# Standardized Interpretable Fairness Measures for Continuous Risk Scores

**Ann-Kristin Becker** [1] **Oana Dumitrasc** [1] **Klaus Broelemann** [1]

## Abstract

We propose a standardized version of fairness measures for continuous scores with a reasonable interpretation based on the Wasserstein distance. Our measures are easily computable and well suited for quantifying and interpreting the strength of group disparities as well as for comparing biases across different models, datasets, or time points. We derive a link between the different families of existing fairness measures for scores and show that the proposed standardized fairness measures outperform ROC-based fairness measures because they are more explicit and can quantify significant biases that ROC-based fairness measures miss.

## 1. Introduction

In recent years, many decision-making processes in areas such as finance, education, social media or medicine have been automated, often at least in part with the goal of making those decisions more comparable, objective, and non-discriminatory (Esteva et al., 2017; Holstein et al., 2018; Alvarado & Waern, 2018; Bucher, 2017; Rader & Gray, 2015). For high-risk business transactions between individuals and companies (e.g. in the lending industry), often predictions of machine learning algorithms are incorporated into those decisions. Such algorithms aim to differentiate individuals as optimally as possible based on historical data and in terms of future behavior. They assign risk scores or risk categories to individuals. Even with good intentions, the approach runs the risk of directly or indirectly discriminating against individuals on the basis of protected characteristics, such as gender, ethnicity, political background or sexual orientation (Larson et al., 2016; Datta et al., 2014; Köchling & Wehner, 2020). That may be the case, if the data reflects biased social circumstances or include prejudicial historical decisions.

[1]SCHUFA Holding AG, Wiesbaden, Germany. Correspondence to: Ann-Kristin Becker <ann-kristin.becker@schufa.de>, Oana Dumitrasc <oana.dumitrasc@schufa.de>, Klaus Broelemann <klaus.broelemann@schufa.de>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

Such discriminatory predictions manifest as disparities among protected groups and may occur in different forms and for various reasons. For example, individuals belonging to different protected groups may be assigned different scores even if they have the same outcome, or predictions may turn out to have different levels of consistency with the ground-truth risk. Unfortunately, in most cases different notions of algorithmic fairness are incompatible (Barocas et al., 2019; Kleinberg et al., 2018; Saravanakumar, 2021; Chouldechova, 2017; Pleiss et al., 2017). Various measures for algorithmic fairness have been developed that aim to quantify different kinds of group disparities (Zafar et al., 2017; Kamishima et al., 2012; Makhlouf & Zhioua, 2021). So far, most of the available literature discusses the problem in the context of binary decision tasks (Mitchell et al., 2021; Barocas et al., 2019; Kozodoi et al., 2022).

However, in many applications, neither a final decision is known, nor is the explicit cost of false predictions. This is especially be the case when score and decision are performed by different entities. A prominent example is the COMPAS Score (Larson et al., 2016) which was developed by one entity to support decisions done by other entities. It may also be that a score is never applied as a pure decision but only as a quantitative prediction that affects, e.g. the cost of a product (risk-based pricing). In these cases, fairness can only be fully assessed if the disparities between groups are summarized across the entire score model.

This paper presents a novel approach to quantifying group disparities for continuous risk score models. It's major contributions are

- a well interpretable and mathematically sound method for quantifying group disparities in continuous risk score models.
- a standardized framework, that allows for monitoring bias over time or between models and populations, even if there is a shift in the score distribution. Furthermore, standardized measures are unaffected by monotonic transformations of the scores, such as logistic / logit transform. This prevents malicious actors from finding a transformation that hides the bias (see section 3.2).
- bridging the gap between common fairness-metrics

stemming directly from three parity concepts (Kleinberg et al., 2018; Hardt et al., 2016; Barocas et al., 2019; Makhlouf & Zhioua, 2021) and ROC-based approaches (Vogel et al., 2021; Kallus & Zhou, 2019; Yang et al., 2022; Beutel et al., 2019).

As not all group disparities arise from discriminatory circumstances - even large disparities between groups may be explainable or justifiable otherwise - assessing whether disparities are unfair should entail a more detailed analysis of their underlying causes and drivers. Thus, to be explicit, we use the term *disparity measure* instead of *fairness measure* throughout the rest of the paper to underline that all discussed measures are purely observational.

The paper is structured as follows: Most of the available quantitative disparity metrics for classifiers reduce down to three main parity concepts that are based on conditional independence: Independence, separation and sufficiency (Barocas et al., 2019; Makhlouf & Zhioua, 2021; Kozodoi et al., 2022). In Section 2, we discuss these concepts and existing related measures in terms of binary classifiers first, and generalize them to continuous risk scores in Section 3. We show that our proposed measures are more flexible than many existing metrics and we discuss their interpretability. In Section 4, we compare the presented measures to ROC-based disparity measures, and we prove that our proposed measures impose a stronger objective and are better suited to detect bias. We outline published related work throughout each section. Section 5 contains results of experiments using benchmark data and Section 6 includes final discussion and outlook. All proofs of technical results are deferred to the appendix.

## 2. Parity concepts and fairness measures for classifiers

Let $Y$ denote a binary target variable with favorable outcome class $Y = 0$ and unfavorable class $Y = 1$, and $X$ a set of predictors. Let $S \in \mathcal{S} \subset \mathbb{R}$ denote an estimate of the posterior probability of the favorable outcome of $Y$, $\mathbb{P}(Y = 0 \mid X)$ or some increasing function of this quantity, in the following called *(risk) score*, with cumulative distribution function $F_S$ and density function $f_S$. We assume $\mathcal{S}$ to be bounded with $|\mathcal{S}| = \sup \mathcal{S} - \inf \mathcal{S}$ denoting the length of the score range. Let $A$ be a (protected) attribute of interest defining two (protected) groups ($A \in \{a, b\}$ binary w.l.o.g.). We choose $A = b$ as the group of interest, e.g. the expected discriminated group. All discussed measures are purely observational and based on the joint distribution of $(S, A, Y)$. They can be easily calculated if a random sample of the joint distribution is available.

Note that each continuous score $S$ induces an infinite set of binary classifiers by choosing a threshold $s \in \mathcal{S}$ and accepting every sample with $S > s$. We define disparity measures for binary classifiers in dependence of such a threshold value $s$. For a group $A$, the *positive rate* at a threshold $s$ is given by $\mathrm{PR}_A(s) = \mathbb{P}(S > s|A) = 1 - F_{S|A}(s)$, the *true positive* and *false positive rates* by $\mathrm{TPR}_A(s) = 1 - F_{S|A,Y=0}(s)$ and $\mathrm{FPR}_A(s) = 1 - F_{S|A,Y=1}(s)$, respectively. We will write in short $F := F_S$ and $f := f_S$, as well as $S_{ay} := S|A = a, Y = y$, $F_{ay} := F_{S|A=a,Y=y}$ and $S_{by}, F_{by}, f_{ay}, f_{by}$ for the conditional random variables, distribution functions and density functions. For a cumulative distribution function $G$, we denote by $G^{-1}$ the related quantile function (generalized inverse) with $G^{-1}(p) = \inf\{x \in \mathbb{R} : p \leq G(x)\}$ which fulfills $G^{-1}(G(X)) = X$ almost surely. If $G$ is continuous and strictly monotonically increasing, then the quantile function is the inverse.

**Independence (selection rate parity)** The random variables $S$ and $A$ satisfy independence if $S \perp\!\!\!\perp A$, which implies $F_{S|A=a} = F_{S|A=b} = F_S$. Group disparity of classifiers can be quantified by the difference between the positive rates (Makhlouf & Zhioua, 2021; Zafar et al., 2017; Dwork et al., 2012)

$$\begin{aligned} \text{c-bias}_{\text{IND}}(S_a, S_b; s) &= \mathrm{PR}_b(s) - \mathrm{PR}_a(s) \\ &= F_a(s) - F_b(s). \end{aligned} \tag{1}$$

The concept of independence contradicts optimality $S = Y$, if $Y \not\perp\!\!\!\perp A$ and is, thus, not an intuitive fairness measure in most cases. On the other hand, the following two measures, separation and sufficiency, are both compatible with optimality and allow $A \not\perp\!\!\!\perp Y$, as they include the target variable $Y$ in the independence statements and allow for disparities that can be explained by group differences in the ground-truth.

**Separation (error rate parity)** The random variables $S$, $A$ and $Y$ satisfy separation if $S \perp\!\!\!\perp A \mid Y$. For a binary outcome $Y$, the separation condition splits into *true positive rate parity* $F_{S|A=a,Y=0} = F_{S|A=b,Y=0} = F_{S|Y=0}$ (*equal opportunity*, EO) (Zhang & Bareinboim, 2018b; Hardt et al., 2016) and *false positive rate parity* $F_{S|A=a,Y=1} = F_{S|A=b,Y=1} = F_{S|Y=1}$ (*predictive equality*, PE) (Corbett-Davies et al., 2017; Makhlouf & Zhioua, 2021). If both hold, the condition is also known as *equalized odds* (Makhlouf & Zhioua, 2021; Hardt et al., 2016). Group disparity of classifiers can be quantified by the difference between the true and false positive rates

$$\begin{aligned} \text{c-bias}_{\text{EO}}(S_a, S_b; s) &= \mathrm{TPR}_b(s) - \mathrm{TPR}_a(s) \\ &= F_{a0}(s) - F_{b0}(s), \end{aligned} \tag{2}$$

$$\begin{aligned} \text{c-bias}_{\text{PE}}(S_a, S_b; s) &= \mathrm{FPR}_b(s) - \mathrm{FPR}_a(s) \\ &= F_{a1}(s) - F_{b1}(s). \end{aligned} \tag{3}$$

**Sufficiency (predictive value parity)** The random variables $S$, $A$ and $Y$ satisfy sufficiency if $Y \perp\!\!\!\perp A \mid S$ (in words, $S$ is sufficient to optimally predict $Y$). Sufficiency implies group parity of positive and negative predictive values. However, especially in case of continuous scores, usually, *calibration* within each group (Kleinberg et al., 2018) (resp. *test fairness* (Chouldechova, 2017)), as an equivalent concept, is used instead (Barocas et al., 2019). The calibration bias examines if the model's predicted probability deviates similarly strongly from the true outcome rates within each group:

$$\begin{aligned}\text{c-bias}_{\text{CALI}}(S_a, S_b; s) = \mathbb{P}(Y = 0|A = b, S = s) \\ - \mathbb{P}(Y = 0|A = a, S = s).\end{aligned} \tag{4}$$

*Well-calibration* (Kleinberg et al., 2018; Pleiss et al., 2017) additionally requires the prediction of both groups to accurately reflect the ground truth $\mathbb{P}(Y = 0|A, S = s) = s$. For determining the calibration difference, the score range is usually binned into a fixed number of intervals. A high calibration bias reflects the fact that (for a given score $s$) the lower-risk group carries the costs of the higher-risk group. The concept of sufficiency is especially important if the model is applied in a context, where both, the score and the group membership are available to the decision maker. Then, a high calibration bias will evoke a group-specific interpretation and handling of identical score values. On the other hand, sufficiency does not prevent discrimination: high- and low-risk individuals of a group can be mixed and assigned an intermediate risk score without violating sufficiency. Moreover, sufficiency is often naturally fulfilled as a consequence of unconstrained supervised learning, especially if the group membership is (at least to some extent) encoded in the input data. Thus, it is usually not a constraint and not a trade-off with predictive performance (Liu et al., 2019).

