Obtaining Fairness using Optimal Transport Theory

Eustasio del Barrio [*1] **Fabrice Gamboa** [*2] **Paula Gordaliza** [*2 1] **Jean-Michel Loubes** [*2]

Abstract

In the fair classification setup, we recast the links between fairness and predictability in terms of probability metrics. We analyze repair methods based on mapping conditional distributions to the Wasserstein barycenter. We propose a *Random Repair* which yields a tradeoff between minimal information loss and a certain amount of fairness.

1. Introduction

Along the last decade, machine learning methods have become more popular to build decision algorithms. Originally meant for Internet recommendation systems, they are now widely used in a large number of very sensitive areas such as medicine, human ressources with hiring policies, banking and insurance (lending), police and justice with criminal sentencing, see for instance (Berk et al., 2017), (Pedreschi et al., 2012) or (Friedler et al., 2018). The decisions made by what is now referred to as AI have a growing impact on human life. The whole machinery of these techniques relies on the fact that a decision rule can be learnt by looking at a subset of labeled examples, the learning sample, and then is applied to the whole population which is assumed to follow the same underlying distribution. So the decision is highly influenced by the choice of the learning set.

In some cases, this learning sample may present some bias or discrimination that could possibly be learnt by the algorithm and then propagated to the entire population through automatic decisions, providing a mathematical legitimacy for this unfair treatment. When giving algorithms the power to make automatic decisions, the danger may come that the reality may be shaped according to their prediction, thus reinforcing their beliefs in the model which is learnt. Hence, achieving fair treatment is one of the growing fields of interest in machine learning. For a recent survey on this topic we refer to (Zafar et al., 2017) or (Friedler et al., 2018).

Classification algorithms are one particular focus of fairness concerns since classifiers map individuals to outcomes. Some variables, such as sex, age or ethnic origin, are potentially sources of unfair treatment since they enable to create information that should not be processed out by the algorithm. Such variables are called in the literature protected variables. An algorithm is said to be fair with respect to these attributes when its outcome does not allow to make inference on the information they convey. Of course, the naive solution of ignoring these attributes when learning the classifier does not ensure this, since the protected variables may be closely correlated with other features enabling a classifier to reconstruct them.

Two solutions have been considered in the fair learning literature. The first one consists in changing the classifier in order to make it not correlated to the protected attribute. We refer for instance to (Zafar et al., 2017), (Bechavod & Ligett, 2017) or (Donini et al., 2018). Yet, explaining how the classifier is chosen may be seen too intrusive for many companies, or some of them may not even be able to change the way they build their models. Hence, a second solution consists in modifying the input data so that predictability of the protected attribute is impossible, whatever the classifier we train. The idea consists in blurring the value of the protected class trying to obtain a fair treatment. This point of view has been proposed in (Feldman et al., 2015), (Johndrow & Lum, 2017) and (Hacker & Wiedemann, 2017), for instance.

In this paper, we first provide in Section 2 a statistical analysis of the Disparate Impact definition and recast some of the ideas developed in (Feldman et al., 2015) to stress the links between fairness, predictability and the distance between the distributions of the variables given the protected attribute. Then, in Section 3 we provide first in 3.1 some theoretical justifications of the methodology proposed by previous authors (for one-dimensional data) to blur the data using the barycenter of the conditional distribution with respect to the Wasserstein distance. These methods are called either *total* or *partial repair*. Then in Section 3.2, we propose another methodology called *random repair* to transform the data in order to achieve a tradeoff between a minimal information loss of the classification task and still a certain level of fairness. We extend in Section 4 this procedure to the

[*]Equal contribution [1]IMUVA, Universidad de Valladolid, Valladolid, Spain [2]Institut de Mathématiques de Toulouse, Université Paul Sabatier, Toulouse, France. Correspondence to: Paula Gordaliza <paula.gordaliza@math.univ-toulouse.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

multidimensional case and provide a feasible algorithm to achieve the repair using the notion of Wasserstein barycenter. Finally application to simulated data in Section 5 enables to study the efficiency of the proposed procedures.

2. Framework for the fairness problem

Consider the probability space $(\Omega, \mathcal{B}, \mathbb{P})$, with $\mathcal{B}$ the Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geq 1$. In this paper, we tackle the problem of forecasting a binary variable $Y : \Omega \to \{0, 1\}$, using observed covariates $X : \Omega \to \mathbb{R}^d,\ d \geq 1$. We assume moreover that the population can be divided into two categories that represent a bias, modeled by a variable $S : \Omega \to \{0, 1\}$. This variable is called the protected attribute and takes the values $S = 0$ for the *minority* (assumed to be the unfavored class), and $S = 1$ for the *default* (and, usually, favored class). We also introduce also a notion of positive prediction: $Y = 1$ represents a *success* while $Y = 0$ is a *failure*. Hence, the classification problem aims at predicting a success from variables X, using a family $\mathcal{G}$ of binary classifiers $g : \mathbb{R}^d \to \{0, 1\}$. For every $g \in \mathcal{G}$, the outcome of the classification will be the prediction $\hat{Y} = g(X)$. We refer to (Bousquet et al., 2004) for a complete description of classification problems in statistical learning.

In this framework, discrimination or unfairness of the classification procedures, appears as soon as the prediction and the protected attribute are too closely related, in the sense that statistical inference on Y may lead to learn the distribution of the protected attribute S. This issue has received lots of attention in the last years and several ways to quantify this *discrimination bias* have been given. We refer for instance to (Lum & Johndrow, 2016), (Chouldechova, 2017) or (Bechavod & Ligett, 2017) for the analysis of fairness in machine learning. Here we focus on the definition given in (Feldman et al., 2015) or (Berk et al., 2017). A classifier $g : \mathbb{R}^d \to \{0, 1\}$ is said to achieve *statistical parity*, with respect to the joint distribution of (X, S), if

$$\mathbb{P}(g(X) = 1 \mid S = 0) = \mathbb{P}(g(X) = 1 \mid S = 1). \tag{1}$$

This means that the probability of a successful outcome is the same across the groups. Yet, the independence described in (1) is difficult to achieve and may not exist in real data. An index called *disparate impact* (DI) of the classifier g with respect to (X, S) has been introduced in (Feldman et al., 2015) as

$$DI(g, X, S) = \frac{\mathbb{P}(g(X) = 1 \mid S = 0)}{\mathbb{P}(g(X) = 1 \mid S = 1)}. \tag{2}$$

The ideal scenario where g achieves statistical parity is equivalent to $DI(g, X, S) = 1$. As we have metioned, statistical parity is often unrealistic and we can consider instead a certain level of fairness as in the following definition.

