Measuring Bias with Wasserstein Distance

Kweku Kwegyir-Aggrey
Department of Computer Science
Brown University
kweku@brown.edu

Sarah M. Brown
Department of Computer Science
Rhode Island University
brownsarahm@uri.edu

Abstract

In *fair* classification, we often ask:"what does it mean to be fair, and how is fairness measured?" Previous approaches to defining and enforcing fairness rely on a set of statistical fairness definitions, with each definition providing its own unique measurement of bias. In this work, we provide preliminary results showing that the Wasserstein distance can provide a more comprehensive fairness measurement than existing methods, and can also measure fairness with respect to a broad class of fairness notions that can encapsulate the already existing fairness definitions.

1 Introduction

Paralleling the ongoing research centered on discrimination in machine learning [3, 6], is a series of critiques discussing how fair machine learning research has the potential to reinforce normative and potentially harmful ideas of fairness in their implementation [1, 9, 12]. There have been a number of *fair* machine learning solutions that have fallen short when audited for their ability to reduce real-world discrimination, in addition to satisfying some computationally tractable fairness objectives [4, 8, 11].

We examine machine discrimination in the context of historically protected attributes, such as race or gender. Indeed, there is a growing literature developing strategies for the detection and audit of discriminatory machine learning models[2, 8, 14], with respect to the aforementioned attributes. These audits typically rely on the observance of some group-level summary-statistics of a classifier or dataset. For example, the commonly cited *disparate impact* legal doctrine states that a process is biased if an advantaged group receives the positive treatment at a rate disproportionate to that of the disadvantaged group [5]. Under the disparate impact fairness notion, a classifier is fair if it assigns the positive outcome to all groups at the same rate. The difference between classification rates is then typically used as a measurement of the classifier's bias. This style of fairness measurement is not unique to disparate impact. Under the notion of equal opportunity, another popular fairness definition, we check that the group-conditioned true positive rates of a classifier (e.g. the group-wise rate at which those who deserve[1] a loan receive a loan) are the same across protected attributes. In this same context, it is also thought that the difference between true positive rates (TPRs) across groups, is a reasonable measurement for the *bias* produced by the classifier under this fairness notion. We summarize the shortcomings of this style of audit in the follow ways:

1. A classifier may appear unbiased according to one fairness definition, but could reveal bias when measured according some other definition. The question of "which fairness definition to enforce" is left to the discretion of the practitioner, as recent results from Kleinberg et al. [10] show it is generally impossible to satisfy multiple fairness notions simultaneously. Moreover there is no standard metric of bias that is common to all group fairness definitions, therefore making it difficult to interpret these measurements across different definitions.

[1] the notion of deserving is typically represent by ground truth. In this example, we would say those that *deserved* a loan, are those who repaid a loan they received

NeurIPS 2020 Workshop on Dataset Curation and Security

2. These audits can reveal discrimination across protected groups, but reveal very little about the structure of discrimination for various subgroups or at the individual level. For example, the COMPAS algorithm, a popular predictive policing tool, claims equal predictive accuracy across racial groups, but when only considering defendants who have no prior criminal history, the algorithm's accuracy claim no longer holds; under this same algorithm, the error rate for first-time black defendants is nearly 48% higher than the same rate for their white counterparts.

Herein, we provide preliminary results for a novel audit technique that addresses these shortcomings: it can reveal violations of both group and individual level fairness, while also being able to audit a classifier with respect to a flexible class of of fairness notions. We broadly describe *fairness* as a probability distribution over outcomes, and use optimal transport to study the difference between an observed distribution of outcomes and some corresponding fair distribution. We offer that the corresponding distance between distributions, as measured by the Wasserstein metric, represents the total bias of a classifier, or dataset.

2 Related Work

This work differs from existing studies at the intersection of optimal transport and fairness in a few key ways. In Flip-Test [2] the authors use optimal transport maps to test a classifier for bias. The mappings proposed by their optimal transport framework provide a "what-if" fairness report, describing what an individual's features *would be*, if they were to have their protected attribute changed. Under this comparative study, structural differences in classification between groups can be better examined. We extend their work by demonstrating that optimal transport mappings can standardize the measurement of bias for both datasets and classifiers. In addition, a number of works demonstrate the use of optimal transport as a repair method for unfair datasets [13, 15, 7].

