A General Approach to Fairness with Optimal Transport

Silvia Chiappa,[1*] **Ray Jiang,**[1*] **Tom Stepleton,**[1*] **Aldo Pacchiano,**[2] **Heinrich Jiang,**[3] **John Aslanides**[1]
[1]DeepMind London, [2]UC Berkeley, [3]Google Research
{csilvia, rayjiang, stepleton, heinrichj, jaslanides}@google.com, pacchiano@berkeley.edu

Abstract

We propose a general approach to fairness based on transporting distributions corresponding to different sensitive attributes to a common distribution. We use optimal transport theory to derive target distributions and methods that allow us to achieve fairness with minimal changes to the unfair model. Our approach is applicable to both classification and regression problems, can enforce different notions of fairness, and enable us to achieve a Pareto-optimal trade-off between accuracy and fairness. We demonstrate that it outperforms previous approaches in several benchmark fairness datasets.

Introduction

Data used to train machine learning systems often contain human and societal biases that can lead to treat individuals unfavorably (*unfairly*) on the basis of characteristics such as race, gender, disabilities, etc. (referred to as *sensitive attributes*). This has motivated researchers to investigate techniques for ensuring that learned models satisfy fairness properties (Dwork et al. 2012; Feldman et al. 2015; Goh et al. 2016; Chouldechova 2017; Corbett-Davies et al. 2017; Gajane and Pechenizkiy 2017; Kusner et al. 2017; Cotter et al. 2018; Mitchell, Potash, and Barocas 2018; Verma and Rubin 2018; Zhang and Bareinboim 2018; Chiappa and Isaac 2019; Narasimhan et al. 2020). Most often, fairness desiderata are expressed as constraints on the lower order moments or other functions of distributions corresponding to different sensitive attributes. Whilst facilitating model evaluation and design, not accounting for the full shapes of the relevant distributions can be restrictive and problematic (Simoiu, Corbett-Davies, and Goel 2017).

In Jiang et al. (2019), we introduced an approach to fair classification that uses optimal transport theory to match the distributions of the model outputs corresponding to different sensitive attributes to a common distribution. We demonstrated that using the Wasserstein-1 barycenter as common distribution incurs in minimal changes to the predictions obtained by the unfair model. In this paper, we generalize this work to nonlinear models and to regression, and to transporting either in the input or output space to be able to achieve different fairness criteria. The proposed approach has three main properties: it accounts for the full shapes of the distributions, it achieves fairness through minimal deviation from the unfair model, and it is applicable to different fairness criteria. We evaluate its performance on several benchmark fairness datasets.

Related Work

There has been an increasing interest in the use of optimal transport for fairness (Black, Yeom, and Fredrikson 2019; Del Barrio et al. 2019; Jiang et al. 2019; Johndrow and Lum 2019; Risser et al. 2019; Wang, Ustun, and Calmon 2019). With the exception of Jiang et al. (2019) and Risser et al. (2019), optimal transport has been used to transport model inputs and, with the exception of Wang, Ustun, and Calmon (2019) and Black, Yeom, and Fredrikson (2019), in order to achieve strong demographic parity. More specifically, Del Barrio et al. (2019) consider a binary classification/binary sensitive attribute setting, and provide an upper bound to the difference between the minimal risks obtained by the best classifier with original and transported model inputs. The work in Johndrow and Lum (2019) suggests using the Wasserstein distance to transport model inputs to a common distribution, whose choice is left to the user. Wang, Ustun, and Calmon (2019) show how to design target distributions to achieve different fairness criteria, and use optimal transport as a way to match to the target distribution. Instead, our target distribution is always chosen to achieve statistical independence, but in a way that allows as little model deviation as possible through optimal transport. Risser et al. (2019) uses the formulation of the Wasserstein-2 distance in terms of cumulative distribution functions (as in our geodesic method) to propose an efficient penalty method for non-linear classifiers in the binary sensitive attribute setting. Our penalty method is more general, as it applies to both classifications and regression, to general cost functions, and to non-binary sensitive attributes. Finally, Black, Yeom, and Fredrikson (2019) uses optimal transport to create a black-box technique for uncovering discrimination in classifiers using the Wasserstein-1 distance.

[*]Equal contribution.

Fairness Criteria for Classification and Regression

We consider the problem of learning a binary classification or regression model from a dataset $\mathcal{D} = \{(a^n, x^n, y^n)\}_{n=1}^N$. Each datapoint (a^n, x^n, y^n) contains a continuous or categorical outcome y^n of an individual (or community) that we wish to predict, a vector of attributes $a^n \in \mathcal{A} = \mathbb{N}^k$ which are considered sensitive, where element a_i^n might correspond *e.g.* to gender, and a vector of features $x^n \in \mathbb{R}^d$ to be used, possibly together with a^n, to form a prediction $\hat{y}^n$ of y^n. Our goal is to introduce an approach to impose fairness constraints on the model that is applicable to different fairness criteria. We introduce some of them below.

To treat classification and regression, as well as probabilistic and deterministic modelling, in an unified way, we formulate classification and regression as the task of estimating the probability distribution $p(Y|A = a^n, X = x^n)$, where A, X and Y are the random variables corresponding to a^n, x^n, and y^n respectively, and assume that the model outputs the expectation

$$s^n = \mathbb{E}_{\bar{p}(Y|A=a^n, X=x^n)}[Y]\,,$$

where $\bar{p}$ indicates the estimate of p (below we omit this distinction to simplify the notation). Notice that in the classification case $s^n = p(Y = 1|A = a^n, X = x^n)$, *i.e.* s^n is the model estimated probability that individual n belongs to class 1. A prediction $\hat{y}^n$ of y^n is then obtained as $\hat{y}^n = s^n$ for the regression case, and as $\hat{y}^n = \mathbb{1}_{s^n>\tau}$ for the classification case, where $\mathbb{1}_{s^n>\tau} = 1$ if $s^n > \tau$ for a threshold $\tau \in [0, 1]$, and zero otherwise. We call the random variable S, corresponding to s^n, the *output variable*, and denote with S_a the output variable restricted to the group of individuals with sensitive attributes a, *i.e.* with distribution $p(S_a) = p(S|A = a)$.

Strong demographic parity. The simplest and most popular fairness criterion, called *demographic parity*, requires the expectation of $\hat{Y}$ to not depend on A, *i.e.*

$$\mathbb{E}_{p(\hat{Y}|A=a)}[\hat{Y}] = \mathbb{E}_{p(\hat{Y}|A=\bar{a})}[\hat{Y}],\ \forall a, \bar{a} \in \mathcal{A}\,.$$

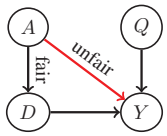
We are interested in a similar but stronger criterion, *i.e.* considering the full shape of the distribution $p(S|A)$, called *strong demographic parity* (SDP) (Jiang et al. 2019). SDP requires statistical independence between S and A, *i.e.*

$$\text{SDP: } p(S_a) = p(S_{\bar{a}}),\ \forall a, \bar{a} \in \mathcal{A}\,.$$

Notice that, for classification, SDP ensures that the class prediction does not depend on the sensitive attribute regardless of the value of the threshold τ used.