Definition 2.1. *The classifier g has disparate impact at level $\tau \in (0, 1]$, with respect to (X, S), if $DI(g, X, S) \leq \tau$.*

The disparate impact of a classifier measures its level of fairness: the smaller the value of τ, the less fair it is. In the following, we denote $a(g) := \mathbb{P}(g(X) = 1 \mid S = 0)$ and $b(g) := \mathbb{P}(g(X) = 1 \mid S = 1)$. In this paper, we will consider classifiers g such that $a(g) > 0$ and $b(g) > 0$ (the classifier is not totally unfair, in the sense that it does not predict the same outcome for a whole level of the protected attribute). Moreover, we assume $b(g) \geq a(g)$ (the default class $S = 1$ is more likely to have a successful outcome). Thus, in the definition above $0 < \tau \leq 1$. We point out that the value $\tau_0 = 0.8 = 4/5$, also known in the literature as the 80% *rule*, has been cited as a legal score to decide whether the discrimination of the algorithm is acceptable or not (see for instance (Feldman et al., 2015)). This rule ensures that "for every 5 individuals with successful outcome in the majority class, 4 in the minority class will have a successful outcome too". It will be useful in the sequel to use the definition in the reverse (positive) sense: a classifier does not have disparate impact at level τ, with respect to (X, S), if $DI(g, X, S) > \tau$.

Finally, another definition has been proposed in the statistical literature on fair learning. Given a classifier $g \in \mathcal{G}$, its *balanced error rate* (BER) with respect to the joint distribution of the random vector (X, S) is defined as the average class-conditional error

$$BER(g, X, S) = \frac{a(g) + 1 - b(g)}{2}. \tag{3}$$

Notice that $BER(g, X, S)$ is the misclassification error of $g \in \mathcal{G}$ for predicting S when the protected classes are equally likely ($\mathbb{P}(S = 0) = \mathbb{P}(S = 1) = 1/2$). This allows to define the notion of $\varepsilon-$predictability of the protected attribute. S is said to be $\varepsilon-$predictable from X if there exists a classifier $g \in \mathcal{G}$ such that $BER(g, X, S) \leq \varepsilon$. Equivalently, S is not $\varepsilon-$predictable from X if $BER(g, X, S) > \varepsilon$, for all classifiers g chosen in the class $\mathcal{G}$. Thus, if $\min_{g \in \mathcal{G}} BER(g, X, S) = \varepsilon^*$ then S is not $\varepsilon-$predictable from X for all $\varepsilon < \varepsilon^*$.

In the following, we recast previous notions of fairness and provide a probabilistic framework to highlight the relationships between the distribution of the observations and the fairness of the classification problem. We denote $\mu_s := \mathcal{L}(X|S = s), s = 0, 1$. The following theorem generalizes the result in (Feldman et al., 2015) showing the relationship between predictability, disparate impact and total variation distance.

Theorem 2.2. *Given r.v.'s $X \in \mathbb{R}^d,\ S \in \{0, 1\}$, the classifier g has disparate impact at level $\tau \in [0, 1]$, if and only if $BER(g, X, S) \leq \frac{1}{2} - \frac{a(g)}{2}(\frac{1}{\tau} - 1)$. Moreover*

$$\min_{g \in \mathcal{G}} BER(g, X, S) = \frac{1}{2}\left(1 - d_{TV}\left(\mu_0, \mu_1\right)\right).$$

As noted in the Introduction, to get rid of the possible dis-

crimination associated to a classifier we could, in principle, either modify the classifier or the input data. If action on the algorithm is not possible (for instance, if we have no access to the values Y of the learning sample) we have to focus on the second option and change the data X to ensure that every classifier trained from the modified data would be fair with respect to S. This transformation aimed at breaking the dependence on the protected attribute, is called *repairing the data*. For this, (Feldman et al., 2015), (Johndrow & Lum, 2017) or (Hacker & Wiedemann, 2017) propose to map the conditional distributions to a common distribution in order to achieve statistical parity. This *total repair* of the data amounts to modifying the input variables X building a repaired version, $\tilde{X}$, such that any classifier g trained from $\tilde{X}$ will have disparate impact $\tau = 1$, with respect to $(\tilde{X}, S)$ (equivalently, every classifier g that predicts Y from the new variable $\tilde{X}$ will achieve statistical parity). As a counterpart, it is clear that the choice of the target distribution should convey as much information as possible on the original variables, otherwise it would hamper the accuracy of the new classification.

In more detail, *total repair* amounts to mapping the original variable X into a new variable $\tilde{X} = T_S(X)$ such that conditional distributions with respect to S are the same, namely,

$$\mathcal{L}\left(\tilde{X} \mid S = 0\right) = \mathcal{L}\left(\tilde{X} \mid S = 1\right). \tag{4}$$

In this case, any classifier g built with such information will be such that $\mathcal{L}\left(g(\tilde{X}) \mid S = 0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S = 1\right)$, guaranteeing full fairness of the classification rule. To accomplish this transformation, the solution detailed in many papers is to map both conditional distributions μ_0 and μ_1 onto a common distribution ν. Actually, the distribution of X is modified using a random map $T_S : \mathbb{R}^d \to \mathbb{R}^d$ that depends on the value of the protected variable S and such that $\mathcal{L}\left(T_0(X) \mid S = 0\right) = \mathcal{L}\left(T_1(X) \mid S = 1\right)$. Consequently, two different problems arise.

- First of all, the choice of the distribution ν should be as similar as possible to both distributions μ_0 and μ_1 at the same time, in order to reduce the amount of information lost with this transformation, and thus still enabling the prediction task using the modified variable $\tilde{X} \sim \nu$ instead of the original X.
- Moreover, once the target ν is selected, we have to find the optimal way of transporting μ_0 and μ_1 into it.

First, from Theorem 2.2, the total variation distance is the natural choice to measure the distances between the conditional distributions in the fairness problem. However, this distance is computationally difficult to handle. Hence, previous works suggest the use of the Wasserstein metric, W_2, which appears as an appropriate tool for comparing probability distributions and arises naturally in optimal transport theory. We refer to (Villani, 2009) for general background on the topic. In this framework, T_S will be a random transport map between the distributions $\mathcal{L}(X \mid S)$ and $\mathcal{L}(\tilde{X})$. Then, when considering an optimal choice for the target distribution for $\mathcal{L}(\tilde{X})$, some authors (see (Feldman et al., 2015)) propose, in the one-dimensional case, to choose the distribution whose quantile is the mean of the quantile functions. In general this corresponds actually to the so-called Wasserstein barycenter of the laws $\mathcal{L}(X \mid S = s)$, as we describe next.

