, g) - T(f(\boldsymbol{X}, g))| d\mu_{X|g}\end{aligned}$$

Let $T_g^{\hat{h}} = F_{\mu_{\hat{h}}} \circ F_{\mu_g}^{-1}$ be the optimal transport maps which minimize the above, and let $\hat{h}(x, g) = T_g^h(f(x, g))$. We can continue

$$\begin{aligned}\sum_{g\in g} p_g \min_{T\in\mathcal{T}_g^h} \int_X |f(\boldsymbol{X}, g) - \hat{h}(\boldsymbol{X}, g))| d\mu_{X|g} &= \mathop{\mathbb{E}}_{g\sim\boldsymbol{G}}\left[\mathop{\mathbb{E}}_{\boldsymbol{X}}[|f(\boldsymbol{X}, g) - \hat{h}(\boldsymbol{X}, g)|]|\boldsymbol{G} = g\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}\left[|f(\boldsymbol{X}, \boldsymbol{G}) - \hat{h}(\boldsymbol{X}, \boldsymbol{G})|\right].\end{aligned}$$

From the above equalities we've shown,

$$\sum_{g\in g} p_g \mathcal{W}_1(\mu_a, \mu_h) = \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}\left[|f(\boldsymbol{X}, \boldsymbol{G}) - \hat{h}(\boldsymbol{X}, \boldsymbol{G})|\right]. \tag{15}$$

and by the presumed optimality of h it follows,

$$\mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}\left[|f(\boldsymbol{X}, \boldsymbol{G}) - \hat{h}(\boldsymbol{X}, \boldsymbol{G})|\right] \geq \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}[|f(\boldsymbol{X}, \boldsymbol{G}) - h(\boldsymbol{X}, \boldsymbol{G})|]. \tag{16}$$

On the other hand suppose T_g^h is an optimal transport plan such that $h(x, g) = T_g^h(f(x, g))$ then, by the optimality of $T_*^{\hat{h}}$ it follows

$$\sum_{g\in g} p_g \int_X |f(\boldsymbol{X}, g) - T_g^{\hat{h}}(f(\boldsymbol{X}, g))| d\mu_{X|g} \leq \sum_{g\in g} p_g \int_X |f(\boldsymbol{X}, g) - T_g^h(f(\boldsymbol{X}, g))| d\mu_{X|g}.$$

Using similar properties as the above derivations we can re-write this relationship as

$$\mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}\left[|f(\boldsymbol{X}, \boldsymbol{G}) - \hat{h}(\boldsymbol{X}, \boldsymbol{G})|\right] \leq \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}[|f(\boldsymbol{X}, \boldsymbol{G}) - h(\boldsymbol{X}, \boldsymbol{G})|]. \tag{17}$$

Therefore by Steps (16) and (17) we have

$$\mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}\left[|f(\boldsymbol{X},\boldsymbol{G}) - \hat{h}(\boldsymbol{X},\boldsymbol{G})|\right] = \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}\left[|f(\boldsymbol{X},\boldsymbol{G}) - h(\boldsymbol{X},\boldsymbol{G})|\right],$$

and combining Step 15 with the above concludes

$$\min_{\nu\in\mathcal{P}_1([0,1])} \sum_{g\in g} p_g \mathcal{W}_1(\mu_a,\nu) \leq \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}\left[|f(\boldsymbol{X},\boldsymbol{G}) - h(\boldsymbol{X},\boldsymbol{G})|\right], \tag{18}$$

where $\mathcal{U}_{\text{PR}}(h) = 0$ by assumption. To prove the other direction, now let

$$\bar{\nu} \leftarrow \mathop{\arg\min}_{\nu\in\mathcal{P}_1([0,1])} \sum_{g\in g} p_g \mathcal{W}_1(\mu_a,\nu)$$

and $T_g^{\bar{\nu}}$ be the optimal transport maps from $\mu_g \to \bar{\nu}$ and $\bar{h}(x,g) = T_g^{\bar{\nu}}(f(x,g))$. Now, if we consider

$$\sum_{g\in g} p_g \mathcal{W}_1(\mu_a,\bar{\nu}) = \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}\left[|f(\boldsymbol{X},\boldsymbol{G}) - \bar{h}(\boldsymbol{X},\boldsymbol{G})|\right]$$

then we can easily conclude by the assumed optimality of h that,

$$\min_{\nu\in\mathcal{P}_1([0,1])} \sum_{g\in g} p_g \mathcal{W}_1(\mu_a,\nu) \geq \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{G}}\left[|f(\boldsymbol{X},\boldsymbol{G}) - h(\boldsymbol{X},\boldsymbol{G})|\right]. \tag{19}$$