Strong path-specific fairness. SDP requires removal of the dependence of Y on A from the prediction $\hat{Y}$. In many cases, we might want to remove only some of the influence that A has on $\hat{Y}$. For example, consider a hypothetical college admission scenario in which applicants are admitted based on qualifications Q, choice of department D, and gender A; and in which female applicants apply spontaneously more often to departments with lower admission rates. Whilst the direct influence of A on admission Y, *i.e.* treating a female applicant and a male applicant with the same qualifications and applying to the same department differently, is unfair, rejecting female applicants more often due to department choice could be considered fair with respect to the college's responsibility. In such a case, we want a fairness criterion that formalizes the requirement that only the unfair influence should be absent from the prediction. This can be obtained using the causal Bayesian networks framework as described below (for a more complete explanation, we refer the reader to the Supplementary Material and to Chiappa and Isaac (2019)).

Causal Bayesian Network (CBN): *A CBN is a directed acyclic graph, with nodes representing random variables and links representing statistical dependence among them, which describes the mechanism underlying the data generation process.*
If Y is an descendant of A, i.e. if there exists a directed — also called causal — path from A to Y (namely a sequence of linked nodes starting at A and ending at Y where links are directed and pointing from preceding towards following nodes in the sequence), then A is a cause of Y.
The causal effect of $A = a$ on Y can be seen as the conditional distribution of Y given $A = a$ restricted to causal paths. We indicate with $Y_{\rightarrow a}$ the random variable with such a conditional distribution — often called the potential outcome variable.



The college admission scenario can be described by the CBN on the left, with joint distribution $p(A, Q, D, Y)$ factorizing as $p(Y|A, Q, D)p(D|A)p(Q)p(A)$, and where the direct causal path $A \rightarrow Y$ represents the direct unfair influence of gender A on admission Y, whilst the indirect causal path $A \rightarrow D \rightarrow Y$ represents the fair influence of A on Y through department choice D.

Potential outcome variables can be used to formalize the requirement that the influence of A along the path $A \rightarrow Y$ should be absent from the model, by setting A to different values along $A \rightarrow Y$ and $A \rightarrow D \rightarrow Y$. More specifically, if we indicate with a and $\bar{a}$ the female and male sensitive attribute respectively, and with $S_{\rightarrow\bar{a}}(D_{\rightarrow a})$ the potential outcome variable with distribution equal to the conditional distribution of S given A restricted to causal paths, with $A = \bar{a}$ along $A \rightarrow Y$ and $A = a$ along $A \rightarrow D \rightarrow Y$, the weak version of the requirement can be expressed as

$$\mathbb{E}_{p(S_{\rightarrow\bar{a}}(D_{\rightarrow a}))}[S_{\rightarrow\bar{a}}(D_{\rightarrow a})] = \mathbb{E}_{p(S_{\rightarrow a})}[S_{\rightarrow a}]\,,$$

obtaining the *path-specific fairness* criterion.

The strong version of the requirement can be expressed as

$$\text{SPSF: } p(S_{\rightarrow\bar{a}}(D_{\rightarrow a})) = p(S_{\rightarrow a})\,.$$

We denote with $s^n_{\rightarrow\bar{a}}(D_{\rightarrow a})$ the *counterfactual outcome* indicating the model estimated probability that a female applicant $\{a^n = a, q^n, d^n\}$ would have been admitted in a

counterfactual world in which she were male along $A \to Y$, *i.e.*

$$s^n_{\to\bar{a}}(D_{\to a}) = p(Y_{\to\bar{a}}(D_{\to a}) = 1 | A = a^n, Q = q^n, D = d^n)\,.$$

As the CBN has no descendant of A along unfair causal paths (Q is not a descendant of A, whilst D is a descendant of A along a fair causal path), the counterfactual outcome reduces to

$$s^n_{\to\bar{a}}(D_{\to a}) = p(Y = 1 | A = \bar{a}, Q = q^n, D = d^n)\,,$$

i.e. it is given by conditioning Y on q^n, d^n and, on A with value equal to $\bar{a}$ to account for setting A to $\bar{a}$ along $A \to Y$. This gives $p(S_{\to\bar{a}}(D_{\to a})) = \int_{Q,D} p(Y = 1 | A = \bar{a}, Q, D) p(D | A = a) p(Q)$.

In the more general case in which a CBN contains descendants of A along unfair causal paths, computing counterfactual outcomes is more challenging. If, *e.g.*, a link from A to Q were present and considered unfair, one way to obtain the counterfactual outcome $s^n_{\to\bar{a}}(Q_{\to\bar{a}}, D_{\to a})$ would be to perform a *correction* $q^n_{\bar{a}}$ of q^n to $\bar{a}$, and then compute

$$s^n_{\to\bar{a}}(Q_{\to\bar{a}}, D_{\to a}) = p(Y = 1 | A = \bar{a}, Q = q^n_{\bar{a}}, D = d^n)\,,$$

as explained in the next section and in the Supplementary Material.

Path-specific counterfactual fairness. Counterfactual outcomes can also be used to require fairness at the individual, rather than population, level, *i.e.* that female applicant $\{a^n = a, q^n, d^n\}$ obtains the same decision $\hat{y}^n$ as the one that would have been taken in a counterfactual world in which she were male along the direct path $A \to Y$ ($s^n_{\to\bar{a}}(D_{\to a}) = s^n_{\to a}$). This criterion is called *path-specific counterfactual fairness* (PSCF) (Chiappa and Isaac 2019), and can be expressed as requiring $p(Y = 1 | A = \bar{a}, Q = q^n, D = d^n) = p(Y = 1 | A = a^n, Q = q^n, D = d^n)$.

Fairness with Optimal Transport

By framing a fairness task as matching empirical distributions either in the space of model outputs (*output transportation*) or in the space of model inputs (or latent representations of the inputs) (*input transportation*), we are able to propose an approach that is applicable to different fairness criteria or to different approaches to the same criterion. Output transportation can be used to achieve SDP or SPSF. Input transportation can be used to achieve SDP through transporting the model inputs (or latent representations) — and thus could be applied to fair representation learning (Zemel et al. 2013) — or more complex criteria such as PSCF. The use of optimal transport theory for matching the distributions enables us to ensure minimal deviation from the unfair model.