Given probability measures $(\mu_j)_{1 \le j \le J}$ with finite second moment and weights $(\omega_j)_{1 \le j \le J}$, the Wasserstein barycenter is a minimizer of

$$\nu \mapsto \sum_{j=1}^{J} \omega_j W_2^2(\nu, \mu_j), \tag{5}$$

see (Agueh & Carlier, 2011). Empirical versions of the barycenter and their properties are analyzed in (Boissard et al., 2015) or (Le Gouic & Loubes, 2017). Similar ideas have also been developed in (Cuturi & Doucet, 2014) or (Del Barrio & Loubes, 2017). In general, the Wasserstein barycenter appears to be a meaningful feature to represent the mean prototype of a set of distributions. Note that in the one dimensional case, the mean of the quantile functions corresponds actually to the minimizer of (5).

In the following section, we present some statistical justifications for this choice. Computation of Wasserstein barycenters may be a difficult issue in the general case. Yet, in this work we only consider the barycenter between two probabilities μ_0, μ_1 on $\mathbb{R}^d$, so we provide some details on how to compute this barycenter in general dimension.

3. Repair with Wasserstein Barycenter

3.1. Learning with Wasserstein Barycenter distribution

In our particular problem, where $J = 2$ in (5), the conditional distributions μ_0 and μ_1 are going to be transformed into the distribution of the Wasserstein barycenter μ_B between them, with weights π_0 and π_1, defined as

$$\mu_B \in \textit{argmin}_{\nu \in \mathcal{P}_2} \left\{\pi_0 W_2^2(\mu_0, \nu) + \pi_1 W_2^2(\mu_1, \nu)\right\}.$$

Let $\tilde{X}$ be the transformed variable with distribution μ_B. For each $s \in \{0, 1\}$, the deformation will be performed through the optimal transport map (o.t.m.) $T_s : \mathbb{R}^d \to \mathbb{R}^d$ pushing each μ_s towards the weighted barycenter μ_B. The existence of μ_B is guaranteed (see Theorem 2.12 in (Villani, 2003)) as soon as μ_s are absolutely continuous (a.c.) with respect to Lebesgue measure. In that case,

$$\mathbb{E}\left(\|X - T_s(X)\|^2 \mid S = s\right) = W_2^2(\mu_s, \mu_B). \tag{6}$$

Remark 3.1. *Note that computing the barycenter of two measures is equivalent to the computation of the o.t.m. be-*

tween them. If μ_0 is a.c. on $\mathbb{R}^d$ and $T : \mathbb{R}^d \to \mathbb{R}^d$ denotes the o.t.m. between μ_0 and μ_1, that is $\mu_1 = \mu_{0\sharp}T$, then $\mu_\lambda = \mu_{0\sharp}((1-\lambda)Id + \lambda T)$ is the weighted barycenter between μ_0 and μ_1, with weights $1-\lambda$ and λ, respectively. The map $(1-\lambda)Id + \lambda T$ is an optimal transport plan for all $\lambda \in [0,1]$. So, the complexity of computing $\mu_B = \mu_{0\sharp}(\pi_0 Id + \pi_1 T)$ is the same as computing T.

Remark 3.2. *Note also that for distributions on the real line, we can write the explicit expression of the barycenter μ_B based on the exact solution to the optimization problem (6). Given $s \in \{0,1\}$ and $X \in \mathbb{R}$, let $F_s : \mathbb{R} \to [0,1]$ denote the cumulative distribution function of X, given $S = s$, and $F_s^{-1} : [0,1] \to \mathbb{R}$ its quantile associated function. The weighted Wasserstein barycenter μ_B of μ_0 and μ_1 is the unique minimizer of the functional (5) and its quantile function can be computed as*

$$F_B^{-1}(t) = \left(\lambda F_0^{-1}(t) + (1-\lambda)F_1^{-1}(t)\right),\ t \in [0,1].$$

Moreover, note that $F_s(X \mid S = s) \sim \mathcal{U}(0,1), s = 0,1$, and the o.t.m. solution to (6) is $T_s = F_B^{-1} \circ F_s$.

To understand the use of the Wasserstein barycenter as the target distribution for μ_0 and μ_1, we will quantify the amount of information lost when replacing the distribution of X by a new and, for the moment, unknown distribution of $\tilde{X}$ obtained by transporting μ_0 and μ_1. Set the random transport plan $T_S : \mathbb{R}^d \to \mathbb{R}^d$, and the modified variable $\tilde{X} = T_S(X)$. We point out that choosing the distribution of $\tilde{X}$ amounts to choosing the transport plans T_0 and T_1. We are facing learning problems in two different settings.

- On the one hand, the full information available are the input variables X and the protected variable S, which play an important role in the classification, since the classifier has a different behavior according to the different classes $S = 0$ and $S = 1$. Hence, we let S play a role in the decision process since it is associated to Y, and possibly giving rise to a different treatment for the two different groups. In this case, the classification risk when the full data (X, S) is available can be computed as $R(g, X, S)$, the risk in the prediction of a classification rule g that depends on both variables X and S, namely $R(g, X, S) := \mathbb{P}(g(X, S) \neq Y)$
- On the other hand, in the repair data only the modified version $\tilde{X}$ of the input is at hand. Hence, the risk when learning a classifier is $R(h, \tilde{X}) := \mathbb{P}(h(\tilde{X}) \neq Y)$.