If separation is violated, the model output includes more information about the group $A$ as is justified by the ground truth $Y$ alone. So, different groups carry different costs of misclassification. It is therefore a reasonable concept for surfacing potential inequities. Conversely, a violation of sufficiency results in a different calibration and a different meaning of identical score values per group. That is the case, if the relation of $A$ and $Y$ is not properly modeled by the score.

In general, independence, separation and sufficiency are opposing concepts. It can be shown that for a given dataset, except for special cases (like perfect prediction or equal base rates), every pair of the three parity concepts is mathematically incompatible (Barocas et al., 2019; Kleinberg et al., 2018; Saravanakumar, 2021; Chouldechova, 2017; Pleiss et al., 2017).

## 3. Generalization to continuous risk scores

We propose to use the expected absolute classifier bias as a disparity measure for scores. Note, that an expected value of zero implies that every classifier derived from the score by choosing a group-unspecific threshold will be bias-free. By evaluating and aggregating the bias across all possible decision thresholds, this generalization serves as a useful diagnostic tool in fairness analyses and follows a similar idea as used in ROC analyses. The two proposed versions can be seen as generalized rate differences. They differ only in the way, in which possible thresholds are weighted. We show, that for the concepts independence and separation, the proposed disparity measures are identical to Wasserstein distances between the groupwise score-distributions.

The use of Wasserstein distance in previous works has focused mainly on independence fairness (e.g. demographic parity), therefore a consideration of all three disparity concepts (independence, separation and sufficiency / calibration) for continuous risk scores is novel to this work.

### 3.1. Expected classifier bias with uniformly weighted thresholds

**Definition 3.1.** By assuming each threshold $s \in \mathcal{S}$ is equally important, we define

$$\begin{aligned}\text{bias}_x^{\mathcal{U}}(S_a, S_b) &:= \mathbb{E}_{S\sim\mathcal{U}}[|\text{ c-bias}_x(S_a, S_b; S)|] \\ &= \frac{1}{|\mathcal{S}|}\int_{\mathcal{S}} |\text{ c-bias}_x(S_a, S_b; s)|\, ds.\end{aligned} \tag{5}$$

**Theorem 3.2.** For the concepts independence and separation, i.e for $x \in \{\text{IND, PE, EO}\}$, it holds:

(i) $\text{bias}_x^{\mathcal{U}}(S|A = a, S|A = b)$ is equal to the normalized Wasserstein-1-distance between the conditional score distributions in the groups over the (finite) score region $\mathcal{S}$ i.e.

$$\text{bias}_x^{\mathcal{U}}(S_a, S_b) = \frac{1}{|\mathcal{S}|} \cdot W_1(S_{ay}, S_{by}), \tag{6}$$

where $y = 0$ for $x = \text{EO}$, $y = 1$ for $x = \text{PE}$, and $y = \cdot$ for $x = \text{IND}$.

(ii) As a consequence, we can derive the disparity between average scores per group (known as *balance for the positive / negative class* (Kleinberg et al., 2018)) as a lower bound, i.e.

$$\text{bias}_x^{\mathcal{U}}(S_a, S_b) \geq \frac{1}{|\mathcal{S}|}\left|\mathbb{E}[S_{by}] - \mathbb{E}[S_{ay}]\right|. \tag{7}$$

A similar version of Theorem 3.2 (i) for independence bias has previously be presented by Jiang et al. (2020), but they did not draw the connection to the *balance for the positive*

*/ negative class*. The Wasserstein distance was recently proposed as a fairness measure (Miroshnikov et al., 2022; Kwegyir-Aggrey et al., 2021; Zhao, 2023) mainly for independence bias, and it was especially used for debiasing purposes earlier (Miroshnikov et al., 2021; Han et al., 2023; Chzhen et al., 2020). Fairness of scores has also been subject for regression tasks (Agarwal et al., 2019; Wei et al., 2023; Zhao, 2023). Again, due to the different target value, only for independence bias. A formal definition and properties of $W_1$ can be found in the appendix. For calibration, the bias $\text{bias}^{\mathcal{U}}_{\text{CALI}}$ is equal to the two-sample version of the the $l_1$-calibration error (Kumar et al., 2019).

### 3.2. Standardized Measures

It can be difficult to compare the expected classifier bias of datasets with distinct score distributions. Especially for imbalanced datasets score distributions are often highly skewed. In this case, disparities in dense score areas may be more critical as they affect more samples. Therefore, we developed a method that standardizes the bias computation, making it independent of data skewness.

Our standardized disparity measures for risk scores are important especially when a monotonic transformation is applied to the score. A good example of such a scenario is given by the logistic regression, where both the probability or the linear term can be used as a score. The risk assessment of both variants is the same. It is also most likely that down-stream tasks would adopt to the score representation (linear term /probability) used. This means, both representations are likely to lead to the same treatment in down-stream tasks and to the same (un)fairness. Without invariance to monotonic transformations, the two representations would have different bias-measures.

In a worst-case scenario an entity could apply a strictly monotonic function to their score, stretching areas with low bias and shrinking areas with high bias. Doing so would allow to mask the bias without any change in accuracy or better ranking of the disadvantaged group. This has already be proposed (Jiang et al., 2020).

That is why we propose an alternative generalization that weights the thresholds by their frequency observed in the population. By this, the resulting disparity measures become independent of the concrete distribution and evaluate the fairness of a bipartite ranking task, similar to ROC measures. Each sample is equally important in this scenario.

As a consequence, this allows for a meaningful comparison of different scores, even of scores with different ranges (e.g. a normally-distributed score that can take any real value and a uniformly-distributed score that only takes probabilities). Our methodology can thus be utilized to assess the effectiveness of debiasing approaches (Hort et al., 2023).

**Definition 3.3.**

$$\text{bias}^S_x(S_a, S_b) := \mathbb{E}_{S\sim F}[|\,\text{c-bias}_x(S_a, S_b; S)|] \tag{8}$$

$$= \int_0^1 |\,\text{c-bias}_x(S_a, S_b; F^{-1}(r))|\,dr$$

$$= \int_{\mathcal{S}} |\,\text{c-bias}_x(S_a, S_b; s)| \cdot f(s)\,ds \tag{9}$$

Note that $\text{bias}^S_x(S|A = a, S|A = b)$ is invariant under monotonic score transformations as it is a purely ranking-based metric, $\text{bias}^{\mathcal{U}}_x(S|A = a, S|A = b)$ is not. If $S \sim \mathcal{U}$ it holds $\text{bias}^S = \text{bias}^{\mathcal{U}}$. We show, that the standardized bias is equal to the Wasserstein-1-distance between quantile-transformed distributions. To our knowledge, this is the first introduction of a fairness measure based on the Wasserstein distance, which is invariant to transformations.

**Theorem 3.4.** For the concepts independence and separation, i.e. for $x \in \{\text{IND, PE, EO}\}$, it holds:

(i) $\text{bias}^S_x$ is equal to the Wasserstein-1-distance using the push-forward by the quantile function $F^{-1}\#\mathcal{L}_1$ as ground metric (with $y = 0$ for $x = \text{EO}$, $y = 1$ for $x = \text{PE}$, and $y = \cdot$ for $x = \text{IND}$)

$$\text{bias}^S_x(S_a, S_b) = W_1(F(S_{ay}), F(S_{by}))$$
$$= \int_0^1 |F_{ay} \circ F^{-1}(t) - F_{by} \circ F^{-1}(t)|\,dt. \tag{10}$$

(ii) We can derive the disparity between the average relative rank per group as a lower bound.

For reasons of simplicity, we will use the notation $W_Z(X, Y) := W_1(F_Z(X), F_Z(Y))$.

### 3.3. Interpretation of the score bias

In general, $\text{bias}^{\mathcal{U}}$ and $\text{bias}^S$ take values in the interval $[0, 1]$ as they are expected values over rate differences. The optimal value, a bias of zero, indicates group parity for all decision thresholds with respect to the analyzed type of classifier error. When comparing multiple score models or one model over multiple populations, a smaller bias is preferable. The standardized method allows direct comparison of models with different score distributions with respect to group parity in bipartite ranking tasks. $\text{bias}^{\mathcal{U}}$ and $\text{bias}^S$ can be interpreted as the classifier bias to be expected at a randomly chosen threshold - either randomly selected from all available score values ($\text{bias}^{\mathcal{U}}$) or by randomly selecting one sample and assigning the favorable label to all samples that are ranked higher ($\text{bias}^S$).

In addition, the separation and independence biases can be interpreted in terms of the Wasserstein distance (or *Earth*

*Mover distance*): The bias is measured as the minimum cost of aligning the two groups with respect to the analyzed type of classifier error. Here, the baseline distance is measured in normalized scores for $\text{bias}^{\mathcal{U}}$ or in ranks for $\text{bias}^{S}$. It indicates what proportion of a group must be scored (how) differently in order to equalize the groups.