For simplicity of exposition we explain output and input transportation in the context of SDP and PSCF respectively, starting with some background on optimal transport theory (Villani 2009; Peyré and Cuturi 2019).

Optimal Transport. The optimal transport problem was originally formulated as the problem of transporting a distribution to another one incurring in minimal cost (Monge 1781). Given two probability density functions (pdfs) p_X and p_Y on $\mathcal{X}$ and $\mathcal{Y}$, the set $\mathcal{T}$ of *transportation maps* from $\mathcal{X}$ to $\mathcal{Y}$ (where each transportation map $T : \mathcal{X} \to \mathcal{Y}$ satisfies $\int_{\mathcal{B}} p_Y(y)dy = \int_{T^{-1}(\mathcal{B})} p_X(x)dx$ for all measurable subsets $\mathcal{B} \subseteq \mathcal{Y}$), and a *cost function* $\mathcal{C} : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$, the optimal transport problem is that of finding the transportation map T^* that minimizes the total transportation cost, *i.e.*

$$T^* = \arg\min_{T \in \mathcal{T}} \int \mathcal{C}(x, T(x)) p_X(x) dx\,.$$

As T^* may not always exist for arbitrary p_X and p_Y, the problem was later reformulated (Kantorovich 1942) as that of finding a pdf γ^* in the set $\Gamma(p_X, p_Y)$ of pdfs on $\mathcal{X} \times \mathcal{Y}$ with marginals p_Y and p_X — called the *optimal coupling* between p_X and p_Y — such that

$$\gamma^* = \arg\min_{\gamma \in \Gamma(p_X, p_Y)} \mathbb{E}_{\gamma(X,Y)}[\mathcal{C}(X, Y)]\,.$$

Under appropriate conditions on $\mathcal{C}$, $\mathcal{W}_{\mathcal{C}}(p_X, p_Y) := \min_{\gamma \in \Gamma(p_X, p_Y)} \mathbb{E}_{\gamma(X,Y)}[\mathcal{C}(X, Y)]$ can be turned into a distance between p_X and p_Y. Specifically, if $\mathcal{X} = \mathcal{Y}$ and $\mathcal{C} = D^p$ for some distance metric $D : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $p \geq 1$, then $\mathcal{W}_{\mathcal{C}}(p_X, p_Y)^{1/p}$ is a valid distance between p_X and p_Y. When $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $\mathcal{C}(x, y) = \|x - y\|_p^p$, where $\|\cdot\|_p$ indicate the L^p norm, $\mathcal{W}_{\mathcal{C}}(p_X, p_Y)$ corresponds to the pth power of the Wasserstein-p distance and we adopt the shorthand $\mathcal{W}_p(p_X, p_Y)$ to denote it.

Output Transportation

In the output transportation for SDP, we would like to transport the distribution p_{S_a} of each group output variable S_a to a common distribution $p_{\bar{S}}$ using a transportation map T^*_a such that $T^*_a(S_a)$ remains close to S_a to retain accuracy. For regression problems this could be achieved by minimizing $\mathbb{E}_{p_{S_a}}[(S_a - T_a(S_a))^2]$, which leads to the T^*_a corresponding to $\min_{T_a \in \mathcal{T}(p_{S_a}, p_{\bar{S}})} \int (s - T_a(s))^2 p_{S_a}(s) ds = \mathcal{W}_2(p_{S_a}, p_{\bar{S}})$. Considering all groups, each weighted by its probability $p_a = p(A = a)$, we obtain that the distribution $p_{\bar{S}}$ inducing the minimal deviation from S is given by

$$p_{\bar{S}} = \arg\min_{p^*} \sum_{a \in \mathcal{A}} p_a \mathcal{W}_2(p_{S_a}, p^*)\,.$$

This distribution coincides with the Wasserstein (Wass)-2 barycenter with weights p_a. For classification problems, using instead the L^1 norm would give the Wass-1 barycenter, which induces the minimal number of class prediction changes in expectation (see Jiang et al. (2019)).

Input Transportation

We discuss input transportation in the PSCF context corresponding to the CBN on the left, in which red links indicate unfair influence from A. As M is a descendant of A along an unfair causal path, whilst L is a descendant of A along both an unfair causal path $A \to M \to L$ and fair causal path $A \to L$, the counterfactual outcome of interest for individual $\{a^n = a, c^n, m^n, l^n\}$ (for the classification case) is given by $s^n_{\to\bar{a}}(M_{\bar{a}}, L_{\to a}(M_{\bar{a}})) = p(Y_{\to\bar{a}}(M_{\bar{a}}, L_{\to a}(M_{\bar{a}})) = 1 | A =$

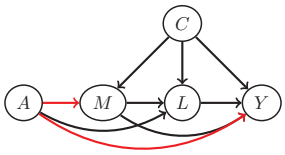


Figure 1: CBN with unfair links colored in red.

$a^n, C = c^n, M = m^n, L = l^n)$, where A is set to a counterfactual value $\bar{a}$ along the unfair paths $A \to M$ and $A \to Y$ and to the factual value a along the fair path $A \to L$. This can be estimated as

$$s^n_{\to\bar{a}}(M_{\bar{a}}, L_{\to a}(M_{\bar{a}})) \\ = p(Y = 1|A = \bar{a}, C = c^n, M = m^n_{\bar{a}}, L = l^n_a(m_{\bar{a}}))\,,$$

where $m^n_{\bar{a}}$ and $l^n_a(m_{\bar{a}})$ are *corrected* versions of m^n and l^n to $\bar{a}$ that can be computed using the procedure described below (see the Supplementary Material for more details).

Let us first assume that $m^n = f_M(A = a, C = c^n) + \epsilon^n_m$ for a function $f_M(\cdot)$ and a statistically independent latent term ϵ^n_m. The corrected version of m^n can be obtained as

$$m^n_{\bar{a}} = f_M(\bar{a}, c^n) + \epsilon^n_m \\ = f_M(\bar{a}, c^n) + m^n - f_M(a, c^n)\,.$$

Similarly, if $l^n = f_L(A = a, C = c^n, M = m^n) + \epsilon^n_l$, the corrected version of l^n can be obtained as

$$l^n_a(m_{\bar{a}}) = f_L(\bar{a}, c^n, m^n_{\bar{a}}) + \epsilon^n_l \\ = f_L(\bar{a}, c^n, m^n_{\bar{a}}) + l^n - f_L(a, c^n, m^n)\,.$$

For the more general case in which, *e.g.*, $m^n = f_M(A = a, C = c^n, \epsilon^n_m)$ for a non-linear function f_M, the counterfactual outcome can be obtained through a generalization of this procedure which uses a Monte-Carlo approximation of $s^n_{\to\bar{a}}(M_{\bar{a}}, L_{\to a}(M_{\bar{a}}))$ based on expressing it as

$$\int_\epsilon p(Y_{\to\bar{a}}(M_{\bar{a}}, L_{\to a}(M_{\bar{a}}) = 1|\epsilon) p(\epsilon|a, m^n, l^n)\,, \qquad (1)$$

where $\epsilon = (\epsilon_m, \epsilon_l)$ (see Chiappa and Isaac (2019) and Chiappa (2019)).