Studying the efficiency of the method requires providing a bound for the difference between the minimal risks obtained for the best classifier with input data $\tilde{X} = T_S(X)$, and for the best classifier with input data (X, S), called g_B. These risks are respectively denoted $R_B(\tilde{X})$ and $R_B(X, S) = \inf_g R(g, X, S) = R(g_B, X, S)$, and then its difference is

$$\mathcal{E}(\tilde{X}) := R_B(\tilde{X}) - R_B(X, S).$$

Note first that, given $X = x$ and $S = s$, $\inf_g R(g, X, S)$ can be computed by mimicking the usual expression of the 2-class classification error as in (Bousquet et al., 2004), for instance. Denoting by $\eta_s(x) := \mathbb{P}(Y = 1 \mid X = x, S = s)$,

$$\begin{aligned}&\mathbb{P}(g(X,S) \neq Y \mid X = x, S = s)\\ &= \mathbb{1}_{g(x,s)\neq 0}(1 - \eta_s(x)) + \mathbb{1}_{g(x,s)\neq 1}\eta_s(x).\end{aligned}$$

So we deduce that $R(g, X, S) = \mathbb{E}\left[\mathbb{1}_{g(X,S)=0}(2\eta_S(X) - 1)\right] + \mathbb{E}\left[1 - \eta_S(X)\right]$. The minimum risk is thus obtained using the Bayes' rule $g_B(x, s) = \mathbb{1}_{\eta_s(x)>1/2}$, showing that

$$\begin{aligned}&R_B(X, S) := min_g R(g, X, S)\\ &= \mathbb{E}\left[\mathbb{1}_{\{2\eta_S(X)-1<0\}}(2\eta_S(X) - 1)\right] + \mathbb{E}\left[1 - \eta_S(X)\right].\end{aligned}$$

Similarly, the risk related to a classifier $h(\tilde{X})$ is given by

$$\begin{aligned}&R(h, \tilde{X}) = R(h, T_S(X))\\ &= \mathbb{E}\left[\mathbb{1}_{h\circ T_S(X)=0}(2\eta_S(X) - 1)\right] + \mathbb{E}\left[1 - \eta_S(X)\right]. \quad (7)\end{aligned}$$

Hence, the amount of information lost due to modifying the data is controlled by the following theorem.

Theorem 3.3. *Consider $X \in \mathbb{R}^d$ and $S \in \{0,1\}$. Let $T_S : \mathbb{R}^d \to \mathbb{R}^d,\ d \geq 1$ be a random transformation such that $\mathcal{L}(T_0(X) \mid S = 0) = \mathcal{L}(T_1(X) \mid S = 1)$, and consider $\tilde{X} = T_S(X)$. Assume that $\eta_s(X)$ is Lipschitz with constant $K_s > 0$, $s = 0, 1$. Then, if $K = \max\{K_0, K_1\}$,*

$$\mathcal{E}(\tilde{X}) \leq 2\sqrt{2}K\left(\sum_{s=0,1}\pi_s W_2^2(\mu_s, \mu_{s\sharp}T_s)\right)^{\frac{1}{2}}. \quad (8)$$

Theorem 3.3 provides some justification to the use of the Wasserstein barycenter as the distribution of the modified variable. Similar inequalities in the framework of domain adaptation are given in (Redko et al., 2017). In fact, minimizing the upper bound in (8) with respect to the function $T_S : \mathbb{R}^d \to \mathbb{R}^d,\ d \geq 1$, leads to consider the transport plan carrying the conditional distributions μ_0 and μ_1 towards their Wasserstein barycenter μ_B with weights π_0, π_1, that is, $\mu_{S\sharp}T_S = \mu_B$. Hence, this provides some understanding on the choice of the Wasserstein barycenter advocated in the work (Feldman et al., 2015) and leads to the following bound

$$\begin{aligned}&\inf_{T_S}\{R(g_B \circ T_S, X) - R(g_B, X, S)\}\\ &\leq 2\sqrt{2}K\left(\sum_{s=0,1}\pi_s W_2^2(\mu_s, \mu_B)\right)^{\frac{1}{2}} \leq \frac{K}{\sqrt{2}}W_2^2(\mu_0, \mu_1).\end{aligned}$$

This upper bound only provides some guidelines on the choice of the target distribution. Nevertheless, choosing the Wasserstein barycenter provides a reasonable and, more important, feasible solution to achieve fairness. Recently in

(del Barrio et al., 2018) a CLT for L_p transportation cost in $\mathbb{R}$ is provided, which enables to build two sample tests and confidence intervals to certify the similarity between two distributions. We also point out that we only deal with the case of 2 classes for S, a majority and a minority, which is one of the main concerns in fair learning. Yet, the result could be generalized to multiclass where $S \in \mathcal{S}$ with several labels since it only relies on the defintion of the Wasserstein barycenter. In this case, computing the barycenter becomes a harder issue.

As pointed out previously, the *total repair* process ensures full fairness but at the expense of the accuracy of the classification. A solution for this could be found in (Feldman et al., 2015), called *geometric repair*. The authors propose not to move the conditional distributions to the barycenter but only partly towards it along the Wasserstein's geodesic path between μ_0 and μ_1. We analyze next this procedure and propose an alternative method for the partial repair.

3.2. A new algorithm for partial repair

Let $\lambda \in [0,1]$ be the parameter representing the amount of repair desired for X. Let Z be a target variable with distribution μ. Set $R_s = T_s^{-1},\ s = 0,1$, where T_s is the o.t.m. pushing each μ_s towards the target μ. Note that $R_s(Z)$ follows the original conditional distribution μ_s.

Definition 3.4 (Random repair)**.** *Let B be a Bernoulli variable with parameter λ. With the above notation, we define for $s \in \{0,1\}$, and $\lambda \in (0,1)$ the repaired distributions*

$$\begin{aligned} \tilde{\mu}_{s,\lambda} &= \mathcal{L}(BZ + (1-B)R_s(Z)) \\ &= \mathcal{L}(BT_s(X) + (1-B)X \mid S = s). \end{aligned} \quad (9)$$

This repair procedure consists in randomly changing the distribution of the original X by either selecting the target μ or the original conditional distributions. The degree of repair is governed by the Bernoulli parameter λ: note that for $\lambda = 0$ $\tilde{\mu}_{s,0} = \mathcal{L}(X \mid S = s)$ and for $\lambda = 1$ $\tilde{\mu}_{s,1} = \mathcal{L}(Z) = \mu$. The value of λ should come from a trade-off between *(i)* the accuracy of the new classification result, that leads to little changes in the initial distributions; and *(ii)* the non-predictability of the protected variable, which implies that the two conditional distribution should stay close with respect to the total variation distance. In fact, (see e.g. (Massart, 2007)), the distance in total variation between two probabilities P and Q can be computed as

$$d_{TV}(P,Q) = \min_{\pi \in \Pi(P,Q)} \pi(x \neq y). \quad (10)$$

This leads to

$$\begin{aligned} &d_{TV}(\tilde{\mu}_{0,\lambda}, \tilde{\mu}_{1,\lambda}) \leq \mathbb{P}(BZ + (1-B)R_0(Z) \\ &\neq BZ + (1-B)R_1(Z)) = 1 - \mathbb{P}(BZ + (1-B)R_0(Z) \\ &= BZ + (1-B)R_1(Z)) \leq 1 - \mathbb{P}(B = 1) = 1 - \lambda. \end{aligned}$$

This bound suggests that λ should be close to 1 to ensure non-predictability of S. Finally, observe that the misclassification error using the randomly repaired data is a mixture of the two errors with the totally repaired variable $T_S(X)$ and the original X since $R(g, \tilde{X}_\lambda) = (1-\lambda)\mathbb{P}(g(X) \neq Y) + \lambda \mathbb{P}(g(T_S(X)) \neq Y)$. Hence, from Theorem 3.3 the use of the Wasserstein barycenter $Z \sim \mu_B$ is justified.