Finally, recalling that $\bar{h}$ satisfies $\mathcal{U}_{\text{PR}}(\bar{h}) = 0$ since $\bar{h}$ is a Barycenter (Corollary 3.1). Combining Steps 18 and 19 to yield the proof. □

B Proofs

B.1 Proof of Proposition 3.1

Proposition 3.1. For $\mu_a, \mu_b \in \mathcal{P}_2([0,1])$ which are the groupwise score distributions of f, then $\mathcal{W}_2(\mu_a,\mu_b) = 0$ if and only if $\mathcal{U}_{\text{PR}}(f) = 0$.

Proof. Let μ_a, μ_b be the groupwise score distributions of some regressor f. Since W_p is a metric on $\mathcal{P}_p([0,1])$ (according to Proposition 2.3 in (Peyré and Cuturi 2018)) if $\mathcal{W}_2(\mu_a,\mu_b)$ then $\mu_a = \mu_b$. Similarly, by the same property we know that $\mathcal{W}_2(\mu_a,\mu_b) = \mathcal{W}_1(\mu_a,\mu_b) = 0$. Showing that $\mathcal{W}_1(\mu_a,\mu_b) = \mathcal{U}_{\text{PR}}(f)$ completes the proof. To show this equality, recall by definition that

$$\begin{aligned}
\gamma_g(\tau) &= \Pr[f(\boldsymbol{X},\boldsymbol{G}) \geq \tau | \boldsymbol{G} = g] && (20)\\
&= 1 - \Pr[f(\boldsymbol{X},\boldsymbol{G}) \leq \tau | \boldsymbol{G} = g] && (21)\\
&= 1 - F_g(\tau) && (22)
\end{aligned}$$

Plugging this into the expression for $\mathcal{U}_{\text{PR}}$

$$\begin{aligned}
\mathcal{U}_{\text{PR}}(f) = \mathop{\mathbb{E}}_{\tau\in U([0,1])} |\gamma_a(\tau) - \gamma_b(\tau)| &= \int_{[0,1]} |\gamma_a(\tau) - \gamma_b(\tau)|^p d\tau && (23)\\
&= \int_{[0,1]} |(1 - F_a(\tau)) - (1 - F_b(\tau))| d\tau && (24)\\
&= \int_{[0,1]} |F_a(\tau) - F_b(\tau)| d\tau && (25)\\
&= \int_0^1 |F_a^{-1}(t) - F_b^{-1}(t)| dt = \mathcal{W}_1(\mu_a,\mu_b) && (26)
\end{aligned}$$

where the second to last equality was proven in Lemma 6 from (Jiang et al. 2020). □

B.2 Proof of Lemma 3.1

Lemma 3.1. Let f be some learned regressor. For all $\lambda \in [0,1]$ the set of optimally fair regressors for λ-relaxations of f with respect to risk and distributional parity for $\gamma = \text{PR}$ are given by

$$f_\lambda \leftarrow \mathop{\arg\min}_{\hat{f}\in\mathcal{F}} \lambda R(\hat{f}) \quad \text{s.t.} \quad \mathcal{U}_{\text{PR}}(\hat{f}) = (1-\lambda)\mathcal{U}_{\text{PR}}(f) \tag{27}$$

Proof. By Definition of f^* we know that $\mathcal{R}(f^*)$ is

$$\min_{\hat{f}\in\mathcal{F}} \mathcal{R}(\hat{f}) \quad \text{s.t.} \quad \mathcal{U}_{\text{PR}}(\hat{f}) = 0.$$