Crucial to the validity of this correction approach is that ϵ^n_m and ϵ^n_l must be statistically independent from A. Whilst the distributions of ϵ_m and ϵ_l satisfies $p(\epsilon_m|A) = p(\epsilon_m)$ and $p(\epsilon_l|A) = p(\epsilon_l)$ by construction, due to inaccuracies, the *empirical distributions*[1] $\hat{p}(\epsilon_m)$ and $\hat{p}(\epsilon_l)$ based on the estimates $\epsilon^n_m, \epsilon^n_l$ will most often depend on A, *i.e.*

$$\hat{p}(\epsilon_m|A = a) := \frac{1}{N_a} \sum_{n \text{ s.t. } a^n = a} \delta_{\epsilon_m = \epsilon^n_m} \neq \hat{p}(\epsilon_m|A = \bar{a})\,,$$

where N_a is he number of individuals with sensitive attributes a, and similarly for ϵ_l.

A way to maintain validity of the correction is to transport all $\hat{p}_{\epsilon_a} = \hat{p}(\epsilon|A = a)$ to a common distribution $p_{\bar{S}}$. We would like to use transportation maps T^*_a such that this process incurs in the minimal overall deviation from S. To highlight dependence on c^n and ϵ^n, we use $s(c^n, \epsilon^n)$ to indicate the model output $s^n = \mathbb{E}_{p(Y|A=a^n, C=c^n, M=m^n, L=l^n)}[Y]$. Using a similar reasoning and notation as in the Output Transportation Section, where the integral is approximated with a Monte-Carlo approach due to the use of empirical distributions, we want the T^*_a corresponding to

$$\min_{T_a \in \mathcal{T}(p_{S_a}, p_{\bar{S}})} \frac{1}{N_a} \sum_{\substack{\epsilon, \epsilon^n = \epsilon \\ n \text{ s.t. } a^n = a}} \left[s(c^n, \epsilon^n) - s(c^n, T(\epsilon^n))\right]^2,$$

where the sum is over all different ϵ in group a — we indicate this minimum as $\mathcal{W}_{\mathcal{C}_\epsilon}(\hat{p}_{\epsilon_a}, p_{\bar{S}})$. By considering all groups, each weighted by its probability $\frac{N_a}{N}$, we obtain

$$p_{\bar{S}} = \arg\min_{\hat{p}^*} \sum_{a \in \mathcal{A}} \frac{N_a}{N} \mathcal{W}_{\mathcal{C}_\epsilon}(\hat{p}_{\epsilon_a}, \hat{p}^*)\,.$$

This distribution corresponds to a barycenter with customized cost function $\mathcal{C}(\epsilon, T(\epsilon)) = \sum_{\epsilon^n = \epsilon} \left[s(c^n, \epsilon^n) - s(c^n, T(\epsilon^n))\right]^2$. The cost function aggregates over different c^n for the same value of ϵ^n, since we transport only in the ϵ space $\mathbb{R}^2$ (*i.e.* $T : \mathbb{R}^2 \to \mathbb{R}^2$) to avoid any transportation on values of c^n.

Since in the cases of interest s is a Lipschitz function of ϵ for any fixed c, we observe that $|s(c^n, \epsilon^n) - s(c^n, T(\epsilon^n))| \leq L_s \|\epsilon^n - T(\epsilon^n)\|_2$ for some universal constant L_s. Consequently $|s(c^n, \epsilon^n) - s(c^n, T(\epsilon^n))|^2 \leq L_s^2 \|\epsilon^n - T(\epsilon^n)\|_2^2$ for all ϵ^n and therefore $\mathcal{W}_{\mathcal{C}_\epsilon}(\hat{p}_{\epsilon_a}, \hat{p}^*) \leq L_s^2 \mathcal{W}_2(\hat{p}_{\epsilon_a}, \hat{p}^*)$ for all $a \in \mathcal{A}$ and fixed $\hat{p}^*$. In our experimental results we make use of this approximation, and use the Wass-2 barycenter as a computationally efficient surrogate for our customized objective.

Pareto-Optimal Fairness-Accuracy Trade-Off

We have derived barycenters as target distributions to ensure fairness whilst incurring in minimal model deviation. Given a trained model and an estimate of a barycenter, we could perform a post-processing operation by applying the derived optimal transport maps on the relevant variables (such as S_a or ϵ_a). Whilst this ensures that we retain as much accuracy as possible, in some cases we might want to trade-off a certain amount of fairness for higher accuracy. In the remainder of this section, we explain how to obtain an optimal trade-off, focusing on output transportation for SDP.

Not achieving SDP in output transportation implies that each p_{S_a} is transported to a distribution $p_{S^*_a}$ that does not match the barycenter $p_{\bar{S}}$. A valid measure of deviation from SDP is $d_{\text{pair}} = \sum_{a \neq \bar{a}} \mathcal{W}_2(p_{S^*_a}, p_{S^*_{\bar{a}}})$ since $d_{\text{pair}} = 0 \Leftrightarrow p_{S^*_a} = p_{S^*_{\bar{a}}}, \forall a, \bar{a} \in \mathcal{A}$. This measure has the additional merit of being indifferent to the choice of the resulting matched distribution that achieves SDP, and of being interpretable. For any distribution p, by the triangle inequality

[1] Throughout the paper we use $\hat{p}(\cdot)$ to indicate the empirical counterpart of $p(\cdot)$.

and Young's inequality,

$$\begin{aligned} d_{\text{pair}} &\leq \sum_{a\neq\bar{a}} \left(\sqrt{\mathcal{W}_2(p_{S_a^*}, p)} + \sqrt{\mathcal{W}_2(p, p_{S_{\bar{a}}^*})}\right)^2 \\ &\leq \sum_{a\neq\bar{a}} 2\Big(\mathcal{W}_2(p_{S_a^*}, p) + \mathcal{W}_2(p, p_{S_{\bar{a}}^*})\Big) \\ &= 4(|\mathcal{A}|-1)\sum_{a\in\mathcal{A}} \mathcal{W}_2(p_{S_a^*}, p)\,. \end{aligned}$$

By the definition of the barycenter $p_{\bar{S}}$, this upper bound reaches its minimum when $p = p_{\bar{S}}$. We call this tightest upper bound pseudo-d_{pair} and use it to derive optimal trade-off solutions.