In the literature (for instance (Zafar et al., 2017)), another partial repair procedure is used, called *geometric repair*. As before, μ is chosen as the barycenter μ_B and the partially repaired conditional distributions are defined as

$$\begin{aligned} \mu_{s,\lambda} &= \mathcal{L}(\lambda Z + (1-\lambda)R_s(Z)) \\ &= \mathcal{L}(\lambda T_s(X) + (1-\lambda)X \mid S = s),\ s \in \{0,1\}. \end{aligned}$$

Observe that $\lambda = 1$ yields the fully repaired variable, and $\lambda = 0$ leaves the conditional distributions unchanged. So the parameter λ governs how close the distributions are to the barycenter. Such procedure sounds appealing since the conditional distributions are moved on the geodesic path between the original distributions which warrants an optimal prediction and the barycenter which warrants fairness. Controling this distance is the key of the *geometric repair*. Yet, reasoning among the lines of previous argument to obtain an upper bound for the classification risk using the partially repaired distributions $\mu_{0,\lambda}$ and $\mu_{1,\lambda}$ does not lead to a satisfying result. This comes from the fact that the *geometric repair* moves the original distributions according to the Wasserstein distance, while fairness is measured through the total variation distance, and they are of different nature. So if $\lambda \in (0,1)$, using (10) implies that

$$\begin{aligned} &d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) \leq \mathbb{P}(\lambda Z + (1-\lambda)R_0(Z) \\ &\neq \lambda Z + (1-\lambda)R_1(Z)) = \mathbb{P}(R_0(Z) \neq R_1(Z)). \end{aligned} \quad (11)$$

The previous bound means that the amount of repair quantified by λ does not affect the TV distance between the modified conditional distributions. Moreover, in some situations, (11) turns out to be an equality. Consider, for instance,

$$\mu_{0,0} = U(K, K+1) \qquad \mu_{1,0} = U(-K-1, -K) \quad (12)$$

as the distributions of X in each class. Then, the barycenter is $\mu_{0,1} = \mu_{1,1} = U(-1/2, 1/2)$ and $\mu_{0,\lambda} = U\left(-\frac{\lambda}{2} + (1-\lambda)K, -\frac{\lambda}{2} + (1-\lambda)K + 1\right)$, $\mu_{1,\lambda} = U\left(-\frac{\lambda}{2} - (1-\lambda)(K+1), -\frac{\lambda}{2} - (1-\lambda)(K+1) + 1\right)$. In this case, the TV distance can be easily computed as

$$d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) = \min(1, (1-\lambda)(2K+1)). \quad (13)$$

We see from equation (13) that $d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) = 1$, if $\lambda \leq 2K/(2K+1)$, which means that the protected attribute could be perfectly predicted from the partially repaired data set for values of λ arbitrarily close to 1. Thus, the bound (11) provides some argument against the *geometric* method

since the repair should favour small distances between the original distributions to ensure a certain desired level of fairness. Hence, rather than using a displacement along the Wasserstein geodesic between the distributions, we propose the *random repair* which enables a better control of their total variation distance, enhancing the disparate impact while not hampering too much the efficiency of the classification.

In the next section, we propose a new algorithm for the *total repair* which in practice attains full fairness in contrast with the existing in the literature. Based on it, we design a scheme to perform the *random repair*.

4. Computational aspects for Repairing Datasets in General Dimension

Let $\{(X_i, S_i, Y_i), i = 1, \dots, N\}$ be an observed sample of (X, S, Y), and denote by n_0 and n_1 the number of instances in each protected class. Without loss of generality, we assume that the observations are ordered by the value of S,

$$x_{0,i} := X_i, \text{ if } s_i = 0,\ i = 1, \dots, n_0,$$
$$x_{1,j-n_0} := X_j, \text{ if } s_j = 1,\ j = n_0 + 1, \dots, N = n_0 + n_1.$$

Generally, the sizes of the samples $\mathcal{X}_0 = \{x_{0,1}, \dots, x_{0,n_0}\}$ and $\mathcal{X}_1 = \{x_{1,1}, \dots, x_{1,n_1}\}$ are different and Monge maps may not even exist between an empirical measure to another. This happens when their weight vectors are not compatible, which is always the case when the target measure has more points than the source measure. Hence, the solution to the optimal transport problem does not correspond to finding an optimal transport map, but an optimal transport distribution. The cuadratic cost function becomes discrete as it can be written as a matrix $C = (c_{ij})$, with $c_{ij} = \|x_{0,i} - x_{1,j}\|^2, 1 \le i \le n_0,\ 1 \le j \le n_1$. When $\mu_{0,n} = \sum_{i=1}^{n_0} \frac{1}{n_0} \delta_{x_{0,i}}$ and $\mu_{1,n} = \sum_{j=1}^{n_1} \frac{1}{n_1} \delta_{x_{1,j}}$, the Wasserstein distance $W_2(\mu_{0,n}, \mu_{1,n})$ between them is the squared root of the optimum of a net-work flow problem known as the *transportation problem*. It consists in finding a matrix $\gamma \in \mathcal{M}_{n_0 \times n_1}(\mathbb{R})$ which minimizes the transportation cost between the two distributions as follows

$$\begin{cases} \min_\gamma \sum_{\substack{1 \le i \le n_0 \\ 1 \le j \le n_1}} c_{ij}\gamma_{ij}, & \text{subject to:} \\ \gamma_{ij} \ge 0, & \\ \sum_{i=1}^{n_0} \gamma_{ij} = \frac{1}{n_1}, & \text{for all } j, \\ \sum_{j=1}^{n_1} \gamma_{ij} = \frac{1}{n_0}, & \text{for all } i. \end{cases} \tag{14}$$

If $\hat{\gamma}$ is a solution to the linear program (14) then, the measure $\mu_{B,n} = \sum_{\substack{1 \le i \le n_0 \\ 1 \le j \le n_1}} \hat{\gamma}_{ij} \delta_{\{\pi_0 x_{0,i} + \pi_1 x_{1,j}\}}$ is a barycenter of $\mu_{0,n}$ and $\mu_{1,n}$, with weights π_0 and π_1, according to Remark 3.1. See (Cuturi & Doucet, 2014) for details on the discrete Wasserstein and Optimal Transport computation.