It follows that

$$\lambda(\min_{\hat{f}\in\mathcal{F}} \mathcal{R}(\hat{f})) = \min_{\hat{f}\in\mathcal{F}} \lambda R(\hat{f}) = \lambda\mathcal{R}(f^*). \tag{28}$$

By definition of f_λ it is straightforward to show that $\mathcal{R}(f_\lambda) = \lambda\mathcal{R}(f^*)$. Under Proposition A.2, it is straightforward to show that $\mathcal{U}_{\text{PR}}(f_\lambda) = (1-\lambda)\mathcal{U}_{\text{PR}}(f)$. Combining these two facts proves the result. □

B.3 Proof of Corollary 2.1

Corollary 2.1. Let μ_* be the ρ_b-weighted barycenter of μ_a, μ_b then the transport plan from $\mu_a \to \mu_*$ (wlog) is computed

$$T_a^*(\omega) = (\rho_a F_{\mu_a} + \rho_b F_{\mu_b}) \circ F_{\mu_a}^{-1}(\omega)$$

Proof. Observe that by Theorem A.1 we can express barycenter from μ_a to μ_* (wlog)

$$\mu_* = (\rho_a \mathtt{id} + bT_a^b)\#\mu_a = (\rho_a F_{\mu_a}^{-1} \circ F_{\mu_a} + \rho_b F_{\mu_b}^{-1} \circ F_{\mu_a})\#\mu_a$$

The second equality follows from Remark 2.1. From this expression, we can define $T_a^* = (\rho_a F_{\mu_a}^{-1} \circ +\rho_b F_{\mu_b}^{-1}) \circ F_{\mu_a}$ as the function which computes the transport from $\mu_a \to \mu_*$. □

B.4 Proof of Proposition 4.1

Proposition 4.1 . For any $\lambda \in [0,1]$, a repaired regressor f_λ satisfies the following

$$R(f_\lambda) = \lambda R(f^*) \quad \text{and} \quad \mathcal{U}_{\text{PR}}(f_\lambda) = (1-\lambda)\mathcal{U}_{\text{PR}}(f)$$

Proof. The first equality follows from the definition of R and linearity of expectation. It is easy to show that

$$\begin{aligned} R(f_\lambda) &= R((1-\lambda)f + \lambda f^*) \\ &= (1-\lambda)R(f) + \lambda R(f^*) = \lambda R(f^*) \end{aligned}$$

where the last equality follows by noting that $R(f) = 0$ by definition. The proof that $\mathcal{U}_{\text{PR}}(f_\lambda) = (1-\lambda)\mathcal{U}_{\text{PR}}(f)$ follows from Proposition A.2. □

B.5 Proof of Theorem 4.1.

In the proof of Theorem 4.1, we make use of the fact that this transport plans are bijective, under Assumption 2.1. In order to show that these plans are bijective we show that they are strictly monotone via the following Remark.

Remark B.1 (Santambrogio (2015), p. 55)**.** *For two measures $\mu, \nu \in \mathcal{P}_p([0,1])$, if ν is non-atomic, then the transport plan T from $\mu \to \nu$ is strictly monotone on a closed domain like* $[0,1]$.

It is well known that strictly monotone functions on a closed domain are bijective, and therefore we claim bijectivity as a corollary of the above result.

Corollary B.1. *A transport plan T that is strictly monotone, on a closed domain, is also bijective.*

Now we begin the proof of Theorem A.1.

Theorem 4.1. Fix $\gamma \in \Gamma$. Let $f : X \times G \to [0,1]$ be a regressor, and f_λ be the geometrically repaired regressor for any $\lambda \in [0,1]$. The map $\lambda \mapsto \mathcal{U}_\gamma(f_\lambda)$ is convex in λ.