### 3.4. Positive and negative components of the score bias

Unlike a classifier bias, a score bias does not have to be overall positive or negative for a particular group. Instead, there may be thresholds at which one group is disadvantaged and others at which the opposing group is disadvantaged. To further analyze the bias, we can decompose the total bias into a positive and a negative component (positive and negative from the point of view of the chosen disadvantaged group, here $b$). For this purpose, the classifier bias is divided into a positive and a negative part for each threshold

$$\text{c-bias}^{+}(s) = \max(\text{c-bias}(s), 0) \quad \text{and}$$
$$\text{c-bias}^{-}(s) = -\min(\text{c-bias}(s), 0).$$

This allows to derive a decomposition of both score bias types into two components:

$$\text{pos-bias}_x(S_a, S_b) = \mathbb{E}[\text{c-bias}_x^{+}(S_a, S_b; S)], \tag{11}$$
$$\text{neg-bias}_x(S_a, S_b) = \mathbb{E}[\text{c-bias}_x^{-}(S_a, S_b; S)], \tag{12}$$

where $\text{bias}_x(S_a, S_b) = \text{pos-bias}_x(S_a, S_b) + \text{neg-bias}_x(S_a, S_b)$. By dividing each component by the total bias, a percentage can be calculated. The decomposition helps to interpret, which of the two compared groups is affected predominantly negatively by the observed bias. A similar decomposition of a Wasserstein bias was proposed by Miroshnikov et al. (2022).

## 4. ROC-based fairness measures and relations

Furthermore, there exists a wide variety of (separation) fairness metrics which are calculated based on ROC curves or the area under the curves. We show, that the proposed standardized bias measures outperform these ROC-based measures as they are more explicit, easier to interpret, and can measure biases, that ROC-based fairness measures cannot catch. We define the ROC curve between two arbitrary random variables $G, H$, similar to Vogel et al. (2021). In a bipartite ranking or scoring task, the ROC curve is usually used to evaluate the separability between positive and negative outcome class. In this case, $G = S_0, H = S_1$.

**Definition 4.1** (ROC)**.** Let $G$ and $H$ be two random variables with cumulative distribution functions $F_G, F_H$ on $\mathbb{R}$ with quantile functions $F_G^{-1}, F_H^{-1}$. Then the ROC curve of $G$ and $H$ is the mapping

$$\text{ROC}_{G,H} : p \in [0,1] \mapsto 1 - F_G(F_H^{-1}(1-p)) \tag{13}$$

with the area under the curve (AUROC) and the Gini coefficient defined as

$$\text{AUROC}(G,H) = \int_0^1 \text{ROC}_{G,H}(p)\, dp \quad \text{and}$$
$$\text{Gini}(G,H) = 2 \cdot \text{AUROC}(G,H) - 1. \tag{14}$$

**Definition 4.2.** Similar to the above introduced biases, a ROC-based disparity-measure for score models can be defined as the expected absolute difference between two ROC curves

$$\text{bias}_{\text{ROC}}(S_a, S_b) = \mathbb{E}[|\,\text{ROC}_{S_{b0},S_{b1}} - \text{ROC}_{S_{a0},S_{a1}}\,|]$$
$$= \int_0^1 |\,\text{ROC}_{S_{b0},S_{b1}}(s) - \text{ROC}_{S_{a0},S_{a1}}(s)|\, ds$$

$\text{bias}_{\text{ROC}}(S|A = a, S|A = b)$ is equal to the absolute between ROC area (*ABROCA*) (Gardner et al., 2019). In general, $\text{bias}_{\text{ROC}}(S|A = a, S|A = b) \geq |\,\text{AUROC}(S_{b0}, S_{b1}) - \text{AUROC}(S_{a0}, S_{a1})|$, which is known as *intra-group fairness* and often used as a fairness measure for scores (Vogel et al., 2021; Beutel et al., 2019; Borkan et al., 2019; Yang et al., 2022). If the ROC curves of two groups do not cross (i.e. one group gets uniformly better scores than the other), equality holds. As the thresholds that lead to certain ROC values (pair of FPR and TPR at a certain score threshold) are group-specific, it is not sufficient to compare intra-group ROC curves (Vogel et al., 2021). Thus, we define a second ROC-based measure that compares the discriminatory power across groups and is based on the cross-ROC curve (Kallus & Zhou, 2019).

**Definition 4.3.** We define the cross-ROC bias as the expected difference of the ROC curves across groups

$$\text{bias}_{\text{xROC}}(S_a, S_b) = \mathbb{E}[|\,\text{ROC}_{S_{b0},S_{a1}} - \text{ROC}_{S_{a0},S_{b1}}\,|]$$
$$= \int_0^1 |\,\text{ROC}_{S_{b0},S_{a1}}(s) - \text{ROC}_{S_{a0},S_{b1}}(s)|\, ds$$

The cross-ROC bias evaluates the difference in separability of negatives samples in one group versus positive samples of the other group. $\text{bias}_{\text{xROC}}$ is always greater or equal to the related AUROC-based fairness-measure $|\,\text{AUROC}(S_{a0}, S_{b1}) - \text{AUROC}(S_{b0}, S_{a1})|$, that is known as *subgroup positive background negative* (BPSN) or *inter-group fairness* (Borkan et al., 2019; Vogel et al., 2021; Beutel et al., 2019; Yang et al., 2022).

### 4.1. Relating Wasserstein and ROC biases

We now reveal some connections of the standardized Wasserstein disparity measures with the ROC-based disparity measures. We first consider the general case of the Wasserstein distance between two random variables $X, Y$ quantile-transformed by $Z$.

For the following section, we require $\mathrm{ROC}_{X,X}(r) = r$. This is fulfilled, whenever $F_X$ is continuous and strictly monontonic increasing, so it permits a well-defined inverse, or if the ROC-curve is interpolated linearly from finite data.

**Theorem 4.4.** The quantile-transformed Wasserstein distance can be rewritten in terms of ROC

$$\begin{aligned} W_Z(X,Y) &= \int_0^1 |F_X(F_Z^{-1}(t)) - F_Y(F_Z^{-1}(t))|dt \\ &= \int_0^1 |\,\mathrm{ROC}_{X,Z}(t) - \mathrm{ROC}_{Y,Z}(t)|dt. \end{aligned} \tag{15}$$

Moreover, we easily get the following result.

**Proposition 4.5.** Let $Z_i$, $i = 1, \ldots, n$ be random variables with values in $\mathcal{S}$ and with densities $f_i$. Let $Z_K$ be their mixture, where $K$ is a random variable with values in $\{1, \ldots, n\}$. Then their joint density is given by $f_{Z_K}(x) = \sum_{i=1}^n \mathbb{P}(K = i) f_i(x)$ and it holds

$$W_{Z_K}(X,Y) = \sum_{i=1}^n \mathbb{P}(K = i) W_{Z_i}(X,Y). \tag{16}$$

Formulating $S$ as a mixture of the two groups and two outcome classes $S_{a0}, S_{a1}, S_{b0}, S_{b1}$, we get

$$\begin{aligned} W_S(S_{ay}, S_{by}) = w_{a0} \cdot W_{S_{a0}}(S_{ay}, S_{by}) \\ + w_{b0} \cdot W_{S_{b0}}(S_{ay}, S_{by}) \\ + w_{a1} \cdot W_{S_{a1}}(S_{ay}, S_{by}) \\ + w_{b1} \cdot W_{S_{b1}}(S_{ay}, S_{by}). \end{aligned} \tag{17}$$

By looking at the different mixture components, we can reveal a connection to the ROC-based disparity measures.

**Lemma 4.6.** $W_{S_{ay}}(S_{ay}, S_{by})$ and $W_{S_{a\tilde{y}}}(S_{ay}, S_{by})$ for $\tilde{y} \neq y$ can be rewritten in terms of ROC

$$W_{S_{ay}}(S_{ay}, S_{by}) = \int_0^1 \left| \mathrm{ROC}_{S_{by},S_{ay}}(r) - r \right| dr, \tag{18}$$

$$\begin{aligned} W_{S_{a\tilde{y}}}(S_{ay}, S_{by}) = \int_0^1 \Big| \mathrm{ROC}_{S_{by},S_{a\tilde{y}}}(r) - \\ \mathrm{ROC}_{S_{ay},S_{a\tilde{y}}}(r) \Big| dr. \end{aligned} \tag{19}$$

**Lemma 4.7.** From Jensen inequality, it follows

$$\begin{aligned} &W_{S_{ay}}(S_{ay}, S_{by}) \geq |\,\mathrm{AUROC}(S_{by}, S_{ay}) - \tfrac{1}{2}| \\ &= \tfrac{1}{2} \cdot |\,\mathrm{Gini}(S_{by}, S_{ay})|. \end{aligned} \tag{20}$$

If the ROC curve does not cross the diagonal, then equality holds.

**Theorem 4.8.** We can now decompose each separation bias into a sum of four ROC statements. Let $w_{ay} = \mathbb{P}(Y = y, A = a)$ and $w_{by} = \mathbb{P}(Y = y, A = b)$, as well as $w_y = \mathbb{P}(Y = y)$, then it holds:

$$\begin{aligned} &\mathrm{bias}^S_{\mathrm{EO}}(S_a, S_b) = w_{a0} \int_0^1 |\mathrm{ROC}_{S_{b0},S_{a0}}(r) - r|\, dr \\ &+ w_{b0} \int_0^1 |\mathrm{ROC}_{S_{a0},S_{b0}}(r) - r|\, dr \\ &+ w_{a1} \int_0^1 |\,\mathrm{ROC}_{S_{a0},S_{a1}}(r) - \mathrm{ROC}_{S_{b0},S_{a1}}(r)|dr \\ &+ w_{b1} \int_0^1 |\,\mathrm{ROC}_{S_{a0},S_{b1}}(r) - \mathrm{ROC}_{S_{b0},S_{b1}}(r)|dr, \end{aligned} \tag{21}$$

and analogously for $\mathrm{bias}^S_{\mathrm{PE}}(S_a, S_b)$ by exchanging $w_{a0}$ with $w_{a1}$, $w_{b0}$ with $w_{b1}$, $S_{a0}$ with $S_{a1}$ and $S_{b0}$ with $S_{b1}$.

**Corollary 4.9.** From Theorem 4.8 we can infer upper bounds of the separation biases and their sum

$$\begin{aligned} &\mathrm{bias}^S_{\mathrm{EO}}(S_a, S_b) \leq 1 - \frac{w_0}{2} \text{ and} \\ &\mathrm{bias}^S_{\mathrm{PE}}(S_a, S_b) \leq 1 - \frac{w_1}{2} \end{aligned} \tag{22}$$

$$\Rightarrow \mathrm{bias}^S_{\mathrm{EO}}(S_a, S_b) + \mathrm{bias}^S_{\mathrm{PE}}(S_a, S_b) \leq \frac{3}{2}. \tag{23}$$

Moreover, we show that the sum of the separation biases is an upper bound (up to population-specific constants) to both ROC biases and the separability of the groups within each outcome class.

