Measuring Bias with The Wasserstein Distance

Kweku Kwegyir-Aggrey, Brown University, IMSD
Sarah M. Brown, University of Rhode Island

Statistical Fairness

We overview some common statistical fairness definitions:

1. **Demographic Parity**: Requires that the fraction of individuals who receive the positive classification is the same across all groups. In other words, given a protected attribute, the probability of a positive decision should be the same across groups.

$$\Pr(Y = 1 \mid A = 0) = \Pr(Y = 1 \mid A = 1)$$

2. **Equalized Opportunity**: Requires that the TPR be the same across all groups.

$$\Pr(Y = 1 \mid Y^* = 1, A = 0) = \Pr(Y = 1 \mid Y^* = 1, A = 1)$$

3. **Equalized Odds**: Requires that the TPR and the FPR be the same across all groups.

$$\Pr(Y = 1 \mid Y^* = 1, A = 0) = \Pr(Y = 1 \mid Y^* = 1, A = 1)$$
$$\Pr(Y = 1 \mid Y^* = 1, A = 0) = \Pr(Y = 1 \mid Y^* = 1, A = 1)$$

4. **Disparate Impact**: A classifier is said to have disparate impact if one group receives the positive treatment more often than the other group beyond some threshold τ.

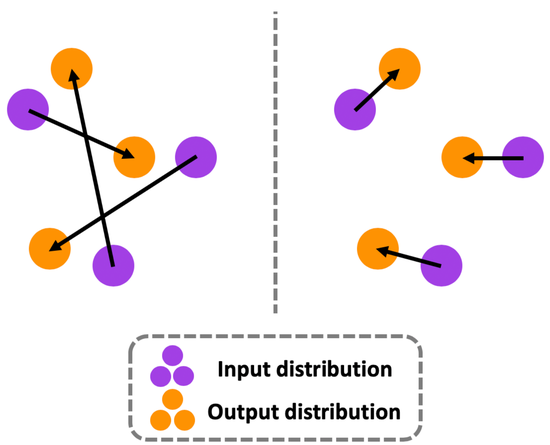
$$\frac{\Pr(Y = 1 \mid A = 0)}{\Pr(Y = 1 \mid A = 1)} \leq \tau \tag{1}$$

Wasserstein Distance

We use the Wasserstein Distance to measure distances between probability distributions on a given metric space $(\mathcal{Y}, c)$.
The *pth* Wasserstein distance between probability measures μ and ν over $\mathcal{Y}$ is given by:

$$W_p(\mu, \nu) \equiv \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{Y}\times\mathcal{Y}} c(y_1, y_2)^p d\gamma(y_1, y_2) \right)^{\frac{1}{p}}$$

where Γ is the set of couplings over distributions μ and ν. A coupling is a joint distribution over $Y \times Y$ with marginals μ and ν.

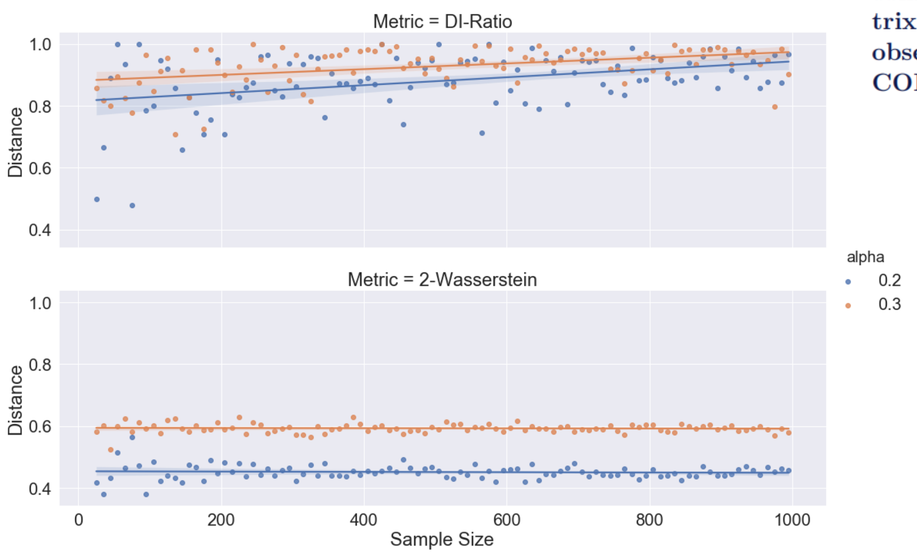


Measuring Bias

Consider the following toy example. Blue College is looking to audit their admissions data. We assume that the college can estimate the likelihood $P(Y = 1|X)$ for a given student with GPA $x \in X$ to be admitted into the college. In general Blue's applicants come from two secondary schools: expensive private School A, and public School B. Blue College would like to check that their admissions policy is not biased with respect to an applicant's school. Looking at their admissions data, Blue College observes the following:

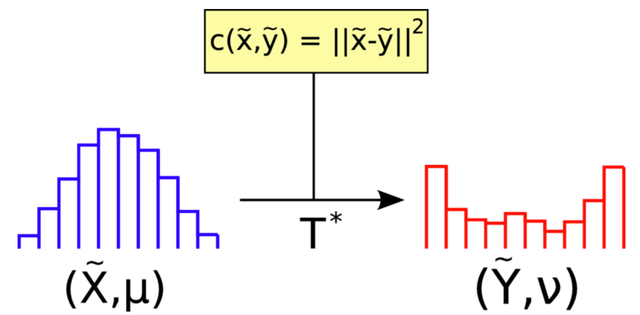
1. For students in School A, the probability of being accepted $P(Y = 1|X = x) = P(Y = 1) = 30\%$.
2. For students in School B, the probability of being accepted $P(Y = 1|X = x) = 1$ for the top 30% of applicants, and $P(Y = 1|X = x) = 0$ for the bottom 70% of applicants.

Suppose the school wishes to be fair according to the disparate impact rule. Given this generative model, we see that Blue college accepts roughly 30% of the applicants from both schools. This implies that $\frac{P(Y=1|School=A)}{P(Y=1|School=B)} = 1$, and so the model does not admit disparate impact and would be considered "fair" with respect to this fairness definition. In fact, if we assume WLOG $P(Y = 1|School = B) \leq P(Y = 1|School = A)$ then $\tau = 1$ represents the *fairest* possible outcome. Quite clearly, this model is not fair as it puts the bottom 70% of students from School B at a disadvantage, while also offering an advantage to the top 30% of students at school B.



Optimal Transport Plan

The coupling $\gamma \in \Gamma(\mu, \nu)$ also provides important information when computing the Wasserstein Distance. This transportation plan, based on the cost function $c : Y \times Y \to \mathbb{R}$ indicates which objects are being "mapped" in the optimal transport plan. In practice, we can interpret this mappings as revealing the structure of some underlying bias.



Auditing COMPAS

The COMPAS dataset represents several predictions of recidivism for some 6k arrested individuals in Southern Florida. We learn the distribution of predicted recidivism outcomes based on the COMPAS risk model, and compare this to the ground truth outcomes of the individuals in the dataset, that is, if they did indeed recidivize. We examine the structure of bias of the COMPAS model under this optimal transport framework. By computing the group-wise transport plans from the coupling matrix along several relevant dimensions of discrimination, we can observe some interesting properties of the bias present in the COMPAS risk model.