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):

$$\mathcal{U}_\gamma(f_\lambda) \triangleq \mathop{\mathbb{E}}_{\tau\sim U([0,1])} |\gamma_a(\tau) - \gamma_b(\tau)|.$$

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

Recall the definition of $\gamma_g(\tau; f_\lambda)$

$$\gamma_g(\tau; f_\lambda) = \Pr[f_\lambda(\boldsymbol{X}, \boldsymbol{G}) \geq \tau | \boldsymbol{G} = g].$$

by Proposition 4.2 we know that $\mu_{g,\lambda}$ is the score distribution associated with $f_\lambda(\cdot, g)$ and so we re-write this expression as a conditional expectation

$$\Pr[f_\lambda(\boldsymbol{X}, \boldsymbol{G}) \geq \tau | \boldsymbol{G} = g] = \int_{[0,1]} \mathbb{1}_{[\tau,1]} d\mu_{g,\lambda} \tag{29}$$

In order to take this derivative, we need to invoke several change of variables to convert this Lebesgue integral to a Riemann integral. We'll proceed for $a \in G$ without loss of generality. Also note for brevity, we present the proof for $\mu_{g,\lambda}$,i.e., the measure associated with $\gamma = \text{PR}$. Similarly, if we condition the l.h.s. of Eq. 29 on $\boldsymbol{Y}$, our results follow similarly for corresponding probability measures associated with this conditional probability, .e.g, we would let $\mu_{g|\boldsymbol{Y},\lambda}$ be the measure associated with the conditional probability $\Pr[f_\lambda(\boldsymbol{X}, \boldsymbol{G}) | \boldsymbol{G} = g, \boldsymbol{Y} \geq \tau]$ for which setting $\boldsymbol{Y}$ computes TPR and FPR respectively.

Following Claim A.1 can re-write $\mu_{a,\lambda} \coloneqq ((1-\lambda)id + \lambda T_a^*)\#\mu_a$. For notational ease, define $\pi_{a,\lambda} \coloneqq (1-\lambda)id + \lambda T_a^*$. Using these substitutions, we have that $\mu_{a,\lambda} = (\pi_{a,\lambda})\#\mu_a$, so γ_a can be equivalently written

$$\gamma_a(\tau) = \int_{[0,1]} \mathbb{1}_{[\tau,1]} d(\pi_{a,\lambda}\#\mu_a).$$

By definition of the push-forward operator

$$\int_{[0,1]} \mathbb{1}_{[\tau,1]} d(\pi_{a,\lambda}\#\mu_a) = \int_{\pi_{a,\lambda}^{-1}([0,1])} \mathbb{1}_{[\tau,1]}(\pi_{a,\lambda}) d\mu_a = \int_{[0,1]} \mathbb{1}_{[\tau,1]}(\pi_{a,\lambda}) d\mu_a.$$

We note that the domain of integration is unchanged in the last equality because π is a bijective mapping from $[0,1] \to [0,1]$ by Corollary B.1, and so $\pi_{g,\lambda}^{-1}([0,1]) = [0,1]$.

For the last change of variables, Let ℓ be the Lebesgue measure. By Assumption 2.1 μ_a is absolutely continuous with respect to ℓ meaning that by the Radon Nikodym-Theorem

$$\int_{[0,1]} \mathbb{1}_{[\tau,1]}(\pi_{g,\lambda}) d\mu_a = \int_{[0,1]} \sigma_a \mathbb{1}_{[\tau,1]}(\pi_{g,\lambda}) d\ell$$

where σ_a is the Radon-Nikodym Derivative, i.e., the probability density function associated with μ_a.

We'll also need to define γ_b similarly. To do this we invoke Lemma A.1 which yields that $\mu_{b,\lambda} = \mu_{a,\frac{1-\lambda}{\rho_a}+\lambda}$. Using this substitution we get

$$\gamma_b(\tau) = \int_{[0,1]} \rho_{\mu_a} \mathbb{1}_{[\tau,1]}(\pi_{b,\frac{1-\lambda}{\rho_a}+\lambda}) d\ell$$