4.1. Total repair

In practice, the implementation of the repair scheme in Section 3 is based on the transport matrix $\hat{\gamma}$ from $\mathcal{X}_0$ to $\mathcal{X}_1$. As we have pointed out, in this transport scheme the major difficulty comes from the fact that the sizes of these sets are different and the transport is not a one-by-one mapping. Each point in the source set could be transported (with weights) into several points of the target, or various points in the source could be moved into the same point of the target. As a consequence, we must adapt the algorithm that produces the repaired data set, denoted by $\tilde{\mathcal{X}}$.

We detail next two different methods. The first one is similar to some existing in the literature and does not achieve total fairness in practice, while the second one is a novelty and does guarantee this property for the new data $\tilde{\mathcal{X}}$.

(A) As depicted in Figure 1(A), each original point in $\mathcal{X}_0, \mathcal{X}_1$ is changed by a unique point given by

$$\tilde{x}_{0,i} = \pi_0 x_{0,i} + n_0 \pi_1 \sum_{j=1}^{n_1} \gamma_{ij} x_{1,j},\ 1 \le i \le n_0,$$
$$\tilde{x}_{1,j} = n_1 \pi_0 \sum_{i=1}^{n_0} \gamma_{ij} x_{0,i} + \pi_1 x_{1,j},\ 1 \le j \le n_1.$$

The set $\tilde{\mathcal{X}}$ will be a collection of exactly $n_0 + n_1$ points. This approach generalizes to higher dimensions the idea in (Feldman et al., 2015) and (Johndrow & Lum, 2017), which only considered the unidimensional case, where the transport is written in terms of the distribution funtions. Yet, in practice it builds two different sets $\tilde{\mathcal{X}}_0 = \{x_{0,i},\ 1 \le i \le n_0\}$ and $\tilde{\mathcal{X}}_1 = \{x_{1,j},\ 1 \le j \le n_1\}$ that do not ensure (4).

(B) To ensure total fairness, each point will split its mass to be transported into several modified versions. This generates an extended set $\tilde{\mathcal{X}} = \tilde{\mathcal{X}}_0 \cup \tilde{\mathcal{X}}_1$, which is formed by the complete distribution $\mu_{B,n}$. As shown in Figure 1(B), if $\hat{\gamma}_{ij} > 0, 1 \le i \le n_0,\ 1 \le j \le n_1$, we define two points

$$\tilde{x}_{0,i,j} := \tilde{x}_{1,j,i} = \pi_0 x_{0,i} + \pi_1 x_{1,j}, \tag{15}$$

and sets $\tilde{\mathcal{X}}_0 := \bigcup_{i=1}^{n_0} \{\tilde{x}_{0,i,j} / \ \hat{\gamma}_{ij} > 0, 1 \le j \le n_1\}$, and $\tilde{\mathcal{X}}_1 := \bigcup_{j=1}^{n_1} \{\tilde{x}_{1,j,i} \ / \ \hat{\gamma}_{ij} > 0, 1 \le i \le n_0\}$. The rebuilt distributions have sizes equal to the number of non zero elements in $\hat{\gamma}$, and each point has weight $\hat{\gamma}_{ij}$. Unlike the previous, this approach does achieve total impredictability, as it manages to produce repaired conditional distributions equally distributed.

Example 4.1. *We have simulated two samples $\mathcal{X}_0$ and $\mathcal{X}_1$ of points in $\mathbb{R}$ of sizes $n_0 = 4$ and $n_1 = 7$. The optimal*

matrix solution to the problem (14) is

$$\hat{\gamma} = \begin{bmatrix} \frac{1}{7} & \frac{1}{4} - \frac{1}{7} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2}{7} - \frac{1}{4} & \frac{1}{7} & \frac{1}{14} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{14} & \frac{1}{7} & \frac{2}{7} - \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} - \frac{1}{7} & \frac{1}{7} \end{bmatrix}$$

*If $\mathcal{X}_0$ and $\mathcal{X}_1$ are realizations of μ_0 and μ_1, respectively, then the left part of Figure 1 represents procedure **(A)** that produces the repaired sets $\tilde{\mathcal{X}}_0 = \{\tilde{x}_{0,1}, \ldots, \tilde{x}_{0,4}\}$ (rounded green points) and $\tilde{\mathcal{X}}_1 = \{\tilde{x}_{1,1}, \ldots, \tilde{x}_{1,7}\}$ (squared green points). As we can observe, the two sets are clearly different and the statistical parity can not be reached. Otherwise, procedure **(B)** on the right yields to $\tilde{\mathcal{X}}_0 = \tilde{\mathcal{X}}_1$.*

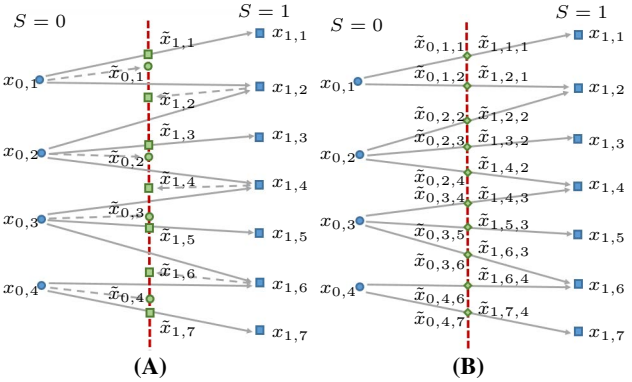

Figure 1. Example of the performance of procedures (A) and (B)