For any $r \in \mathbb{R}_+$, we say that pseudo-d_{pair} satisfies the *r-fairness* constraint when it is smaller than r. To reach optimal trade-offs, we are interested in transporting p_{S_a} to $p_{S_a^*}$ under the *r-fairness* constraint while minimizing the model deviation from S, $\min_{p^*_{S_a}} \sum_{a\in\mathcal{A}} p_a \mathcal{W}_2(p_{S_a}, p_{S_a^*})$ (following the same derivation as in the Output Transportation Section). Assuming disjoint groups, we can optimize each group transportation in turn independently. The r-fairness constraint on a single group a becomes $\mathcal{W}_2(p_{S_a^*}, p_{\bar{S}}) \leq r' - d'$, where $r' = r/(4|\mathcal{A}|-4)$ and $d' = \sum_{\bar{a}\in\mathcal{A}\setminus\{a\}} \mathcal{W}_2(p_{S_{\bar{a}}^*}, p_{\bar{S}})$. Satisfying this constraint corresponds to transporting p_{S_a} to the ball with center $p_{\bar{S}}$ and radius $r'-d'$ in the Wass-2 metric space. To achieve the optimal trade-off, we need to transport p_{S_a} to a destination $p_{S_a^*}$ with minimal $\mathcal{W}_2(p_{S_a}, p_{S_a^*})$. Thus we want

$$\begin{aligned} p_{S_a^*} &= \underset{p^* \text{ s.t. } \mathcal{W}_2(p^*, p_{\bar{S}}) \leq r'-d'}{\arg\min} p_a \mathcal{W}_2(p_{S_a}, p^*) \\ &= \underset{p^* \text{ s.t. } \mathcal{W}_2(p^*, p_{\bar{S}}) \leq r'-d'}{\arg\min} \mathcal{W}_2(p_{S_a}, p^*) \end{aligned}$$

since p_a is constant with respect to p^*. As $\mathcal{W}_2(p_{S_a}, p^*) \geq \left(\sqrt{\mathcal{W}_2(p_{S_a}, p_{\bar{S}})} - \sqrt{\mathcal{W}_2(p^*, p_{\bar{S}})}\right)^2$ by triangle inequality, $\mathcal{W}_2(p_{S_a}, p^*)$ reaches its minimum if and only if p^* lies on a shortest path between p_{S_a} and $p_{\bar{S}}$. Therefore it is optimal to transport p_{S_a} along any shortest path between itself and $p_{\bar{S}}$ in the Wass-2 metric space.

Notice that, as the argument above only relies on $\mathcal{W}_2$ being the square of a distance, the same conclusion applies to any $\mathcal{W}_\mathcal{C}$ that is a square of a distance metric.

Methods

In this section, we introduce specific methods for implementing the fairness approach described above.

Geodesic Method. For the case of univariate Wass-p distances, we are able to propose a simple post-processing method that achieves a Pareto-optimal trade-off between accuracy and SDP by transporting p_{S_a} along any shortest path to $p_{\bar{S}}$ based on geodesics.

Geodesic: *A curve $\psi : I \to \Omega$ from an interval I of the real numbers to a metric space Ω equipped with metric D is a geodesic iff there exist $v \geq 0$ such that for $\forall t \in I$, there exists a neighborhood (t_-, t_+) of t such that $D(\psi(t_1), \psi(t_2)) = v|t_1 - t_2|$ for any $t_1, t_2 \in (t_-, t_+)$.*

A geodesic curve in Ω is therefore everywhere locally a distance minimizer.

Let the Wass-p space $\mathcal{P}_p(\mathbb{R})$ be defined as the space of all pdfs $p(\cdot)$ on the metric space $\mathbb{R}$ with finite[2] absolute p-th moments, *i.e.* $\mathbb{E}_{p(s_1)}[|s_1 - s_0|^p] < \infty$ for $\forall s_0 \in \mathbb{R}$, equipped with the Wass-p metric. As $\mathbb{R}$ is a geodesic space, *i.e.* there exists a geodesic between every pair of points in that space, then so is $\mathcal{P}_p(\mathbb{R})$ (Lisini 2007). Whilst geodesics are only locally shortest paths, shortest paths are always geodesics if they exist. In the case of $\mathcal{P}_p(\mathbb{R})$, the geodesic between p_{S_a} and $p_{\bar{S}}$ is unique and can be parametrized by

$$P^{-1}_{S_a,t} = (1-t)P^{-1}_{S_a} + tP^{-1}_{\bar{S}}, \quad t \in [0,1]\,,$$

where P_{S_a} and $P_{\bar{S}}$ are the cumulative distribution functions of S_a and $\bar{S}$ (Peyré and Cuturi 2019). This geodesic, by its uniqueness, is therefore the shortest path.

An implementation of this method is described in Algorithm 1, where $\mathcal{D}_a = \{(a^n, x^n, y^n) \in \mathcal{D} \text{ s.t. } a^n = a\}$ denotes the subset of $\mathcal{D}$ corresponding to the group of N_a individuals with sensitive attributes a. The parameter t controls the level to which p_{S_a} is moved toward the barycenter $p_{\bar{S}}$, with $t = 1$ corresponding to total matching.

Algorithm 1 Wass-p Geodesic

Input: Dataset $\mathcal{D} = \{(a^n, x^n, y^n)\}_{n=1}^N$, number of bins B, model outputs $\{s^n\}$, trade-off parameter t.
Compute group datasets $\{\mathcal{D}_a\}$ and barycenter dataset $\bar{\mathcal{D}}$.
Define the i-th quantile of $\mathcal{D}_a$, as

$$q_{\mathcal{D}_a}(i) := \sup\left\{s : \frac{1}{N_a}\sum_{n \text{ s.t. } a^n=a} \mathbb{1}_{s^n \leq s} \leq \frac{i-1}{B}\right\},$$

and its inverse as $q^{-1}_{\mathcal{D}_a}(s) := \sup\{i \in [B] : q_{\mathcal{D}_a}(i) \leq s\}$.
Define $q^{-1}_{\mathcal{D}_{a,t}}(s) := (1-t)q^{-1}_{\mathcal{D}_a}(s) + t\, q^{-1}_{\bar{\mathcal{D}}}(s)$giving

$$q_{\mathcal{D}_{a,t}}(i) = \sup\{s \in [0,1] : (1-t)q^{-1}_{\mathcal{D}_a}(s) + t\, q^{-1}_{\bar{\mathcal{D}}}(s) \leq i\}.$$

Return: $\{q_{\mathcal{D}_{a,t}}\left(q^{-1}_{\mathcal{D}_a}(s^n)\right)\}$.