Next, let $h_{a,\tau}(\lambda)$ be the mapping $\lambda \mapsto \mathbb{1}_{[\tau,1]}(\mu_{a,\lambda})$ and $h_{b,\tau}(\lambda)$ be $\lambda \mapsto \mathbb{1}_{[\tau,1]}(\mu_{a,\frac{1-\lambda}{\rho_a}+\lambda})$. Taking the first derivative of this difference, we get

$$\begin{aligned}
&\frac{d}{d\lambda}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)] = \\
&\frac{d}{d\lambda} \int_{[0,1]} \rho_{\mu_a} \cdot [\mathbb{1}_{[\tau,1]}(\pi_{a,\lambda}) - \mathbb{1}_{[\tau,1]}(\pi_{b,\frac{1-\lambda}{\rho_a}+\lambda})] d\ell = \\
&\int_{[0,1]} \rho_{\mu_a} \cdot \left[\frac{d}{d\lambda} \left(\mathbb{1}_{[\tau,1]}(\pi_{a,\lambda}) - \mathbb{1}_{[\tau,1]}(\pi_{b,\frac{1-\lambda}{\rho_a}+\lambda}) \right) \right] d\ell
\end{aligned}$$

where the second equality follows from Leibniz Rule. To finish the derivative, we remind the reader that the derivative of $\frac{d}{d\lambda}\mathbb{1}_{[\tau,1]}(\pi_{g,\lambda})$ is the Dirac delta function $\delta(\pi_{g,\lambda} - \tau)$. It follows that

$$\begin{aligned}
\int_{[0,1]} \rho_{\mu_a} \cdot \left[\frac{d}{d\lambda} \left(\mathbb{1}_{[\tau,1]}(\pi_{a,\lambda}) - \mathbb{1}_{[\tau,1]}(\pi_{b,\frac{1-\lambda}{\rho_a}+\lambda}) \right) \right] d\ell = \int_{[0,1]} \rho_{\mu_a} \cdot \Bigg[& T_a^*(\delta(\pi_{a,\lambda} - \tau)) + \left(\frac{1-\rho_a}{\rho_a} \right) \delta(\pi_{b,\frac{1-\lambda}{\rho_a}+\lambda} - \tau) T_b^* \\
& - \left(\frac{\rho_a - 1}{\rho_a} \right) \delta(\pi_{b,\frac{1-\lambda}{\rho_a}+\lambda} - \tau)\mathtt{id} - \delta(\pi_{a,\lambda} - \tau)\mathtt{id}) \Bigg]
\end{aligned}$$

and by definition of δ of the delta function, we at last obtain

$$\begin{aligned}
\int_{[0,1]} \rho_{\mu_a} \cdot \Bigg[T_a^*(\delta(\pi_{a,\lambda} - \tau)) + \left(\frac{1-\rho_a}{\rho_a} \right) \delta(\pi_{b,\frac{1-\lambda}{\rho_a}+\lambda} - \tau) T_b^* - \left(\frac{\rho_a - 1}{\rho_a} \right) \delta(\pi_{b,\frac{1-\lambda}{\rho_a}+\lambda} - \tau)\mathtt{id} - \delta(\pi_{a,\lambda} - \tau)\mathtt{id}) \Bigg] \\
= \left[T_a^* + \left(\frac{1-\rho_a}{\rho_a} \right) T_b^* - \left(\frac{\rho_a - 1}{\rho_a} \right) \mathtt{id} - \mathtt{id} \right] \circ \tau.
\end{aligned}$$

To summarize, we have just shown that

$$\frac{d}{d\lambda}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)] = \left[T_a^* + \left(\frac{1-\rho_a}{\rho_a}\right) T_b^* - \left(\frac{\rho_a - 1}{\rho_a}\right) \mathtt{id} - \mathtt{id}\right] \circ \tau.$$

To prove convexity we must also compute the second derivative of the above. Since the above does not depend on λ, taking another derivative yields

$$\frac{d^2}{d\lambda^2}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)] = 0. \tag{30}$$

Now, to prove the convexity of $\mathcal{U}_\gamma(f_\lambda)$ we take the second derivative of the absolute value of this difference, i.e.,

$$\frac{d}{d^2\lambda}|h_{a,\tau} - h_{b,\tau}| = \text{sign}(h_{a,\tau} - h_{b,\tau}) \underbrace{\frac{d^2}{d\lambda^2}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)]}_{=0} \tag{31}$$

$$+ 2\underbrace{\delta(h_{a,\tau} - h_{b,\tau})}_{\simeq 0 \text{ or } 1} \underbrace{(\frac{d}{d\lambda}[h_{a,\tau}(\lambda) - h_{b,\tau}(\lambda)])^2}_{\geq 0}. \tag{32}$$