3 Preliminaries

We study fairness in the context of binary classification. Allow $X \subseteq \mathbb{R}^d$ to be a set of covariates, $A = \{0, 1\}$ to denote the protected attribute, and $Y = \{0, 1\}$ to be some binary decision variable. In general, we say that $Y = 1$ denotes the positive classification, and $Y = 0$ denotes the negative outcome. We are interested in learning a classifier f over some dataset $D = (x_i, a_i, y_i)$. We write predictions from the model as $f(x) = \hat{Y}$. Finally, allow μ, ν to be a distribution over classification vectors $\mathcal{Y} = \{y \in \mathbb{R}^d_+ : \sum_{i=1}^d y_i\}$. In general, every y_i denotes the likelihood of an individual obtaining the i^{th} outcome. In our binary classification setup there are only two possible outcomes, however this model can easily be extended to a multi-class classification task.

4 Wasserstein Distance

We use the Wasserstein Distance to measure distances between probability distributions on a given metric space $(\mathcal{Y}, c)$.

Definition 4.1 (pth-Wasserstein Distance)**.** 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. Intuitively, we interpret the Wasserstein distance as the minimum *cost* of transforming one distribution to another, in this case, transforming μ to ν. The Wasserstein Distance, is closely related to the solution to the Monge-Kantorovich optimal transport problem, described below.

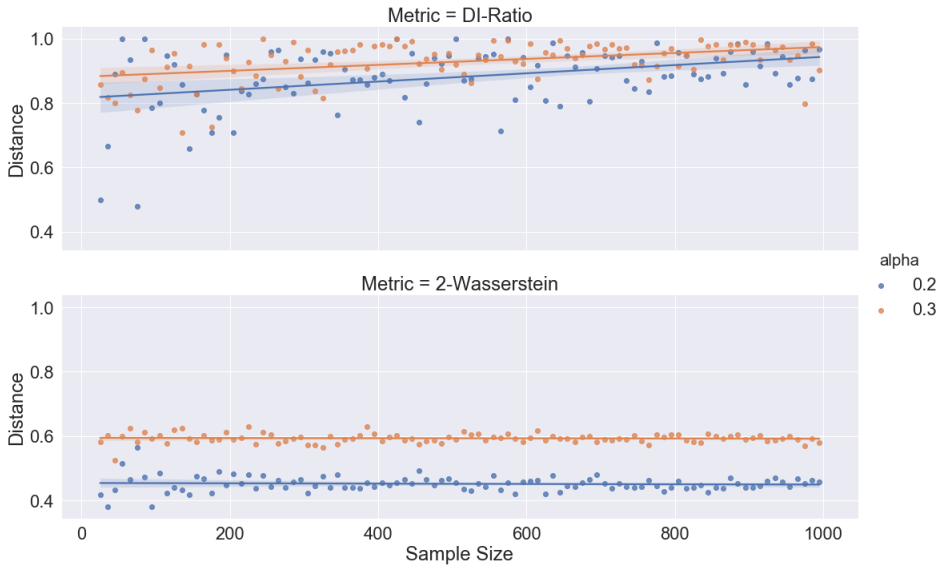

Figure 1: We compare the Wasserstein Metric to the Disparate Impact (DI) ratio for two alpha levels. The Wasserstein metric reveals less bias as we decrease alpha towards the (trivially) fair scenario when all applicants are rejected, whereas the DI-Ratio increases towards perfect fairness for both alpha levels.

4.1 Optimal Transport Maps

Consider the following discretized optimal transport problem between distributions μ, ν:

$$\gamma^* \equiv \underset{\gamma \in \Gamma(\mu,\nu)}{\arg\inf} \sum_{y_2 \in \mathcal{Y}_2} \sum_{y_1 \in \mathcal{Y}_1} c(y_1, y_2)\gamma(y_1, y_2)$$

Similarly to the Wasserstein distance, the goal of the optimal transport problem is to describe a precise plan, as to how to reconfigure the distribution objects in $\mathcal{Y}_1$, so that it "looks like" the distribution of objects in $\mathcal{Y}_2$. Recall, The cost function $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a metric specifying the *cost* of each potential pairwise transformation between objects in $\mathcal{Y}_1$ and $\mathcal{Y}_2$. Thus, we will call γ^* the optimal transport mapping; from this mapping, the Wasserstein distance can be computed. Intuitively, we may think of the mapping as a matrix $\gamma \in \mathbb{R}^{d \times d}$ where each $\gamma_{i,j}$ describes the magnitude of shift for the ith object in $\mathcal{Y}_1$ to the jth in $\mathcal{Y}_2$, in some arbitrary transportation plan.