Remark 4.2. *When the two samples $\mathcal{X}_0$ and $\mathcal{X}_1$ have equal size n and the weights $\gamma_{ij} = \frac{1}{n}$, $1 \leq i, j \leq n$, are uniform, the mass conservation constraint implies that γ is a bijection and the Monge problem is equivalent to the optimal matching problem $\min_{\sigma \in Perm(n)} \frac{1}{n} \sum_{i=1}^{n} c_{i,\sigma(i)}$. Both repairing procedures **(A)** and **(B)** perfom the same generating $\tilde{x}_{0,i} = \tilde{x}_{1,i} = \frac{1}{2}(x_{0,i} + x_{1,i})$, $1 \leq i \leq n$, as depicted in Figure 2. Then, total fairness is always achieved.*

4.2. Random repair

As previously noted, trying to build the set $\tilde{\mathcal{X}}$ satisfying the goal (4) may compromise too much the accuracy of the classification with these new data. In this sense, the *random repair* procedure proposed in this paper aims at setting a tradeoff between fairness and accuracy through the parameter λ, that models the amount of repair desired. We detail next how to compute the randomly repaired set denoted by $\tilde{\mathcal{X}}_\lambda$, with $\lambda \in [0, 1]$. According to (9), we will randomly select either the points in the original sets $\mathcal{X}_0$ and $\mathcal{X}_1$ or their repaired sequels with procedure **(B)**. For this, consider a sample $b_1, \ldots, b_{n_0+n_1} \sim B(\lambda)$, and define

$$\tilde{\mathcal{X}}_{0,\lambda} := \bigcup_{i=1}^{n_0} R_{0,i,\lambda} \qquad \tilde{\mathcal{X}}_{1,\lambda} := \bigcup_{j=1}^{n_1} R_{1,j,\lambda}, \tag{16}$$

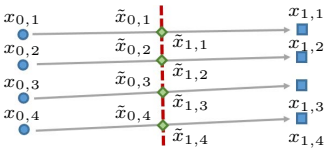

Figure 2. Repairing process when $n_0 = n_1$

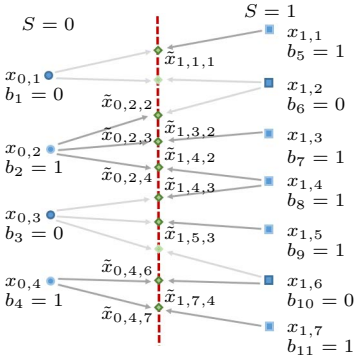

Figure 3. Example of the *random repair* with $\lambda = \frac{1}{2}$.

where $R_{0,i,\lambda}$ and $R_{1,j,\lambda}$ are the repaired sets of the points $x_{0,i}$ and $x_{1,j}$, respectively:

$$R_{0,i,\lambda} := \begin{cases} \{x_{0,i}\} & \text{if } b_i = 0 \\ \{\tilde{x}_{0,i,j} \,/\, \hat{\gamma}_{ij} > 0, 1 \leq j \leq n_1\} & \text{if } b_i = 1 \end{cases}$$

$$R_{1,j,\lambda} := \begin{cases} \{x_{1,j}\} & \text{if } b_{n_0+j} = 0 \\ \{\tilde{x}_{1,j,i} \,/\, \hat{\gamma}_{ij} > 0, 1 \leq i \leq n_0\} & \text{if } b_{n_0+j} = 1 \end{cases}$$

with $\tilde{x}_{0,i,j}$ and $\tilde{x}_{1,j,i}$ given in (15), with weights $\hat{\gamma}_{i,j}$.

Example 4.3. *Consider the situation in Example 4.1. Figure 3 represents the random repair procedure for $\lambda = \frac{1}{2}$. For $l = 1, \ldots, n_0 + n_1 = 11$, we have simulated values $b_l \sim \mathcal{B}(\frac{1}{2})$. From (16) we have the randomly repaired sets*

$$\tilde{\mathcal{X}}_{0,\lambda} = \{x_{0,1}, \tilde{x}_{0,2,2}, \tilde{x}_{0,2,3}, \tilde{x}_{0,2,4}, x_{0,3}, \tilde{x}_{0,4,6}, \tilde{x}_{0,4,7}\}$$
$$\tilde{\mathcal{X}}_{1,\lambda} = \{\tilde{x}_{1,1,1}, x_{1,2}, \tilde{x}_{1,3,2}, \tilde{x}_{1,4,2}, \tilde{x}_{1,4,3}, \tilde{x}_{1,5,3}, x_{1,6}, \tilde{x}_{1,7,4}\}.$$

5. Application with simulated data

In this section, we present an application of the repairing procedures in Section 3 to some simulated data to illustrate their performance. We also provide an example in which the *geometric repair* fails to remove the bias in the data.

To introduce some bias in the simulated dataset $\mathcal{X}$ we have taken $n_0 = 600$ and $n_1 = 400$ examples from two multivariate normal distributions on $\mathbb{R}^5$ with vector of means $\mu_0 = (3, 3, 2, 2.5, 3.5)$ and $\mu_1 = (4, 4, 3, 3.5, 4.5)$ and equal covariance matrices $\Sigma = diag(1, 1, 0.5, 0.5, 1)$. Then, in order to simulate the classification Y, we have chosen parameters $\beta_0 = (1, -1, -0.5, 1, -1, 1)$ and $\beta_1 = (1, -0.4, 1, -1, 1, -0.5)$ to build a logit model for each group with different probability of success for $s = 0, 1$, $\pi_s(x) = \frac{e^{X\beta_s}}{1+e^{X\beta_s}}$, higher for the class $S = 1$.

Then, a new logit classifier has been trained from this simulated data, splitting the set into the learning and the test sample using the ratio 300 / 700. In the first row of Table 1 we can see a summary of the performance of the logit with

the original data. We have estimated the disparate impact using its empirical counterpart and provided a confidence interval which was established in (Besse et al., 2018). Before the repair, we can say with a confidence of 95% that the logit rule has DI at level 0.53 with respect to S. Then, we have made the repair in $\mathbb{R}^5$ in the testing sample using the different procedures studied in this paper. We have used the previous logit model, which was trained from biased data, to classify such repaired observations. In the remaining rows of Table 1 a summary of the performance of the logit with the repaired data using procedures **(A)** and **(B)** is presented. We note that in the experiments with procedure **(A)** the estimated value for DI is not exactly 1, as we have already anticipated. On the other hand, procedure **(B)** manages to change the data to attain statistical parity. The error in the logit classification done with the repaired data sets is a bit higher for the second procedure.