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

$$\frac{d}{d^2\lambda} \underbrace{\mathop{\mathbb{E}}_{\tau\sim U([0,1])} |h_{a,\tau} - h_{b,\tau}|}_{\mathcal{U}_\gamma(f_\lambda)} = \mathop{\mathbb{E}}_{\tau\sim U([0,1])} \left| \underbrace{\frac{d}{d^2\lambda}[h_{a,\tau} - h_{b,\tau}]}_{\geq 0 \text{ by (31)}} \right|.$$

This indicates that $\mathcal{U}_\gamma(f_\lambda)$ is convex (i.e., we have shown that the second derivative is non-negative). □

B.6 Proof of Proposition 4.2

Proposition 4.2. Let $\lambda \in [0, 1]$. Let $\mu_{g,\lambda}$ be the λ-weighted barycenter between μ_g and μ_*, i.e.,

$$\mu_{g,\lambda} \leftarrow \mathop{\arg\min}_{\nu\in\mathcal{P}_2([0,1])} (1-\lambda)\mathcal{W}_2^2(\mu_g, \nu) + \lambda\mathcal{W}_2^2(\mu_*, \nu),$$

then $\mu_{g,\lambda} = Law(f_\lambda(\boldsymbol{X}, \boldsymbol{G})|\boldsymbol{G} = g)$.

Proof. First we recall the definition of geometric repair

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

It is easy to show that for T_g^* we have that (wlog) $f^*(x, a) = T_g^*(f(x, a))$

$$f^*(x, a) = (\rho_a \mathtt{id} + \rho_b T_a^b) \circ F_{\mu_a}(f(x, a)) = \tag{33}$$

$$F_{\mu_*}^{-1}(F_{\mu_a}(f(x, a))) = \tag{34}$$

$$T_g^*(f(x, a)). \tag{35}$$

Where the first equality follows from Theorem 2.3. in (Chzhen et al. 2020), and the second equality is the definition of μ_*. Using this equality in the definition of geometric repair we get

$$f_\lambda(x, g) = (1-\lambda)f(x, g) + \lambda T_g^*(f(x, g)) \tag{36}$$

$$= \left((1-\lambda)\mathtt{id} + \lambda T_g^*\right) \circ f(x, g) \tag{37}$$

If we let μ_g be the groupwise score distribution for group g then we know $\mu_g = Law(f(\boldsymbol{X}, \boldsymbol{G})|\boldsymbol{G} = g)$ by definition. If we pushfoward μ_g using $\left((1-\lambda)\mathtt{id} + \lambda T_g^*\right)$, i.e.,

$$\left((1-\lambda)\mathtt{id} + \lambda T_g^*\right) \#\mu_g = \mathop{\arg\min}_{\nu\in\mathcal{P}_2([0,1])} (1-\lambda)\mathcal{W}_2^2(\mu_g, \nu) + \lambda\mathcal{W}_2^2(\mu_*, \nu)$$

by Claim A.1 and the uniqueness of $\mathcal{W}_2$ barycenters. Noticing the $\mu_{g,\lambda}$ is the score distribution for $f_\lambda(\boldsymbol{X}, \boldsymbol{G})|\boldsymbol{G} = g$ completes the proof. □

B.7 Proof of Theorem 4.2

Theorem 4.2. For all $\lambda \in [0,1]$, the repaired regressor f_λ is pareto optimal in the multi-objective minimization of $\mathcal{R}(\cdot)$ and $\mathcal{U}_{\text{PR}}(\cdot)$

Proof. It is clear from the definition of f_λ that $\{f_\lambda\}_{\lambda\in[0,1]}$ forms a pareto front. Indeed, recall that for any level of unfairness, say $\lambda\mathcal{U}_{\text{PR}}(f^*)$, that f_λ is the regressor which minimizes risk, i.e.,

$$f_\lambda \leftarrow \arg\min_{\hat{f}\in\mathcal{F}} \lambda R(\hat{f}) \quad \text{s.t.}\mathcal{U}_{\text{PR}}(f_\lambda) = (1-\lambda)\mathcal{U}_{\text{PR}}(f).$$