5 Case Studies

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.

Definition 5.1 (Disparate Impact (80% Rule))**.** A model is said to admit disparate impact if:

$$\frac{P(Y = 1|School = B)}{P(Y = 1|School = A)} \leq \tau = 0.8 \tag{1}$$

In the above 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.

5.1 Auditing Disparate Impact

We evoke the above scenario in the following example. Consider a dataset where all GPAs $x \sim U(0, 4)$ but $P(Y = 1|School = A) \sim Bernouili(\alpha)$, $P(Y = 1|School = B, X > 4(1 - \alpha)) = 1$, and $P(Y = 1|School = B, X \leq 4(1 - \alpha)) = 0$ where $\alpha \in \{.2, .3\}$. We run an experiment where we generate labels according to the above model, and compare the disparate impact ratio $\frac{P(Y=1|School=A)}{P(Y=1|School=B)}$ to the Wasserstein distance. We use the ℓ^2-norm between classification vectors[2] as a cost function in our Wasserstein computation. In Figure 1 we see that as we increase the number of samples in our experiment, the disparate impact ratio increases towards one, whereas the Wasserstein distance remains relatively constant. This experiments demonstrates an unfair setting: the lower $1-\alpha$ of applicants from School B are disadvantaged compared to their School A counterparts. However, according to the disparate impact measurement, the admissions decision becomes perfectly fair as more samples are drawn, whereas the Wasserstein distance does not change.

5.2 Fair Policy

This optimal transport framework can serve as a unifying framework for studying all group fairness definitions. For example, consider equal opportunity:

Definition 5.2. Equalized Opportunity We say that a generative model $P(Y|X)$ satisfies equal opportunity, if for ground truth label Y^* the following holds

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

In order to measure bias with respect to equal opportunity, one could construct a transport mapping between individuals in protected attribute groups with ground truth label $Y^* = 1$ that were assigned the positive classification. Recall, we assume that one group can be designated as the advantaged group, and the other as the disadvantaged group. The Wasserstein distance measures the difference in outcomes between the advantaged group and the disadvantaged group. In a realistic scenario, whenever a machine learning practitioner can designate a sample of individuals who are considered to have been treated fairly, with this optimal transport framework, the Wasserstein distance from this sample to any other sample can be computed, thus revealing the difference between the fair treatment group to the observed group.

6 Future Work

We've shown evidence that the Wasserstein distance can be used to measure fairness with respect to several existing fairness definitions. In section 5.1 we also show that the Wasserstein distance can reveal bias that is often hidden by current measurements. In future work, we would like to explore the following:

1. An optimal transport map reveals pairs of individual where the individuals have similar classification outcomes. The Wasserstein metric, when considered for individual points, can also provide insight into violations in individual fairness as we can compare the distance in feature space of two individuals who have similar outcomes in the classification space.
2. The optimal transport map can be used to extend notions of subgroup fairness to the feature space. Indeed, we may observe which subsets of the feature space are being discriminated against, and study how this relates to some observed discrimination with respect to a protected attribute.

[2] an m-dimensional classification vector is a vector $y \in \mathbb{R}^m$ such that $\sum_{i=1}^m y_i = 1$.

References

[1] Rediet Abebe and Maximilian Kasy. Fairness, equality, and power in algorithmic decision-making, Oct 2020.

[2] Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[4] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[5] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[6] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016.

[7] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365, 2019.

[8] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[9] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, 10:113–174, 04 2019.

[10] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016.

[11] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. Does mitigating ml's impact disparity require treatment disparity?, 2019.

[12] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019.

[13] Chiappa Silvia, Jiang Ray, Stepleton Tom, Pacchiano Aldo, Jiang Heinrich, and Aslanides John. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3633–3640, 2020.

[14] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel J. Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Discovering unwarranted associations in data-driven applications with the fairtest testing toolkit. *CoRR*, abs/1510.02377, 2015.

[15] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery*, 34(1):163–200, 2020.