**Theorem 4.10.** The following inequality holds [1]

$$\begin{aligned} &\mathrm{bias}^S_{\mathrm{EO}}(S_a, S_b) + \mathrm{bias}^S_{\mathrm{PE}}(S_a, S_b) \\ &= W_S(S_{a0}, S_{b0}) + W_S(S_{a1}, S_{b1}) \\ &\geq \frac{\min(w_{a0}, w_{a1}, w_{b0}, w_{b1})}{2} \cdot (\mathrm{bias}_{\mathrm{ROC}} + \mathrm{bias}_{\mathrm{xROC}} \\ &\quad + \mathrm{Gini}(S_{a0}, S_{b0}) + \mathrm{Gini}(S_{a1}, S_{b1})). \end{aligned} \tag{24}$$

Note, that the constant $\min(w_{a0}, w_{a1}, w_{b0}, w_{b1})/2$ is fixed for each dataset. Thus, decreasing both separation biases leads to a decrease of the sum of both ROC biases as well as the separability of the groups within each outcome class. Especially, separation biases of zero also diminish both ROC biases.

**Corollary 4.11.** Zero separation biases imply zero ROC biases

$$\begin{aligned} &\mathrm{bias}^S_{\mathrm{EO}}(S_a, S_b) = \mathrm{bias}^S_{\mathrm{PE}}(S_a, S_b) = 0 \\ &\Rightarrow \mathrm{bias}_{\mathrm{ROC}}(S_a, S_b) = \mathrm{bias}_{\mathrm{xROC}}(S_a, S_b) = 0. \end{aligned} \tag{25}$$

The inverse does not hold.

[1] Note, that if $F_{ay}$ and $F_{by}$ have identical supports and permit an inverse, then $\mathrm{Gini}(S_{ay}, S_{by}) = \mathrm{Gini}(S_{by}, S_{ay})$. If this symmetry is not fulfilled, the minimum of both must be used on the right side.

**Theorem 4.12.** Moreover, if only one separation bias is zero, ROC and cross-ROC bias become equal

$$\begin{aligned} &\text{bias}^S_{\text{EO}}(S_a, S_b) = 0 \text{ or } \text{bias}^S_{\text{PE}}(S_a, S_b) = 0 \\ \Rightarrow &\text{bias}_{\text{ROC}}(S_a, S_b) = \text{bias}_{\text{xROC}}(S_a, S_b). \end{aligned} \tag{26}$$

## 5. Experiments

We use the COMPAS dataset[2], the Adult dataset[3] and the German Credit dataset[4] to demonstrate the application of the fairness measures for continuous risk scores. For each bias, we perform permutation tests to determine statistical significance under the null hypothesis of group parity (Di-Ciccio et al., 2020; Schefzik et al., 2021). The core of this paper is our novel bias evaluation metric, therefore the focus of our experiments is not on achieving a low bias, but on demonstrating where and how detecting bias is useful, for example while comparing different models and analyzing debiasing approaches. In addition, we perform an experiment with synthetic datasets where the equal opportunity bias is controllable by one parameter. Experimental details and complete results including all presented bias types can be found in appendix. The code used for the experiments in this study is online available [5]. The repository includes detailed instructions for reproducing the results.

### 5.1. COMPAS

We calculate the different types of biases for the famous COMPAS decile score ($n = 7214$), which predicts the risk of violent recidivism within two years following release. We choose race as protected attribute and set African-America as the expected discriminated group versus Caucasian race. To be consistent with the notation in this paper, we calculate the counter-score, so that a high score stands for the favorable outcome. In contrast to the original analysis (Larson et al., 2016) we calculate the bias over the entire score area. Results (Table 1) show a significant separation bias against the African-American and in favor of the Caucasian race. The disadvantaged group experiences a much lower true-positive rate (rate difference in average $\text{bias}^S_{\text{EO}} = 0.16$) as well as false positive rate (rate difference in average $\text{bias}^S_{\text{PE}} = 0.15$). The calibration bias is lower and not statistically significant but predominantly in favor of the African-American race. While the ROC bias is also low (implying that the separability is equally good in both groups considered independently), the cross-ROC bias is again high. In this case, there is not much difference between $\text{bias}^{\mathcal{U}}$ and $\text{bias}^S$ (complete results can be found in appendix).

Table 1: Bias of COMPAS score of African-American vs. Caucasian.

| type of bias | total | pos. | neg. | p-value |
|---|---|---|---|---|
| $\text{bias}^S_{\text{EO}}$ | 0.161 | 0% | 100% | <0.01 |
| $\text{bias}^S_{\text{PE}}$ | 0.154 | 0% | 100% | <0.01 |
| $\text{bias}^S_{\text{CALI}}$ | 0.034 | 79% | 21% | 0.30 |
| $\text{bias}_{\text{ROC}}$ | 0.016 | 46% | 54% | 0.31 |
| $\text{bias}_{\text{xROC}}$ | 0.273 | 0% | 100% | <0.01 |

### 5.2. German Credit Data

Moreover, we trained two logistic regression scores on the German Credit Risk dataset ($n = 1000$) to predict if a borrower belongs to the good risk class. The first model *LogR* uses all available nine predictors including the feature *sex*, which we choose as protected attribute. For the second score *LogR (debiased)*, the protected attribute was removed from the model input. We set *female* as the expected discriminated group. The scores achieve an AUROC of 0.772 and 0.771.

Compared to COMPAS, the separation biases of both models are lower (all below 0.1) whereas the calibration biases are higher (close to 0.3). Removing the attribute decreases the separation bias (Table 2), while it slightly increases the calibration bias. Note that while *LogR* contains bias to the detriment of female, the debiased model predominantly favors female over male. This demonstrates the use and importance of the split into positive and negative components introduced in 3.4.

### 5.3. UCI Adult

Moreover, we used the UCI Adult dataset ($n = 32561$) to train three different scores that predict the probability of the income being above 50k$. Again, we choose *sex* as the protected attribute and *female* as the expected discriminated group. As before, a logistic regression was trained including (*logR*) and excluding (*logR (debiased)*) the protected attribute *sex*. Moreover, an XGBoost model (*XGB*), was trained with the complete feature set. XGB is known as one of the best performing methods on tabular data (Shwartz-Ziv & Armon, 2021). The logistic regression achieved an AUROC of 0.898 with and of 0.897 without the protected attribute, the XGB model achieved an AUROC of 0.922 on the testset. Resulting biases are shown in Table 3, with the lowest bias in bold.

Removing the protected attribute from the model input im-

[2]https://raw.githubusercontent.com/propublica/compas-analysis/master/compas-scores-two-years.csv

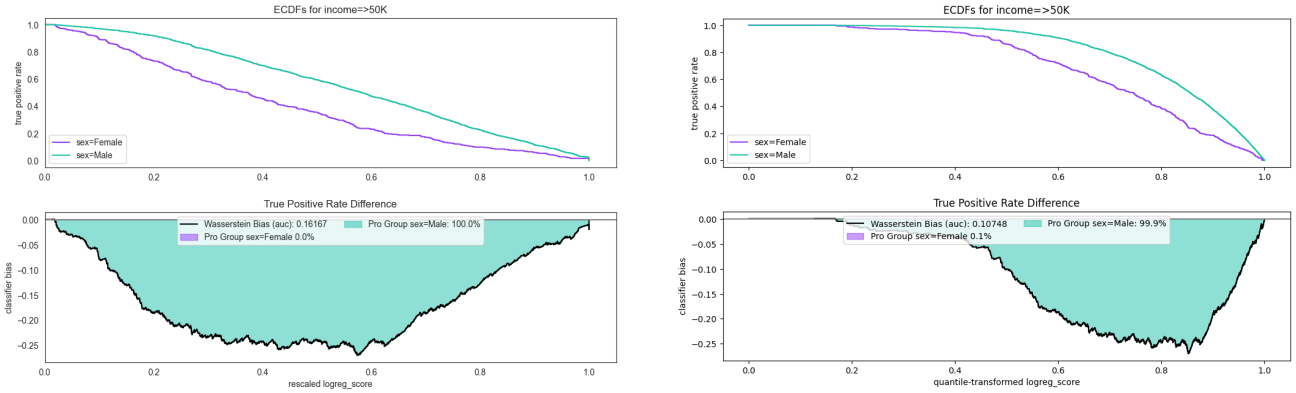
[3]https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

[4]https://www.kaggle.com/datasets/uciml/german-credit?resource=download

[5]https://github.com/schufa-innovationlab/fair-scoring

Table 2: Gender bias of logistic regression (trained with and without sex) scores on German Credit Risk dataset; positive and negative component from the point of view of female persons.

| | LogR | | | | LogR (debiased) | | | |
|---|---|---|---|---|---|---|---|---|
| type of bias | total bias | pos. | neg. | p-value | total bias | pos. | neg. | p-value |
| $\text{bias}^S_{\text{EO}}$ | 0.083 | 1% | 99% | 0.04 | 0.048 | 93% | 7% | 0.32 |
| $\text{bias}^S_{\text{PE}}$ | 0.092 | 0% | 100% | 0.09 | 0.025 | 62% | 38% | 0.99 |
| $\text{bias}^S_{\text{CALI}}$ | 0.291 | 46% | 54% | 0.35 | 0.299 | 58% | 42% | 0.26 |




(a) $\text{bias}^{\mathcal{U}}_{\text{EO}}$ (b) $\text{bias}^S_{\text{EO}}$

Figure 1: Equal opportunity biases $\text{bias}^{\mathcal{U}}_{\text{EO}}$ and $\text{bias}^S_{\text{EO}}$ of the logistic regression model trained on the Adult dataset. Each of the biases is equal to the area under the curve of the true positive rate difference. The area is colored according to the group for which the bias part is favorable.

proves all biases of LogR except $\text{bias}_{\text{ROC}}$ but separation biases are still against female while the calibration bias of the debiased model is predominantly in favor of female. XGB outperforms the logistic regression model that was trained on the same data in terms of fairness. In half of the cases, the bias of the XGB model is even smaller than the bias of logR (debiased). Here, due to the high sample size, all biases are statistically significant. We see a difference between $\text{bias}^{\mathcal{U}}$ and $\text{bias}^S$ that is due to the skewed score distributions on the imbalanced dataset (appendix Fig. C1-C3): in general rate differences in the range of low scores are weighted higher for $\text{bias}^S$ as they effect more people (Fig. 1). Note that $\text{bias}_{\text{ROC}}$ is in favor of female persons: Looking only at groupwise ROC curves ($\text{bias}_{\text{ROC}}$) suggests an advantage for females. However, female persons experience lower true- and false positive rates at every possible threshold that is chosen independently of the group, as $\text{bias}^S_{\text{EO}}$ and $\text{bias}^S_{\text{PE}}$ clearly show.