Due to the above, it is easy to see that no classifier can have risk less than f_λ, without decreasing λ, which in turn increase $\mathcal{U}(\cdot)$, proving the pareto optimality of f_λ. Now, suppose for contradiction, $\{f_\lambda\}_{\lambda\in[0,1]}$ did not form a pareto front, i.e., there exists some $h \notin \{f_\lambda\}_{\lambda\in[0,1]}$ such that $h \succ f_\lambda$ for some $\lambda \in [0,1]$. Since $h \succ f_\lambda$ then clearly (WLOG) $\mathcal{R}(h) < \mathcal{R}(f_\lambda)$. However if we select $\lambda_h = \frac{\mathcal{R}(h)}{\mathcal{R}(f_\lambda)}$ then $\mathcal{R}(f_{\lambda_h}) = \mathcal{R}(h)$ and subsequently $\mathcal{U}_{\text{PR}}(f_{\lambda_h}) = \mathcal{U}_{\text{PR}}(h)$, which by definition means $h \in \{f_\lambda\}_{\lambda\in[0,1]}$. In the other case where $\mathcal{U}(h) < \mathcal{U}(f_\lambda)$ the proof follows identically. In both cases, we arrive at a contradiction indicating that $\{f_\lambda\}_{\lambda\in[0,1]}$ is indeed a Pareto Frontier. □

B.8 Proof of Theorem 5.1

Theorem 5.1. As $n_g \to \infty$ the empirical distribution of $\hat{f}_\lambda(x,g)$ converges to to $\mu_{g,\lambda}$ in $\mathcal{W}_2$ almost surely.

Proof. To complete this proof, it will be convenient to consider the following mixture distribution

$$\mathcal{P} = \sum_{g\in G} \rho_g \delta_{\mu_g}$$

and its empirical variant $\hat{\mathcal{P}}$ using $\hat{\rho_g}$ and $\hat{\mu}_g$. Relying on the barycenters uniqueness under Assumption 2.1 in $\mathcal{W}_2$ (proven by Agueh and Carlier (2011)) and the consistency of the Wasserstein barycenter (Le Gouic and Loubes 2017)[Theorem 3], proving that $\hat{\mathcal{P}} \to \mathcal{P}$ in the Wasserstein Distance is sufficient to prove the convergence $\hat{\mu}_{g,\lambda}$.

We now begin the proof. Recall that we can express $\mu_{g,\lambda}$ as λ-weighted barycenter between μ_g, μ_* or as a $\lambda\rho_b$ weighted barycenter between μ_a and μ_b. Consider the latter formulation, i.e.

$$\mathcal{P}_\lambda = (1-\lambda)\rho_b\delta_{\mu_a} + \lambda\rho_b\delta_{\mu_b}$$

Thus via the consistency of Wasserstein barycenters, as stated above, we must only show that $\hat{\rho}_g$ converges to ρ_g, and that $\hat{\mu}_g \to \mu_g$ in $\mathcal{W}_2$. The convergence of $\hat{\rho}_g$ follows by the law of large numbers. The convergence of $\hat{\mu}_g$ follows from the well known facts that the Wasserstein Distance metrizes the weak convergence of probability measures (Villani 2008, Theorem 6.9), and that an empirical measure $\hat{\mu}_k \to \mu$ almost surely, (Varadarajan 1958). From these facts it follows that $\mathcal{W}_2(\hat{\mu}_g, \mu_g) \to 0$ almost surely, completing the proof. □

References

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

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

Brent, R. P. 2013. *Algorithms for minimization without derivatives*. Courier Corporation.

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.; Lee, K.; Barocas, S.; Sa, C. D.; Sen, S.; and Zhang, B. 2023. Is My Prediction Arbitrary? Measuring Self-Consistency in Fair Classification. arXiv:2301.11562.

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*.

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.

Hébert-Johnson, U.; Kim, M.; Reingold, O.; and Rothblum, G. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, 1939–1948. PMLR.

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.

Le Gouic, T.; and Loubes, J.-M. 2017. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3): 901–917.

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

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.

Varadarajan, V. S. 1958. On the Convergence of Sample Probability Distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2): 23–26.

Villani, C. 2008. Optimal Transport: Old and New.