Penalty Method. In the multivariate case and arbitrary $\mathcal{W}_\mathcal{C}$, the analytical geodesic computations are not feasible. When $\mathcal{X}$ and $\mathcal{Y}$ are discrete sets of cardinality n and m (as in our empirical approximation of pdfs), the optimal coupling γ^* can be identified with a $m \times n$ doubly stochastic matrix whose marginals agree with p_X and p_Y. The optimal coupling γ^* and $\mathcal{W}_\mathcal{C}$ can obtained via the following linear program:

$$\min \mathbb{E}_{\gamma(X,Y)}\left[\mathcal{C}(X,Y)\right], \text{ s.t. } \mathbf{1}_m^\top \gamma = p_Y, \gamma \mathbf{1}_n = p_X\,, \quad (2)$$

where $\mathbf{1}_m \in \mathbb{R}^m$ and $\mathbf{1}_n \in \mathbb{R}^n$ are all-ones vectors. If m, n are large, solving (2) can be computationally expensive. This issue can be addressed by regularizing (2) by an

[2]This condition is satisfied as we use empirical approximations to pdfs.

	NLSY			
	MSE	Corr	nWass2	KS
Unconstrained	1.210	1.048	4.870	1.989
Corr Penalty	1.726	0.027	**0.052**	0.327
Corr Lag	1.756	0.114	0.127	0.541
GF Lag	2.183	0.462	0.771	0.868
MMD Penalty	1.422	0.089	0.109	0.381
Wass-2 Geo (t=1)	1.374	0.079	0.090	**0.303**
Wass-2 Penalty	1.407	**0.026**	0.090	0.382
	C&C			
	MSE	Corr	nWass2	KS
Unconstrained	0.020	2.896	51.857	11.475
Corr Penalty	0.059	0.257	2.775	2.802
Corr Lag	0.048	0.518	5.330	3.608
GF Lag	0.053	2.970	35.340	9.801
MMD Penalty	0.071	0.211	0.554	1.881
Wass-2 Geo (t=1)	0.063	**0.129**	0.902	**1.820**
Wass-2 Penalty	0.073	0.187	**0.525**	1.832
	LSAC			
	MSE	Corr	nWass2	KS
Unconstrained	0.059	0.657	3.761	1.750
Corr Penalty	0.335	0.147	0.513	0.833
Corr Lag	1.104	0.674	3.233	1.765
GF Lag	1.153	0.092	0.382	0.861
MMD Penalty	0.215	0.092	0.276	0.533
Wass-2 Geo (t=1)	0.211	**0.034**	**0.036**	**0.203**
Wass-2 Penalty	0.202	0.066	0.279	0.481

Table 1: Output transportation results.

entropy term with regularization parameter λ. We denote the optimal coupling and corresponding expected cost as γ^*_λ and $\mathcal{W}^\lambda_\mathcal{C}(p_X, p_Y)$ respectively. Solving this regularized objective can be done efficiently via the Sinkhorn algorithm (Cuturi and Doucet 2014), which achieves a linear convergence rate (Altschuler, Weed, and Rigollet 2017).

The optimal coupling can be used to obtain an in-processing method that achieves a Pareto-optimal trade-off between accuracy and fairness by approximating the shortest paths between p_{S_a} and $p_{\bar{S}}$ with hyperparameter tuning of a gradient descent method. As discussed above, if the transportation cost gives rise to a distance squared, this procedure is valid (we argue that this is the case for $\mathcal{W}_{\mathcal{C}_\epsilon}$ in the Supplementary Material).

For output transportation, we propose to penalize the baseline loss $L(\theta)$ (*e.g.* the Mean Square Error (MSE) loss for regression problems) by enforcing the empirical group distributions $\hat{p}_{S_a}$ to be close to the empirical barycenter $\hat{p}_{\bar{S}}$ through the following weighted objective function:

$$L_P(\theta) = \alpha L(\theta) + \beta \sum_{a \in A} \mathcal{W}^\lambda_\mathcal{C}(\hat{p}_{S_a}(\theta), \hat{p}_{\bar{S}})\,. \qquad (3)$$

If p_X and p_Y are two distributions parametrized by θ, the gradient of $\mathcal{W}^\lambda_\mathcal{C}(p_X, p_Y)$ with respect to θ, $\nabla_\theta \mathcal{W}^\lambda_\mathcal{C}(p_X, p_Y)$,

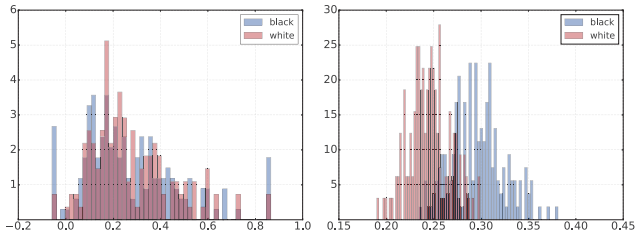


Figure 2: Histograms of the model outputs for two of the eight race groups for (left) Wass-2 Geo and (right) GF Lag.

can be computed via the chain rule

$$\nabla_\theta \mathcal{W}^\lambda_\mathcal{C}(p_X, p_Y) = [\nabla_\theta \mathcal{C}] \cdot \left[\nabla_\mathcal{C} \mathcal{W}^\lambda_\mathcal{C}(p_X, p_Y)\right]\,.$$

The gradient of $\mathcal{W}^\lambda_\mathcal{C}(p_X, p_Y)$ with respect to the cost $\mathcal{C}$ satisfies:

Lemma 1. *For any $\lambda \geq 0$, $\nabla_\mathcal{C} \mathcal{W}^\lambda_\mathcal{C}(p_X, p_Y) = \gamma^*_\lambda$.*

Proof. The proof follows the same argument as in Theorem 3 of Arjovsky, Chintala, and Bottou (2017). As a function of $\mathcal{C}$, $\mathcal{W}^\lambda_\mathcal{C}$ is a pointwise minimum of linear functions. A standard result of convex optimization states that for any function f defined as the pointwise minimum of a set of convex functions $\{f_i\}_{i \in I}$ for some possibly uncountable index set I (in other words, if for any x in the domain $f(x) = \min_{i \in I} f_i(x)$, for example a set of linear functions as in the case of $\mathcal{W}^\lambda_\mathcal{C}$), the subgradient at any point x contains the subgradient set of the function achieving the argmin. In other words, if $i^* = \arg\min_{i \in I} f_i(x)$, then $\partial f(x) \subseteq \partial f_{i^*}(x)$. In the case of differentiable functions, this is equivalent to $\nabla f(x) = \nabla f_*(x)$. As a consequence of this result, it immediately follows that $\nabla_\mathcal{C} \mathcal{W}^\lambda_\mathcal{C}(p_X, p_Y) = \gamma^*_\lambda$ as desired. □

We therefore obtain $\nabla_\theta \mathcal{W}^\lambda_\mathcal{C}(p_X, p_Y) = [\nabla_\theta \mathcal{C}] \cdot \gamma^*_\lambda$, which can be written as the expectation

$$\nabla_\theta \mathcal{W}^\lambda_\mathcal{C}(p_X, p_Y) = \mathbb{E}_{(x,y) \sim \gamma^*_\lambda} \left[\nabla_\theta \mathcal{C}(x, y)\right]\,.$$

Thus $\nabla_\theta \mathcal{W}^\lambda_\mathcal{C}(\hat{p}_{S_a}(\theta), \hat{p}_{\bar{S}}) = \mathbb{E}_{(x,y) \sim \gamma^*_{\hat{p}_{S_a}(\theta), \hat{p}_{\bar{S}}}} [\nabla_\theta \mathcal{C}(x, y)]$.