### 5.4. Synthetic Data

In order to evaluate how different metrics change when the bias changes, we make use of synthetic datasets. This allows

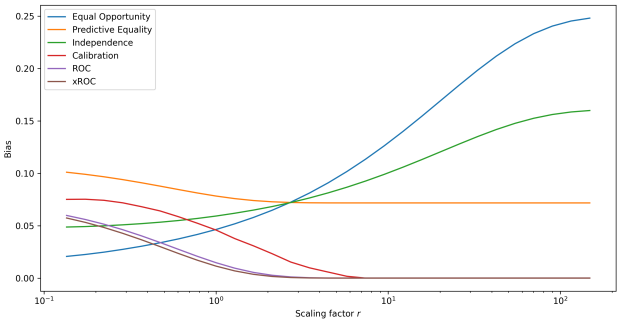



Figure 2: Changing bias measures with increasing distance between the groups and classes.

us to change the bias and observe the effect on the different metrics. For this reason, we sample $S_{a0}, S_{a1}, S_{b0}$ and $S_{b1}$ independently from four Gaussian distributions.

Utilizing a scaling factor $r > 0$, we set the following distributions: $S_{a0} \sim \mathcal{N}(1 \cdot r, 0.6^2 \cdot r)$, $S_{a1} \sim \mathcal{N}(-1, 0.5)$, $S_{b0} \sim \mathcal{N}(1.2 \cdot r, 0.75^2 \cdot r)$ and $S_{b1} \sim \mathcal{N}(-1.3, 0.6)$. Note that scores of the positive class of both groups move further away from each others with increasing $r$ (i.e. an increas-

Table 3: Gender bias of logistic regression (trained with and without sex) and XGBoost on Adult dataset; positive and negative component from the point of view of female persons. Each permutation tests gives $p < 0.01$.

| | LogR | | | LogR (debiased) | | | XGB | | |
|---|---|---|---|---|---|---|---|---|---|
| type of bias | total bias | pos. | neg. | total bias | pos. | neg. | total bias | pos. | neg. |
| $\text{bias}^{S}_{\text{EO}}$ | 0.107 | 0% | 100% | 0.069 | 0% | 100% | **0.057** | 1% | 99% |
| $\text{bias}^{S}_{\text{PE}}$ | 0.164 | 0% | 100% | **0.121** | 0% | 100% | 0.143 | 0% | 100% |
| $\text{bias}^{S}_{\text{CALI}}$ | 0.052 | 22% | 78% | **0.045** | 55% | 45% | 0.050 | 52% | 48% |
| $\text{bias}_{\text{ROC}}$ | 0.050 | 98% | 2% | 0.051 | 98% | 2% | **0.033** | 98% | 2% |
| $\text{bias}_{\text{xROC}}$ | 0.205 | 0% | 100% | 0.151 | 0% | 100% | **0.129** | 0% | 100% |
| $\text{bias}^{\mathcal{U}}_{\text{EO}}$ | 0.161 | 0% | 100% | 0.104 | 0% | 100% | **0.087** | 0% | 100% |
| $\text{bias}^{\mathcal{U}}_{\text{PE}}$ | 0.118 | 0% | 100% | **0.098** | 0% | 100% | 0.101 | 0% | 100% |
| $\text{bias}^{\mathcal{U}}_{\text{CALI}}$ | 0.105 | 20% | 80% | **0.102** | 50% | 50% | 0.138 | 62% | 38% |

ing equal opportunity bias), while the negative class stays unchanged. The effect of this increasing difference can be seen in Fig. 2.

We chose this setting to demonstrate the implications of Theorem 4.10 and Corollary 4.11. Even though the difference between $S_{a0}$ and $S_{b0}$ grows, both ROC and xROC are unable to detect this disparity.

## 6. Discussion and Outlook

In this paper, we introduced a family of standardized group disparity measures for continuous risk scores that have an intuitive interpretation and theoretical grounding based on the Wasserstein distance. We derived their relation to well-established parity concepts and to ROC-based measures and we proved, that reducing the proposed separation biases is a stronger objective than reducing ROC-based measures and, hence, is better suited to cover different sorts of bias. Moreover, we demonstrated the practical application on fairness benchmark datasets. Our results show that removing information about the attribute influences the fairness of a model and also which group is affected by it. They also show that debiasing often leads to a shift between different bias types and should be monitored carefully. XGBoost results may indicate that flexible models can produce fairer results than simpler models. The results of our experiments can serve as a starting point for a comprehensive comparison of score models (in terms of bias) and debiasing methods for such models. This work would then provide evaluation metrics for such a comparison.

The proposed measures generalize rate differences from classification tasks to entire score models. As a future extension, a generalization of rate ratios is another option that is to be explored. Moreover, the discussed decision model errors (TPR/FPR/Calibration) could be summed or related to each other (i.e., $^{\text{TPR}}/_{\text{FPR}}$) to create further disparity measures.

Note also, that the given definitions of the classifier biases are based on the $l_1$-norm. Especially when used for bias mitigation, that we did not cover here, it may also be useful to replace the $l_1$-norm by $l_p$ with $p > 1$, especially $l_2$ or $l_\infty$, to penalize large disparities more than small ones. However, the score bias is then no longer a Wasserstein-distance. Another option is to use the Wasserstein-$p$-distance with $p > 1$. Typically, the outcome of fairness analyses is to assess whether certain groups are discriminated against by a score model. All the proposed disparity measures can be used to assess the group disparity of the errors made by the model. While parity, i.e. a small bias, can be taken as a sign that there is no algorithmic unfairness in a sample with respect to a particular type of error, not all disparities are discriminatory. For practical applications we propose not to use hard thresholds to decide whether a model is fair or unfair. If needed, such thresholds can be chosen similarly to the thresholds for classification biases and should be task-specific. Once a high bias is detected, the causes of the disparities should be analyzed in detail to decide for follow-up actions. The relation to the field of causal fairness criteria (i.e. (Nilforoshan et al., 2022; Zhang & Bareinboim, 2018a; Makhlouf et al., 2020)) is out of scope of this manuscript. Further studies should investigate the relation and how they can be used to perform follow-up analyses in case of significant group disparities.

## Impact Statement

This paper extends the existing ways of measuring bias in the context of continuous scores. The aim is to report existing bias, particularly in situations where the score itself must be considered, such as credit scores, rather than just a binary decision based on it. This work has the potential to contribute to the discussion of bias in scoring systems and lead to the development and use of fairer, bias-reduced scores.

## References


Agarwal, A., Dudík, M., and Wu, Z. S. Fair regression: Quantitative definitions and reduction-based algorithms. *ArXiv*, abs/1905.12843, 2019. URL https://api.semanticscholar.org/CorpusID:170079300.

Alvarado, O. and Waern, A. Towards Algorithmic Experience: Initial Efforts for Social Media Contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, Montreal QC Canada, April 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173860. URL https://dl.acm.org/doi/10.1145/3173574.3173860.

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

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., and Goodrow, C. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2212–2220, Anchorage AK USA, July 2019. ACM. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330745. URL https://dl.acm.org/doi/10.1145/3292500.3330745.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500, San Francisco USA, May 2019. ACM. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3317593. URL https://dl.acm.org/doi/10.1145/3308560.3317593.

Bucher, T. The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, 20(1): 30–44, January 2017. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2016.1154086. URL https://www.tandfonline.com/doi/full/10.1080/1369118X.2016.1154086.

Chouldechova, A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, June 2017. ISSN 2167-6461, 2167-647X. doi: 10.1089/big.2016.0047. URL http://www.liebertpub.com/doi/10.1089/big.2016.0047.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Fair Regression with Wasserstein Barycenters. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, {NIPS'20}. {Curran Associates Inc.}, June 2020. doi: 10.5555/3495724.3496338. URL http://arxiv.org/abs/2006.07286.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, June 2017. doi: 10.1145/3097983.309809. URL http://arxiv.org/abs/1701.08230.

Datta, A., Tschantz, M. C., and Datta, A. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

DiCiccio, C., Vasudevan, S., Basu, K., Kenthapadi, K., and Agarwal, D. Evaluating Fairness Using Permutation Tests. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1467–1477, August 2020. URL https://doi.org/10.1145/3394486.3403199.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness Through Awareness. *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012. doi: 10.1145/2090236.2090255. URL http://arxiv.org/abs/1104.3913.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature21056. URL http://www.nature.com/articles/nature21056.

Gardner, J., Brooks, C., and Baker, R. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pp. 225–234, Tempe AZ USA, March 2019. ACM. ISBN 978-1-4503-6256-6. doi: 10.1145/3303772.3303791. URL https://dl.acm.org/doi/10.1145/3303772.3303791.

Han, X., Jiang, Z., Jin, H., Liu, Z., Zou, N., Wang, Q., and Hu, X. Retiring $\Delta$DP: New distribution-level metrics for demographic parity. In *arXiv:2301.13443*. arXiv, January 2023. doi: 10.48550/arXiv.2301.13443. URL http://arxiv.org/abs/2301.13443.

Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 29, pp. 9, 2016. doi: 10.5555/3157382.3157469.

Holstein, K., McLaren, B. M., and Aleven, V. Student Learning Benefits of a Mixed-Reality Teacher Awareness Tool in AI-Enhanced Classrooms. In Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H. U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., and Du Boulay, B. (eds.), *Artificial Intelligence in Education*, volume 10947, pp. 154–168. Springer International Publishing, Cham, 2018. ISBN 978-3-319-93842-4 978-3-319-93843-1. doi: 10.1007/978-3-319-93843-1_12. URL http://link.springer.com/10.1007/978-3-319-93843-1_12.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., and Sarro, F. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 2023.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. 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*, pp. 862–872. PMLR, 22–25 Jul 2020.

Kallus, N. and Zhou, A. The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the XAUC Metric. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-Aware Classifier with Prejudice Remover Regularizer. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Flach, P. A., De Bie, T., and Cristianini, N. (eds.), *Machine Learning and Knowledge Discovery in Databases*, volume 7524, pp. 35–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-33485-6 978-3-642-33486-3. doi: 10.1007/978-3-642-33486-3_3. URL http://link.springer.com/10.1007/978-3-642-33486-3_3.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *arXiv:1609.05807*, 2018. URL http://arxiv.org/abs/1609.05807.

