Obtaining fairness using optimal transport theory

Eustasio del Barrio, Fabrice Gamboa, Paula Gordaliza, Jean-Michel Loubes

Obtaining fairness using optimal transport theory

Eustasio del Barrio[a], Fabrice Gamboa[b], Paula Gordaliza[a,b], and Jean-Michel Loubes[b]
[a] *IMUVA, Universidad de Valladolid* and [b] *Institut de Mathématiques de Toulouse*

June 4, 2018

Abstract

Statistical algorithms are usually helping in making decisions in many aspects of our lives. But, how do we know if these algorithms are biased and commit unfair discrimination of a particular group of people, typically a minority? *Fairness* is generally studied in a probabilistic framework where it is assumed that there exists a protected variable, whose use as an input of the algorithm may imply discrimination. There are different definitions of Fairness in the literature. In this paper we focus on two of them which are called Disparate Impact (DI) and Balanced Error Rate (BER). Both are based on the outcome of the algorithm across the different groups determined by the protected variable. The relationship between these two notions is also studied. The goals of this paper are to detect when a binary classification rule lacks fairness and to try to fight against the potential discrimination attributable to it. This can be done by modifying either the classifiers or the data itself. Our work falls into the second category and modifies the input data using optimal transport theory.

Keywords : Fairness in Machine Learning, Optimal Transport, Wasserstein barycenter.

1 Introduction

Along the last decade, Machine Learning methods have become more popular to build decision algorithms. Originally meant for recommendation algorithms over the Internet, they are now widely used in a large number of very sensitive areas such as medicine, human ressources with hiring policies, banks and insurance (lending), police, and justice with criminal sentencing, see for instance in [BHJ+17] [PRT12] or [FSV+18] and references therein. The decisions made by what is now referred to as IA have a growing impact on human's life. The whole machinery of these technics relies on the fact that a decision rule can be learnt by looking at a set of labeled examples called the learning sample and then this decision will be applied for 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 by automatic decisions and, even worse, 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. Classification algorithms are one particular focus of fairness concerns since classifiers map individuals to outcomes.

Hence, achieving fair treatment is one of the growing fields of interest in Machine Learning. We refer for instance to [ZVGRG17] or [FSV$^+$18] for a recent survey on this topic. For this, several definitions of fairness have been considered. In this paper we focus on the notion of disparate impact for protected variables introduced in [FFM$^+$15]. Actually, 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 called 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 Machine 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 [ZVGRG17], [BL17] or [DOB$^+$18] and references therein. Yet changing the way a model is built or explaining how the classifier is chosen may be seen too intrusive for many companies or some may not be able to change the way they build the model. Hence a second solution consists in changing the input data so that predictability of the protected attribute is impossible. The data will be blurred in order to obtain a fair treatment of the protected class. This point of view has been proposed in [FFM$^+$15], [JL17] or [HW17] for instance.

In the following, we first provide a statistical analysis of the Disparate Impact definition and recast some of the ideas developed in [FFM$^+$15] to stress the links between fairness, predictability and the distance between the distributions of the variables given the protected attribute. Then we provide 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 full or partial repair. We extend this reparation procedure to the case of multidimensional data and provide a feasible algorithm to achieve this fairness reparation using the notion of Wasserstein barycenter. Finally, we propose another methodology called *Random Repair* to transform the data in order to achieve a tradeoff between a minimal loss of information with respect to the classification task and still a certain level of fairness for classification procedures that could be used with this transformed data. Applications to real data enable to study the efficiency of previous procedures.

The paper falls into the following parts. Section 2 presents the relationships between the notions of Disparate Impact, the predictability of a protected attribute and distance between the distributions conditionally to this attribute. Section 3 is devoted to a probabilistic framework to transform the data to obtain fair classifiers. The following section, Section 4, provides some insight to understand the use of the Wasserstein's barycenter and its limitation. Applications to a real data set are shown in Section 5, while the proofs are postponed to the Appendix.

2 Fairness using Disparate Impact assessment

Consider the probability space $(\Omega, \mathcal{B}, \mathbb{P})$, with $\mathcal{B}$ the Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geqslant 1$. In this paper, we tackle the problem of forecasting a binary variable $Y : \Omega \to \{0, 1\}$, using observed covariates $X : \Omega \to \mathbb{R}^d,\ d \geqslant 1$. 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, which takes the values $S = 0$ for the "minority"class and supposed to be the unfavored class; and $S = 1$ for the "default", and usually favored class. We also introduce also a notion of positive prediction in the sense that $Y = 1$ represents a success while $Y = 0$ is a failure.