For input transportation, in the additive ϵ scenario we can replace $\sum_{a \in A} \mathcal{W}^\lambda_\mathcal{C}(\hat{p}_{S_a}(\theta), \hat{p}_{\bar{S}})$ in Eq. (3) with $\sum_{a \in A} \mathcal{W}^\lambda_{\mathcal{C}_\epsilon}(\hat{p}_{\epsilon_a}(\theta), \hat{p}_{\bar{S}})$. In most cases, learning the model parameters through this penalty term could however not be sufficient to impose independence of ϵ on A. In such cases, a parametrized function for ϵ can be learned. An alternative, more sound approach, is to use a latent variable approach as in Chiappa (2019).

Experiments

We evaluated our approach on the National Longitudinal Survey of Youth (NLSY) 1979 regression dataset, on the UCI Communities & Crime (C&C) (Lichman 2013) regression dataset, on the Law School Admission Council (LSAC) regression dataset, and on the UCI Adult binary classification dataset. As sensitive attributes, for the NLSY dataset we considered age (binned into the two categories of under and over 18 years old) and binary gender (female and

male), obtaining four groups. For the C&C dataset, we considered race (white, black, asian and hispanic), thresholded at the median, obtaining eight groups. For the LSAC dataset, we considered race (white and non-white) and binary gender, obtaining four groups. For the Adult dataset, we considered binary gender, obtaining two groups. Details about the datasets and the experiments are given in the Supplementary Material.

As the baseline objective, we used maximum log-likelihood $\mathcal{L}(\theta) = \log \prod_{n=1}^{N} p(y^n|a^n, x^n; \theta)$, obtaining the MSE loss $L(\theta) = \frac{1}{N}\sum_{n=1}^{N}(y^n - s^n)^2$ for regression (assuming $p(Y|a^n, x^n; \theta) = \mathcal{N}(s^n, 1)$), and the logistic loss $L(\theta) = -\frac{1}{N}\sum_{n=1}^{N} y^n \log(s^n) + (1-y^n)\log(1-s^n)$ where $s^n = p(Y=1|a^n, x^n; \theta)$ for classification. For PSCF, $\mathcal{L}(\theta)$ was adjusted to enable learning the conditional distributions of the CBN.

SDP through Output Transportation

We evaluated the output transportation approach for strong demographic parity on the the regression datasets. Table 1 shows test performance for $s^n = \theta^\top(x^n, a^n, 1)$ (similar results were obtained with nonlinear models). More specifically, we compare Wass-2 Geo (Algorithm 1 with $t = 1$) and Wass-2 Penalty (Eq. (3) with $L_P = \alpha L + \beta \sum_a \mathcal{W}_2(\hat{p}_{S_a}, \hat{p}_{\bar{S}})$) with the following methods:

Unconstrained: Loss L only.

Corr Penalty: $L_P = L + \beta L_C$ where L_C is the squared correlation between sensitive attributes and predictions.

Corr Lag: Regression with constraint on L_C enforced using Lagrange multipliers.

GF Lag: Group fairness method of Berk et al. (2017) with constraint enforced using Lagrange multipliers.

MMD Penalty: $L_P = \alpha L + \beta \sum_a L_{\mathrm{M}}(\hat{p}_{S_a}, \hat{p}_S)$ where $L_{\mathrm{M}}(\hat{p}_{S_a}, \hat{p}_S)$ is the maximum mean discrepancy (MMD) (Gretton et al. 2012) between the empirical distribution of the model output for group a and that for the full dataset.

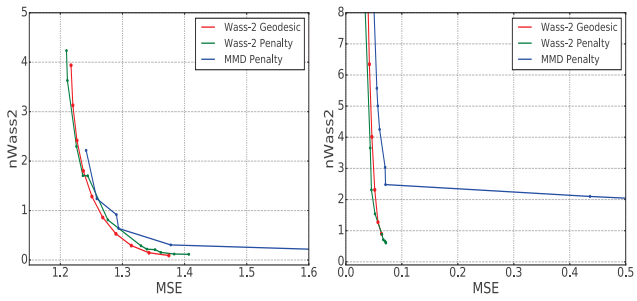
As evaluation metrics we used:

Corr: $\sum_{a\in\mathcal{A}} \left| \frac{\mathrm{cov}(S, \mathbb{1}_{A=a})}{\sigma(S)\sigma(\mathbb{1}_{A=a})} \right|$, where $\mathbb{1}_{A=a}$ indicates a random variable taking values $\mathbb{1}_{a^n=a}$ and $\sigma(\cdot)$ standard deviation.

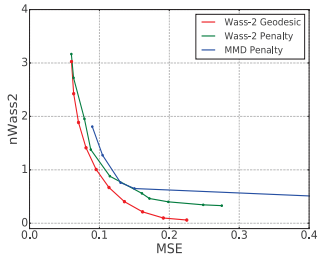
nWass2: $\frac{1}{2}\sum_{\substack{a,\bar{a}\in\mathcal{A} \\ s.t.\ a\neq\bar{a}}} \frac{\mathcal{W}_2(\hat{p}_{S_a}, \hat{p}_{S_{\bar{a}}})}{\sigma(S_a)\sigma(S_{\bar{a}})}$.

KS: $\frac{1}{2}\sum_{\substack{a,\bar{a}\in\mathcal{A} \\ s.t.\ a\neq\bar{a}}} \chi_{a\bar{a}}$, where $\chi_{a\bar{a}} = \sup_{\tau\in[0,1]} |\hat{P}_{S_a}(\tau) - \hat{P}_{S_{\bar{a}}}(\tau)|$, and $\hat{P}_{S_a}$ and $\hat{P}_{S_{\bar{a}}}$ are the empirical cumulative distribution functions for groups a and $\bar{a}$ respectively (Kolmogorov-Smirnov statistics).