Köchling, A. and Wehner, M. C. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development. *Business Research*, 2020.

Kozodoi, N., Jacob, J., and Lessmann, S. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297 (3):1083–1094, 2022.

Kumar, A., Liang, P., and Ma, T. Verified Uncertainty Calibration. In *Advances in Neural Information Processing SystemsNeurIPS*, volume 32. arXiv, 2019. URL http://arxiv.org/abs/1909.10155.

Kwegyir-Aggrey, K., Santorella, R., and Brown, S. M. Everything is Relative: Understanding Fairness with Optimal Transport. *arXiv:2102.10349 [cs]*, February 2021. URL http://arxiv.org/abs/2102.10349.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1):3–3, 2016.

Liu, L. T., Simchowitz, M., and Hardt, M. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pp. 4051–4060. arXiv, January 2019. URL http://arxiv.org/abs/1808.10013.

Makhlouf, K. and Zhioua, S. On the Applicability of Machine Learning Fairness Notions. *ACM SIGKDD Explorations Newsletter*, 1(23):14–23, 2021. URL https://doi.org/10.1145/3468507.3468511.

Makhlouf, K., Zhioua, S., and Palamidessi, C. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

Miroshnikov, A., Kotsiopoulos, K., Franks, R., and Kannan, A. R. Model-agnostic bias mitigation methods with regressor distribution control for Wasserstein-based fairness metrics. *arXiv:2111.11259 [cs, math]*, November 2021. URL http://arxiv.org/abs/2111.11259.

Miroshnikov, A., Kotsiopoulos, K., Franks, R., and Ravi Kannan, A. Wasserstein-based fairness interpretability framework for machine learning models. *Machine Learning*, 111(9):3307–3357, September 2022. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-022-06213-9. URL https://link.springer.com/10.1007/s10994-022-06213-9.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163, March 2021. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-042720-125902. URL https://www.annualreviews.org/doi/10.1146/annurev-statistics-042720-125902.

Nilforoshan, H., Gaebler, J. D., Shroff, R., and Goel, S. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pp. 16848–16887. PMLR, 2022.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Rader, E. and Gray, R. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 173–182, Seoul Republic of Korea, April 2015. ACM. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702174. URL https://dl.acm.org/doi/10.1145/2702123.2702174.

Saravanakumar, K. K. The Impossibility Theorem of Machine Fairness – A Causal Perspective, January 2021. URL http://arxiv.org/abs/2007.06024.

Schefzik, R., Flesch, J., and Goncalves, A. Fast identification of differential distributions in single-cell RNA-sequencing data with waddR. *Bioinformatics*, 37(19): 3204–3211, October 2021. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btab226. URL https://academic.oup.com/bioinformatics/article/37/19/3204/6207964.

Shwartz-Ziv, R. and Armon, A. Tabular data: Deep learning is not all you need. *CoRR*, abs/2106.03253, 2021. URL https://arxiv.org/abs/2106.03253.

Vogel, R., Bellet, A., and Clémençon, S. Learning Fair Scoring Functions: Bipartite Ranking under ROC-based Fairness Constraints. In *arXiv:2002.08159*, pp. 784–792, February 2021. URL http://arxiv.org/abs/2002.08159.

Wei, S., Liu, J., Li, B., and Zha, H. Mean parity fair regression in rkhs. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4602–4628. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/wei23a.html.

Yang, Z., Ko, Y. L., Varshney, K. R., and Ying, Y. Minimax auc fairness: Efficient algorithm with provable convergence. *arXiv preprint arXiv:2208.10451*, 2022.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180, Perth Australia, April 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4913-0. doi: 10.1145/3038912.3052660. URL https://dl.acm.org/doi/10.1145/3038912.3052660.

Zhang, J. and Bareinboim, E. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018a.

Zhang, J. and Bareinboim, E. Equality of Opportunity in Classification: A Causal Approach. 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., 2018b. URL https://proceedings.neurips.cc/paper/2018/file/ff1418e8cc993fe8abcfe3ce2003e5c5-Paper.pdf.

Zhao, H. Costs and benefits of fair regression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=v6anjyEDVW.

# A. Background definitions and results

## A.1. Wasserstein-p-Distance

**Definition A.1** (Wasserstein-p-Distance)**.** The $p^{\text{th}}$ Wasserstein distance between two probability measures $\mu$ and $\nu$ in $\mathcal{P}_p(\mathbb{R}^d)$ is defined as

$$W_p(\mu,\nu) := \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} d(x,y)^p \,\mathrm{d}\gamma(x,y)\right)^{1/p}, \tag{27}$$

where $\Gamma(\mu,\nu)$ denotes the collection of all measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mu$ and $\nu$ on the first and second factors respectively.

**Corollary A.2.** The Wasserstein metric may be equivalently defined by

$$W_p(\mu,\nu) = \left(\inf \mathbb{E}\left[d(X,Y)^p\right]\right)^{1/p}, \tag{28}$$

where $\mathbb{E}[Z]$ denotes the expected value of a random variable $Z$ and the infimum is taken over all joint distributions of the random variables $X$ and $Y$ with marginals $\mu$ and $\nu$ respectively.

If $d = 1$, the Wasserstein distance has a closed form. For this special case, we define $W$ as a measure between two random variables.

**Corollary A.3.** Let $X$ and $Y$ be two random variables on $\mathbb{R}$ and let $F_X$ and $F_Y$ denote their cumulative distribution functions. Then

$$W_p(X,Y) = \left(\int_0^1 |F_X^{-1}(s) - F_Y^{-1}(s)|^p \, ds\right)^{\frac{1}{p}} \tag{29}$$

**Proposition A.4.** Properties of the Wasserstein-Distance for $d = 1$:

1. For any real number $a$, $W_p(aX, aY) = |a|W_p(X,Y)$.
2. For any fixed vector $x$, $W_p(X + x, Y + x) = W_p(X,Y)$.
3. For independent $X_1,\ldots,X_n$ and independent $Y_1,\ldots,Y_n$,

$$W_p\big(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\big) \le \sum_{i=1}^n W_p(X_i,Y_i).$$

## A.2. Special case: One-dimensional Wasserstein-1-Distance

**Corollary A.5.** If $p = 1$ and $X, Y$ are random variables on $\mathbb{R}$ with cumulative distribution functions $F_X$ and $F_Y$, then

$$W_1(X,Y) = \int_0^1 |F_X^{-1}(p) - F_Y^{-1}(p)|dp \tag{30}$$

$$= \int_{\mathbb{R}} |F_X(t) - F_Y(t)|dt. \tag{31}$$

**Remark.** The Wasserstein-1-distance is not invariant under monotone transformations (for instance, under scale tranformations).

**Remark.** The Wasserstein distance is insensitive to small wiggles. For example if $P$ is uniform on $[0,1]$ and $Q$ has density $1 + \sin(2\pi kx)$ on $[0,1]$ then their Wasserstein distance is $\mathcal{O}(1/k)$.

**Theorem A.6** (lower bound of $W_1$)**.** The Wasserstein-distance is always greater or equal to the distance of the means:

$$W_1(X,Y) \ge |\mathbb{E}[X] - \mathbb{E}[Y]| \tag{32}$$

*Proof.* By Jensen inequality, as norm is convex. □

**Theorem A.7** (upper bound of $W_1$)**.** For integers $p \leq q$,

$$W_p(X, Y) \leq W_q(X, Y), \tag{33}$$

especially

$$W_1(X, Y) \leq W_q(X, Y) \quad \forall q \geq 1. \tag{34}$$

*Proof.* By Jensen inequality, as $z \to z^{q/p}$ is convex. □

### A.3. Wasserstein-Distance of Quantile-Transformed Variables

**Definition A.8** (Quantile-Transformed Wasserstein Distance)**.** Let $X, Y, Z$ be random variables on $\mathbb{R}$ and let $F_X, F_Y, F_Z : \mathbb{R} \to [0, 1]$ denote their distribution functions and $f_Z$ denote the density of $Z$. The (by Z) quantile-transformed Wasserstein Distance is then given by:

$$\begin{aligned} W_Z(X, Y) &:= W_1(F_Z(X), F_Z(Y)) && (35)\\ &= \int_0^1 \left|F_{F_Z(X)}(t) - F_{F_Z(Y)}(t)\right| \, dt && (36)\\ &= \int_0^1 \left|F_X(F_Z^{-1}(t)) - F_Y(F_Z^{-1}(t))\right| \, dt && (37)\\ &= \int_{\mathbb{R}} |F_X(s) - F_Y(s)| \, f_Z(s) \, ds && (38) \end{aligned}$$

**Proposition A.9.** Properties of the quantile-transformed Wasserstein-distance

1. For any real number $a \neq 0$, $W_Z(aX, aY) = W_{Z/|a|}(X, Y)$.
2. For any fixed vector $x$, $W_Z(X + x, Y + x) = W_{Z-x}(X, Y)$.

**Remark.** The quantile-transformed Wasserstein-1-distance is invariant under monotone transformations, for instance, under scale tranformations: For $a > 0$:

$$W_Z(X, Y) = W_{aZ}(aX, aY). \tag{39}$$

### A.4. Pushforward

The pushforward of a measure along a measurable function assigns to a subset the original measure of the preimage under the function of that subset.