Finally, we present some results of the performance of the *Geometric* and *random repair*. Figure 4 represents the evolution of the confidence interval for the disparate impact with the amount of repair $0 \leq \lambda \leq 1$. Figure 5 shows the evolution with λ of the error in the classification done from the modified data set. For the experiments concerning the *random repair* procedure, we have repeated it 100 times and then we have computed the mean of the simulations. Clearly, the reached level of DI of the logit rule is higher with the *random repair*. We note that the amount of repair necessary to achieve an estimated DI at level 0.8 for the logit rule is 0.475 with the *random repair*, which entails an error of 0.1537; and 0.7 with the *geometric repair*, which entails an error of 0.1371.

Table 1. Disparate impact of the logit with the original and the repaired datasets

Repair	Error	Difference	$\hat{DI}$	CI 95%
-	0.0943	-	0.5309	$(0.4230, 0.6389)$
(A)	0.1629	0.0686	0.9588	$(0.7641, 1.1535)$
(B)	0.1874	0.0931	1	$(0.8536, 1.1464)$

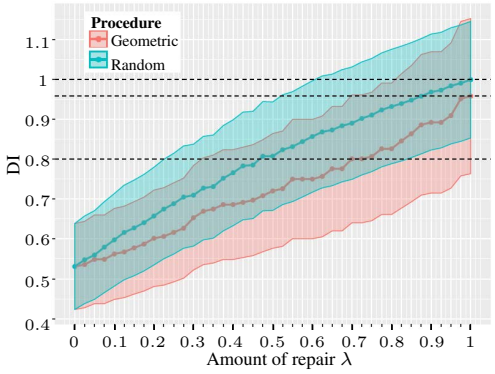

Figure 4. CI at level 95% for DI of the logit

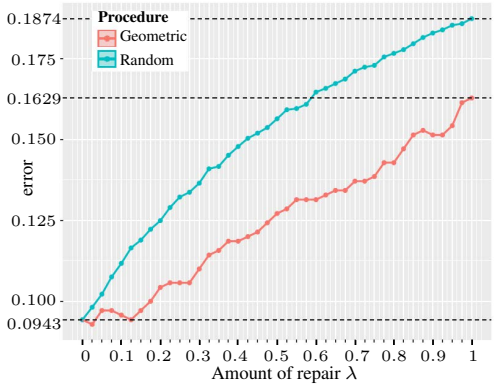

Figure 5. Error of the logit

In order to see the failure of the *geometric repair*, we have simulated $n_0 = n_1 = 500$ observations from uniform distributions as in (12) with $K = 10$. We have trained a random forest classifier with the same ratio $300/700$ for the learning and test sample. In Figure 6 we can see that the evolution of the disparate impact is controlled by the amount of repair only if we use the *random repair*. As pointed out from inequality (13), we observe that for values of $\lambda \leq \frac{20}{21} \approx 0.95$, the DI does not increase with λ for the partially modified distributions with the *geometric repair*. This means that for values of the degree of repair close to 1, this procedure does not manage to remove the bias in the data and consequently, it does not ensure the fairness of every classifier.

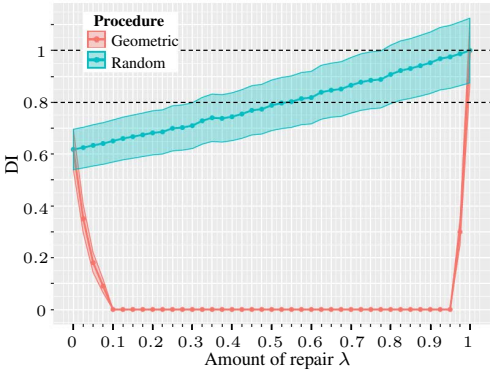

Figure 6. CI at level 95% for DI of the random forest classifier

6. Conclusions

We have provided a multidimensional expansion and a feasible algorithm to repair a learning sample and incorporate fairness to prevent unfair algorithms to be learnt. Moreover this way of correction can be improved using a random reparation as shown in the paper. Yet this way of reparation only deals with disparate impact assessment and other criterion such as conditional accuracy equality for instance will be further incorporated using the same ideas of Wasserstein barycenter of conditional distributions.

Acknowledgements

The authors thank ANITI program (Artificial Natural Intelligence Toulouse Institute). JM Loubes acknowledges the funding by DEEL-IRT.

References

Agueh, M. and Carlier, G. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43:904–924, 2011.

Bechavod, Y. and Ligett, K. Penalizing Unfairness in Binary Classification. *ArXiv e-prints*, June 2017.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

Besse, P., del Barrio, E., Gordaliza, P., and Loubes, J.-M. Confidence intervals for testing disparate impact in fair learning. *arXiv preprint arXiv:1807.06362*, 2018.

Boissard, E., Le Gouic, T., Loubes, J.-M., et al. Distributions template estimate with Wasserstein metrics. *Bernoulli*, 21:740–759, 2015.

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pp. 169–207. Springer, 2004.

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

Cuturi, M. and Doucet, A. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pp. 685–693, 2014.

Del Barrio, E. and Loubes, J.-M. Central limit theorem for empirical transportation cost in general dimension. *arXiv preprint arXiv:1705.01299*, 2017.

del Barrio, E., Gordaliza, P., and Loubes, J.-M. A central limit theorem for l_p transportation cost on the real line with application to fairness assessment in machine learning. *(in-press) Information and Inference*, 2018.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical Risk Minimization under Fairness Constraints. *ArXiv e-prints*, February 2018.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. A comparative study of fairness-enhancing interventions in machine learning. *ArXiv e-prints*, 2018.

Hacker, P. and Wiedemann, E. A continuous framework for fairness. *CoRR*, abs/1712.07924, 2017. URL http://arxiv.org/abs/1712.07924.

Johndrow, J. E. and Lum, K. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*, 2017.

Le Gouic, T. and Loubes, J.-M. Existence and consistency of Wasserstein barycenters. *Probab. Theory Rel.*, 168: 901–917, 2017.

Lum, K. and Johndrow, J. A statistical framework for fair predictive algorithms. *ArXiv e-prints*, October 2016.

Massart, P. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

Pedreschi, D., Ruggieri, S., and Turini, F. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 126–131. ACM, 2012.

Redko, I., Habrard, A., and Sebban, M. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 737–753. Springer, 2017.

Villani, C. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Soc., 2003. ISBN 9780821833124. URL https://books.google.es/books?id=GqRXYFxe0l0C.

Villani, C. *Optimal transport: old and new*. Springer Verlag, 2009.

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. International World Wide Web Conferences Steering Committee, 2017.