We tuned the hyperparameters to minimize nWass2. Overall, our methods reach higher fairness with lower loss in accuracy. The particularly low fairness of GF Lag on C&C is due to shrinking of the distributions. In Fig. 2 we show the histograms of the model outputs for two of the eight race groups for (left) Wass-2 Geo and (right) GF Lag.



(a) NLSY (b) C&C



(c) LSAC

Figure 3: Pareto frontiers for Wass-2 Geo, Wass-2 Penalty and MMD Penalty.

	PSCF – Adult		
	Err-0.5	nWass2	KS
Unconstrained	0.183	N/A	N/A
PSCC	0.195	0.977	0.968
PSCC $\mathcal{W}_{\mathcal{C}_\epsilon}$ Penalty	0.206	**0.620**	**0.462**

Table 2: Input transportation results.

To compare performance on fairness vs accuracy trade-off, in Fig. 3 we plot the Pareto frontiers for Wass-2 Geo, Wass-2 Penalty and MMD Penalty, using nWass2 and MSE. The geodesic trade-off curve is computed by setting $t = [0, 0.1, 0.2, \ldots, 1]$. For Wass-2 and MMD Penalty, each datapoint is computed by fixing the trade-off parameter $\eta = \log(\alpha/\beta)$ and averaging the nWass2 and MSE results over all other hyperparameters swept. Each datapoint thus reflects average performance at a trade-off point controlled by η. On all datasets, the Wass-2 Penalty Pareto frontier lies on or inside the MMD Penalty Pareto frontier, demonstrating that Wass-2 Penalty achieves a more optimal trade-off than MMD Penalty, and is close to the analytically-optimal geodesic trade-off curve. In addition, MMD Penalty produces a wider spread of MSE while achieving similar fairness levels, indicating higher sensitivity to hyperparameters. Similar conclusions hold for KS/MSE and Corr/MSE trade-offs.

PSCF through Input Transportation

We evaluated the input transportation approach for path-specific counterfactual fairness on the UCI Adult dataset.

We assumed the same CBN of Fig. 1, with A correspond-

ing to sex, C to the duple age and nationality, M to level of education, L to the triple capital gain, capital loss, and hours per week, and Y to income. As in Fig. 1, all causal paths from A to Y through M, *i.e.* $A \to M \to L \to Y$, $A \to M \to Y$, and the direct path $A \to Y$ were considered unfair. In Table 2, we show the results obtained with the unconstrained model, with a baseline path-specific correction approach that does not impose independence constraints (PSCC), and with our penalty method (PSCC $\mathcal{W}_{\mathcal{C}_\epsilon}$ Penalty). Err-0.5 indicates classification error with thresholding the model outputs at $\tau = 0.5$.

Conclusions

We have proposed an approach to fairness based on transporting distributions corresponding to different sensitive attributes to a common distribution. The use of optimal transport theory enabled us to devise theoretically sound methods that can achieve fairness through minimal changes to the unfair model. Our approach is widely applicable to all fairness criteria that can be framed as requiring statistical independence with respect to sensitive attributes.

Acknowledgements

The authors would like to thank Mark Rowland for useful feedback on the manuscript.

References

Altschuler, J.; Weed, J.; and Rigollet, P. 2017. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *NeurIPS*, 1964–1974.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *ICML*, 214–223.

Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2017. A convex framework for fair regression. In *Fairness, Accountability, and Transparency in Machine Learning*.

Black, E.; Yeom, S.; and Fredrikson, M. 2019. FlipTest: Fairness auditing via optimal transport. *CoRR* abs/1906.09218.

Chiappa, S., and Isaac, W. S. 2019. *A Causal Bayesian Networks Viewpoint on Fairness*, volume 547 of *IFIP AICT*. Springer Nature Switzerland. 3–20.

Chiappa, S. 2019. Path-specific counterfactual fairness. In *AAAI*, 7801–7808.

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

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *KDD*, 797–806.

Cotter, A.; Jiang, H.; Wang, S.; Narayan, T.; Gupta, M.; You, S.; and Sridharan, K. 2018. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *CoRR* abs/1809.04198.

Cuturi, M., and Doucet, A. 2014. Fast computation of Wasserstein barycenters. In *ICML*, 685–693.

Del Barrio, E.; Gamboa, F.; Gordaliza, P.; and Loubes, J.-M. 2019. Obtaining fairness using optimal transport theory. In *ICML*, 2357–2365.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Feldman, M.; Friedler, S.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *KDD*, 259–268.

Gajane, P., and Pechenizkiy, M. 2017. On formalizing fairness in prediction with machine learning. *CoRR* abs/1710.03184.

Goh, G.; Cotter, A.; Gupta, M.; and Friedlander, M. 2016. Satisfying real-world goals with dataset constraints. In *NeurIPS*, 2415–2423.

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13:723–773.

Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; and Chiappa, S. 2019. Wasserstein fair classification. In *UAI*.

Johndrow, J., and Lum, K. 2019. An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics* 13(1):189–220.

Kantorovich, L. 1942. On the transfer of masses (in Russian). *Doklady Akademii Nauk* 37(2):227–229.

Kusner, M.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *NeurIPS*, 4069–4079.

Lichman, M. 2013. UCI machine learning repository.

Lisini, S. 2007. Characterization of absolutely continuous curves in Wasserstein spaces. *Calculus of Variations and Partial Differential Equations* 28(1):85–120.

Mitchell, S.; Potash, E.; and Barocas, S. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *CoRR* abs/1811.07867:1–17.

Monge, G. 1781. Memoire sur la theorie des déblais et des remblais. *Histoire de l' Académie des Sciences de Paris*.

Narasimhan, H.; Cotter, A.; Gupta, M.; and Wang, S. 2020. Pairwise fairness for ranking and regression. In *AAAI*.

Peyré, G., and Cuturi, M. 2019. Computational optimal transport. *Foundations and Trends in Machine Learning* 11(5-6):355–607.

Risser, L.; Vincenot, Q.; Couellan, N.; and Loubes, J. 2019. Using Wasserstein-2 regularization to ensure fair decisions with neural-network classifiers. *CoRR* abs/1908.05783.

Simoiu, C.; Corbett-Davies, S.; and Goel, S. 2017. The problem of infra-marginality in outcome tests for discrimination. *Annals of Applied Statistics* 11(3):1193–1216.

Verma, S., and Rubin, J. 2018. Fairness definitions explained. In *Proceedings of the International Workshop of Software Fairness*, 1–7.

Villani, C. 2009. *Optimal Transport Old and New*. Springer.

Wang, H.; Ustun, B.; and Calmon, F. 2019. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *ICML*, 6618–6627.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *ICML*, 325–333.

Zhang, J., and Bareinboim, E. 2018. Fairness in decision-making – the causal explanation formula. In *AAAI*, 2037–2045.