**Definition A.10.** Let $(X_1, \Sigma_1)$ and $(X_2, \Sigma_2)$ be two measurable spaces, $f : X_1 \to X_2$ a measurable function and $\mu : \Sigma_1 \to [0, \infty]$ a measure on $(X_1, \Sigma_1)$. The pushforward of $\mu$ is defined as

$$f\#\mu : \Sigma_2 \to [0, \infty], f\#\mu(A) = \mu(f^{-1}(A)) \, \forall A \in \Sigma_2 \tag{40}$$

**Corollary A.11.** Let again $(X_1, \Sigma_1)$ and $(X_2, \Sigma_2)$ be two measurable spaces, $f : X_1 \to X_2$ a measurable function and $\mu : \Sigma_1 \to [0, \infty]$ a measure on $(X_1, \Sigma_1)$. If $g$ is another measurable function on $X_2$, then

$$\int_{X_2} g \circ f \, d\mu = \int_{X_1} g \, d(f\#\mu) \tag{41}$$

## B. Complete proofs

**Lemma B.1.** If we quantile-transform a continuous random variable $X \in \mathbb{R}$ by its own distribution $F_X$, the result will follow a uniform distribution in $[0, 1]$:

$$F_X(X) \sim \mathcal{U}[0, 1], \text{ so } F_{F_X}(x) = x. \tag{42}$$

**Lemma B.2.** Let $X, Y$ be two random variables in $\mathbb{R}$ with cumulative distribution functions $F_X, F_Y$. The cumulative distribution function of a random variable $Z = F_X(Y)$ is given by $F_Y(F_X^{-1}(z))$:

$$\begin{aligned} F_{F_X(Y)}(z) = F_Z(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(F_X(Y) \leq z) \\ = \mathbb{P}(Y \leq F_X^{-1}(z)) = F_Y(F_X^{-1}(z)) \end{aligned} \tag{43}$$

If $F_X$ and $F_Y$ are bijective and have the same support, then

$$F_{F_X(Y)} = F^{-1}_{F_Y(X)}. \tag{44}$$

*Proof of Theorem 3.2.* For $x =$ EO:

$$\text{bias}_x^{\mathcal{U}}(S|A=a, S|A=b) = \frac{1}{|\mathcal{S}|}\int_{\mathcal{S}} |\,\text{c-bias}_x(S|A=a, S|A=b; s)|ds \tag{45}$$

$$= \frac{1}{|\mathcal{S}|}\int_{\mathcal{S}} |\mathbb{P}(S>s|A=b, Y=0) - \mathbb{P}(S>s|A=a, Y=0)|ds \tag{46}$$

$$= \frac{1}{|\mathcal{S}|}\int_{\mathcal{S}} |(1-F_{b0}(s)) - (1-F_{a0}(s))|ds \tag{47}$$

$$= \frac{1}{|\mathcal{S}|}\int_{\mathcal{S}} |F_{a0}(s) - F_{b0}(s)|ds \tag{48}$$

$$\overset{(A.5)}{=} \frac{1}{|\mathcal{S}|} \cdot W_1(S|A=a, Y=0, S|A=b, Y=0). \tag{49}$$

For $x =$ PE and $x =$ *IND* the result follows similary. (ii) follows from Theorem A.6. □

*Proof of Theorem 3.4.* For $x =$ EO:

$$\text{bias}_x^{S}(S|A=a, S|A=b) = \int_{\mathcal{S}} |\,\text{c-bias}_x(S|A=a, S|A=b; s)|f(s)ds \tag{50}$$

$$\overset{(41)}{=} \int_{\mathcal{S}} |\,\text{c-bias}_x(S|A=a, S|A=b; s)|d(F^{-1}\#\mu) \tag{51}$$

$$= \int_0^1 |\,\text{c-bias}_x(S|A=a, S|A=b; F^{-1}(p))|dp \tag{52}$$

$$\overset{(3.2)}{=} W_1(F_{F_S(S_{ay})}, F_{F_S(S_{by})}) \tag{53}$$

$$\overset{(B.2)}{=} W_1(F_{ay} \circ F_S^{-1}, F_{by} \circ F_S^{-1}) \tag{54}$$

For $x =$ PE and $x =$ *IND* the result follows similary. (ii) follows from Theorem A.6.

□

*Proof of Theorem 4.4.*

$$W_Z(X,Y) = \int_0^1 \left|F_{F_Z(X)}(s) - F_{F_Z(Y)}(s)\right| ds \tag{55}$$

$$\overset{(B.2)}{=} \int_0^1 \left|F_X(F_Z^{-1}(s)) - F_Y(F_Z^{-1}(s))\right| ds \tag{56}$$

$$= \int_0^1 \left|(1-F_X(F_Z^{-1}(1-r))) - (1-F_Y(F_Z^{-1}(1-r)))\right| dr \tag{57}$$

$$= \int_0^1 |\text{ROC}_{X,Z}(r) - \text{ROC}_{Y,Z}(r)|\, dr \tag{58}$$

□

*Proof of Theorem 4.5.* Results directly from Def. 3.3 by using the additivity of the density in (9). □

*Proof of Lemma 4.6.* Using Theorem 4.4 and $\mathrm{ROC}_{X,X}(r) = r$. □

Under additional assumptions, we can follow that a quantile-transformation by group $a$ and $b$ result in equal distances:

**Lemma B.3.** If $S_{ay}$ and $S_{by}$ have bijective cdfs and identical support, then

$$W_{S_{ay}}(S_{ay}, S_{by}) = W_{S_{by}}(S_{ay}, S_{by}) = W_{S_y}(S_{ay}, S_{by}). \tag{59}$$

*Proof of Lemma B.3.* We show more general: If $X, Y$ are two random variables on an interval $I$ in $\mathbb{R}$ with cdfs $F_X$ and $F_Y$ that are bijective on $I$

$$W_X(X, Y) = W_Y(X, Y)$$

By Lemma B.1 and by Lemma B.2, it follows

$$\begin{aligned}
W_X(X, Y) &\overset{(38)}{=} \int_0^1 \left| F_{F_X(X)}(t) - F_{F_X(Y)}(t) \right| dt && (60)\\
&\overset{B.1}{=} \int_0^1 \left| t - F_{F_X(Y)}(t) \right| dt && (61)\\
&\overset{B.2}{=} \int_0^1 \left| t - F^{-1}_{F_Y(X)}(t) \right| dt && (62)\\
&\overset{B.1}{=} \int_0^1 \left| F^{-1}_{F_Y(Y)}(t) - F^{-1}_{F_Y(X)}(t) \right| dt && (63)\\
&\overset{(31)}{=} W_Y(X, Y) && (64)
\end{aligned}$$

It follows for $Z = w_1 X + w_2 Y$:

$$\begin{aligned}
W_Z(X, Y) &= w_1 W_X(X, Y) + w_2 W_Y(X, Y) && (65)\\
&= W_X(X, Y) = W_Y(X, Y)
\end{aligned}$$

and Lemma B.3 as a special case. □

Lemma B.3 implies that quantile-transformation can under the above assumptions be performed on either of the two groups or the whole sample with the same result. Under the same assumptions, ROC, AUROC and Gini become symmetrical, i.e. $\mathrm{ROC}_{S_{ay},S_{by}} = \mathrm{ROC}_{S_{by},S_{ay}}$.

*Proof of Theorem 4.8.* Using Proposition 4.5 and Lemma 4.6:

$$\begin{aligned}
&\mathrm{bias}^S_{\mathrm{EO}}(S|A = a, S|A = b) = W_S(S_{a0}, S_{b0}) && (66)\\
&\overset{(4.5)}{=} w_{a0} W_{S_{a0}}(S_{a0}, S_{b0}) + w_{b0} W_{S_{b0}}(S_{a0}, S_{b0}) && (67)\\
&+ w_{a1} W_{S_{a1}}(S_{a0}, S_{b0}) + w_{b1} W_{S_{b1}}(S_{a0}, S_{b0})\\
&\overset{(4.6)}{=} w_{a0} \int |\mathrm{ROC}_{S_{a0},S_{b0}}(r) - r|\, dr + w_{b0} \int |\mathrm{ROC}_{S_{b0},S_{a0}}(r) - r|\, dr && (68)\\
&+ w_{a1} \int |\,\mathrm{ROC}_{S_{a0},S_{a1}}(r) - \mathrm{ROC}_{S_{b0},S_{a1}}(r)|dr\\
&+ w_{b1} \int |\,\mathrm{ROC}_{S_{a0},S_{b1}}(r) - \mathrm{ROC}_{S_{b0},S_{b1}}(r)|dr
\end{aligned}$$

For predictive equality, the result follows similarly.

□

*Proof of Theorem 4.10.* From theorem 4.8 follows by triangle-Inequality:

$$\text{bias}^S_{\text{EO}}(S|A=a,S|A=b)+\text{bias}^S_{\text{PE}}(S|A=a,S|A=b)=W_S(S_{a0},S_{b0})+W_S(S_{a1},S_{b1}) \tag{69}$$

$$\geq w_0\int|\text{ROC}_{S_{a0},S_{b0}}(r)-r|\,dr+w_1\int|\text{ROC}_{S_{a1},S_{b1}}(r)-r|\,dr \tag{70}$$

$$+\min(w_{a0},w_{a1})\cdot\text{bias}_{\text{xROC}}+\min(w_{b0},w_{b1})\cdot\text{bias}_{\text{xROC}}$$

$$\geq w_0\int|\text{ROC}_{S_{a0},S_{b0}}(r)-r|\,dr+w_1\int|\text{ROC}_{S_{a1},S_{b1}}(r)-r|\,dr \tag{71}$$

$$+2\min(w_{a0},w_{a1},w_{b0},w_{b1})\cdot\text{bias}_{\text{xROC}}$$

and also

$$\text{bias}^S_{\text{EO}}(S|A=a,S|A=b)+\text{bias}^S_{\text{PE}}(S|A=a,S|A=b)=W_S(S_{a0},S_{b0})+W_S(S_{a1},S_{b1}) \tag{72}$$

$$\geq w_0\int|\text{ROC}_{S_{a0},S_{b0}}(r)-r|\,dr+w_1\int|\text{ROC}_{S_{a1},S_{b1}}(r)-r|\,dr \tag{73}$$

$$+\min(w_{a1},w_{b1})\cdot\text{bias}_{\text{ROC}}+\min(w_{a0},w_{b0})\cdot\text{bias}_{\text{ROC}}$$

$$\geq w_0\int|\text{ROC}_{S_{a0},S_{b0}}(r)-r|\,dr+w_1\int|\text{ROC}_{S_{a1},S_{b1}}(r)-r|\,dr \tag{74}$$

$$+2\min(w_{a0},w_{a1},w_{b0},w_{b1})\cdot\text{bias}_{\text{ROC}}$$

By additionally using Corollary 4.7, we get

$$\text{bias}^S_{\text{EO}}+\text{bias}^S_{\text{PE}}=W_S(S_{a0},S_{b0})+W_S(S_{a1},S_{b1}) \tag{75}$$

$$\geq\min(w_{a0},w_{a1},w_{b0},w_{b1})\cdot(\text{bias}_{\text{ROC}}+\text{bias}_{\text{xROC}})+\frac{w_0}{2}\,\text{Gini}(S_{a0},S_{b0})+\frac{w_1}{2}\,\text{Gini}(S_{a1},S_{b1})$$

As $\frac{w_i}{2}\leq\frac{\min(w_{a0},w_{a1},w_{b0},w_{b1})}{2}$ for $i=0,1$, we can combine all weights to get

$$\text{bias}^S_{\text{EO}}+\text{bias}^S_{\text{PE}}=W_S(S_{a0},S_{b0})+W_S(S_{a1},S_{b1}) \tag{76}$$

$$\geq\min(w_{a0},w_{a1},w_{b0},w_{b1})\cdot(\text{bias}_{\text{ROC}}+\text{bias}_{\text{xROC}})+\frac{w_0}{2}\,\text{Gini}(S_{a0},S_{b0})+\frac{w_1}{2}\,\text{Gini}(S_{a1},S_{b1}) \tag{77}$$

$$\geq\frac{\min(w_{a0},w_{a1},w_{b0},w_{b1})}{2}(\text{bias}_{\text{ROC}}+\text{bias}_{\text{xROC}}+\text{Gini}(S_{a0},S_{b0})+\text{Gini}(S_{a1},S_{b1})) \tag{78}$$

Note, that if $F_{ay}$ and $F_{by}$ have identical supports and permit an inverse, then $\text{Gini}(S_{ay},S_{by})=\text{Gini}(S_{by},S_{ay})$. If this symmetry is not fulfilled, the minimum of both must be used on the right side. □

*Proof of Theorem 4.12.* Let $\text{bias}_{\text{EO}}(S|A=a,S|A=b)=0$, it follows $F_{b0}=F_{a0}$ almost everywhere. Then

$$\text{bias}_{\text{ROC}}(S|A=a,S|A=b) \tag{79}$$

$$=\int_0^1|F_{b0}(F_{b1}^{-1}(s))-F_{a0}(F_{a1}^{-1}(s))|ds \tag{80}$$

$$=\int_0^1|F_{a0}(F_{b1}^{-1}(s))-F_{b0}(F_{a1}^{-1}(s))|ds \tag{81}$$

$$=\text{bias}_{\text{xROC}}(S|A=a,S|A=b) \tag{82}$$

For predictive equality, the statement follows similarly. □

## C. Experiments

We perform experiments in python using the COMPAS dataset, the Adult dataset and the German Credit dataset. Empirical implementations of Wasserstein-distance (*scipy.wasserstein_distance*), calibration curves (*sklearn.calibration.calibration_curve*) and ROC curves (*sklearn.metrics.roc_curve*) were used.

Table C1: Bias of COMPAS score (complete table)

| type of bias | total | pos. | neg. | p-value |
|---|---|---|---|---|
| $\text{bias}^{\mathcal{S}}_{\text{EO}}$ | 0.161 | 0% | 100% | <0.01 |
| $\text{bias}^{\mathcal{S}}_{\text{PE}}$ | 0.154 | 0% | 100% | <0.01 |
| $\text{bias}^{\mathcal{S}}_{\text{CALI}}$ | 0.034 | 79% | 21% | 0.30 |
| $\text{bias}_{\text{ROC}}$ | 0.016 | 46% | 54% | 0.31 |
| $\text{bias}_{\text{xROC}}$ | 0.273 | 0% | 100% | <0.01 |
| $\text{bias}^{\mathcal{U}}_{\text{EO}}$ | 0.152 | 0% | 100% | <0.01 |
| $\text{bias}^{\mathcal{U}}_{\text{PE}}$ | 0.163 | 0% | 100% | <0.01 |
| $\text{bias}^{\mathcal{U}}_{\text{CALI}}$ | 0.037 | 78% | 22% | 0.23 |

### C.1. Statistical Testing

We perform permutation tests (DiCiccio et al., 2020; Schefzik et al., 2021) with 1000 permutations and one pseudocount to determine the statistical significance of the calculated biases under the null hypothesis of group parity. The calibration biases were calculated using 50 bins.

### C.2. Details on COMPAS experiments

Full results are shown in Table C1.

### C.3. Details on German Credit data experiments

Both models have been trained on 70% of the dataset and evaluated on the remaining samples. We used min-max-scaling on continuous features and one-hot-encoding for categorical features. Full results are shown in Table C2. As the sample size is relatively small, it happens that even the large calibration biases are not statistically significant.

### C.4. Details on Adult experiments

All three models have been trained on 70% of the dataset and evaluated on the remaining samples. We removed the feature *relationship*, which is highly entangled with *sex* through the categories *husband* and *wife* and we engineered the remaining features to merge rare categories. We used min-max-scaling on continuous features and one-hot-encoding for categorical features. Fig. C1-C3 show the score distributions of the three scores on the testset.

Table C2: Bias of models for German Credit data (complete table)

| type of bias | Model | total bias | pos. | neg. | p-value |
|---|---|---|---|---|---|
| $\text{bias}^{S}_{\text{EO}}$ | LogR | 0.083 | 1% | 99% | 0.04 |
| | LogR (debiased) | 0.048 | 93% | 7% | 0.32 |
| $\text{bias}^{S}_{\text{PE}}$ | LogR | 0.092 | 0% | 100% | 0.09 |
| | LogR (debiased) | 0.025 | 62% | 38% | 0.99 |
| $\text{bias}^{S}_{\text{CALI}}$ | LogR | 0.291 | 46% | 54% | 0.35 |
| | LogR (debiased) | 0.299 | 58% | 42% | 0.26 |
| $\text{bias}_{\text{ROC}}$ | LogR | 0.044 | 98% | 2% | 0.80 |
| | LogR (debiased) | 0.050 | 98% | 2% | 0.69 |
| $\text{bias}_{\text{xROC}}$ | LogR | 0.133 | 0% | 100% | 0.02 |
| | LogR (debiased) | 0.057 | 93% | 7% | 0.54 |
| $\text{bias}^{\mathcal{U}}_{\text{EO}}$ | LogR | 0.041 | 3% | 97% | 0.13 |
| | LogR (debiased) | 0.036 | 97% | 3% | 0.23 |
| $\text{bias}^{\mathcal{U}}_{\text{PE}}$ | LogR | 0.078 | 1% | 99% | 0.10 |
| | LogR (debiased) | 0.024 | 74% | 26% | 0.98 |
| $\text{bias}^{\mathcal{U}}_{\text{CALI}}$ | LogR | 0.246 | 40% | 60% | 0.57 |
| | LogR (debiased) | 0.225 | 75% | 25% | 0.84 |

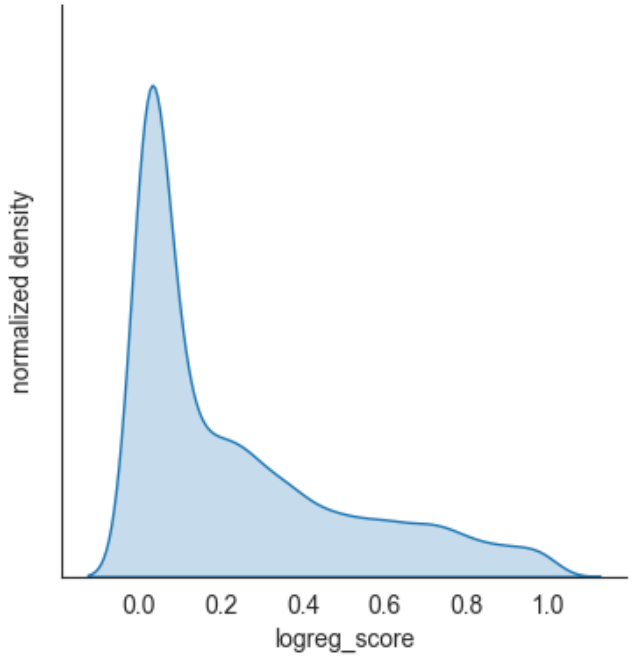



Figure C1: Distribution of logistic regression scores, trained on Adult data.

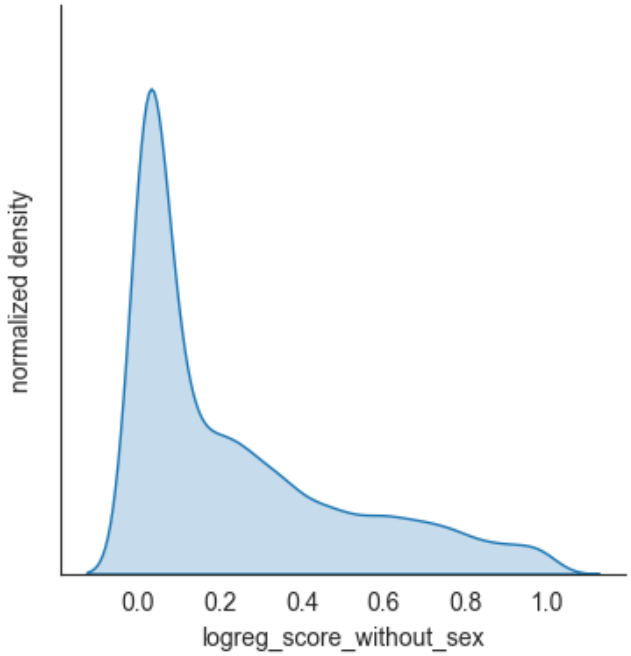



Figure C2: Distribution of logistic regression scores, trained on Adult data without protected attribute.

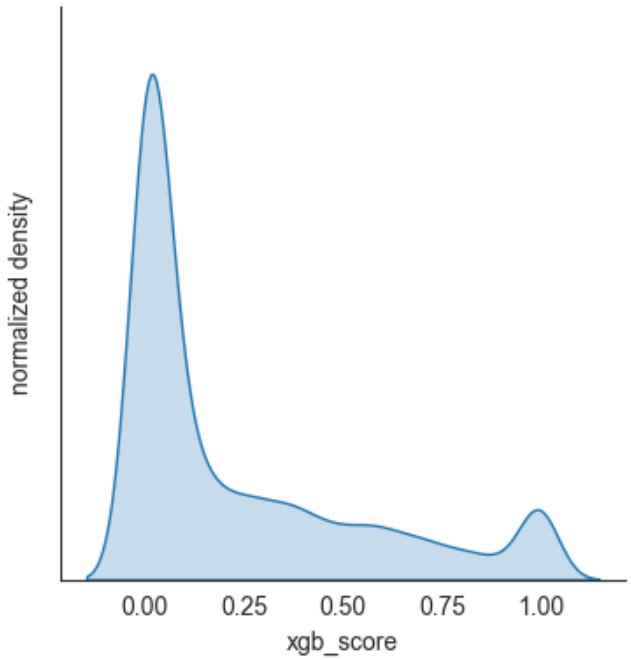



Figure C3: Distribution of XGBoost scores trained on Adult data.