Hence the classification problem aims at predicting a success from the variables X, using a family of binary classifiers $g \in \mathcal{G} : \mathbb{R}^d \to \{0,1\}$. For every $g \in \mathcal{G}$, the outcome of the classification will be the prediction $\hat{Y} = g(X)$. We refer for instance to [BBL04] 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 interest among the last years and several ways to quantify this *discrimination bias* have been given. We highlight two of them, whose interest depends on the particular problem. More precisely, we can deal with two situations, depending whether the true distribution of the label Y is available. If it is known, the definition introduced in [BHJ$^+$17], defines that a classifier $g : \mathbb{R}^d \to \{0,1\}$ achieves *Overall Accuracy Equality*, with respect to the joint distribution of (X, S, Y), if

$$\mathbb{P}(g(X) = Y \mid S = 0) = \mathbb{P}(g(X) = Y \mid S = 1). \tag{2.1}$$

This entails that the probability of a correct classification is the same across groups and, hence, the classification error is independent of the group. This idea can be also found in the [ZVGRG17] as the condition of g having *Disparate Mistreatment*, which happens when the probability of error is different for each group as in (2.1).

Nevertheless, in many problems, the true Y is not available (this data may be very sensitive and the owner of the data may not want to make it available), or the classification methodology can not be changed, so the study of fairness must be based on the outcome $\hat{Y}$. In this situation, following [FFM$^+$15] or [BHJ$^+$17], 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

$$\mathbb{P}(g(X) = 1 \mid S = 0) = \mathbb{P}(g(X) = 1 \mid S = 1). \tag{2.2}$$

This means that the probability of a successful outcome is the same across the groups. For instance, if we consider that the protected variable represents gender, the value $S = 0$ would be assigned to "female"and $S = 1$ to "male", we would say that the algorithm used by a company achieves *Statistical Parity* if a man and a woman have the same probability of success (for instance being hired or promoted).
We will use the following notations

$$a(g) := \mathbb{P}(g(X) = 1 \mid S = 0), \quad b(g) := \mathbb{P}(g(X) = 1 \mid S = 1).$$

In the rest of the paper, we consider classifiers g such that $a(g) > 0$ and $b(g) > 0$, which means that the classifier is not totally fair or unfair in the sense that the classifier does not predict the same outcome for a whole population according to the protected attribute.

The independence described in (2.2) is difficult to achieve and may not exist in real data. Therefore, to assess this kind of fairness, an index called *Disparate Impact of the classifier g with respect to* (X, S), has been introduced in [FFM$^+$15] as

$$DI(g, X, S) = \frac{\mathbb{P}(g(X) = 1 \mid S = 0)}{\mathbb{P}(g(X) = 1 \mid S = 1)}. \tag{2.3}$$

The ideal scenario where g achieves *Statistical Parity* is equivalent to $DI(g, X, S) = 1$. Statistical Parity is often unrealistic, so we will relax it into achieving a certain level of fairness as described in the following definition.

Definition 2.1. The classifier $g : \mathbb{R}^d \to \{0, 1\}$ has Disparate Impact at level $\tau \in (0, 1]$, with respect to (X, S), if $DI(g, X, S) \leq \tau$.

Note the Disparate Impact of a classifier measures its level of fairness: the smaller the value of τ, the less fair it is. The classification rules considered in this framework are such that $b(g) \geqslant a(g) > 0$, because we are assuming that the default class $S = 1$ is more likely to have a successful outcome. Thus, in the definition, the level of fairness τ takes values $0 < \tau \leq 1$. We point out that the value $\tau_0 = 0.8 = 4/5$, which is 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 [FFM$^+$15]). This rule can be explained as "for every 5 individuals with successful outcome in the majority class, 4 in the minority class will have a successful outcome too".

In what follows, to promote fairness, it will be useful to state the definition in the reverse 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

$$BER(g, X, S) = \frac{\mathbb{P}\left(g(X) = 0 \mid S = 1\right) + \mathbb{P}\left(g(X) = 1 \mid S = 0\right)}{2}. \tag{2.4}$$

Notice that $BER(g, X, S)$ is the general misclassification error of $g \in \mathcal{G}$ in the particular case when we have $\mathbb{P}(S = 0) = \mathbb{P}(S = 1) = 1/2$, which consists in the ideal situation when both protected classes have the same probability of occurence. This quantity enables 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 said not to be $\varepsilon-$predictable from X if $BER(g, X, S) > \varepsilon$, for all classifiers g chosen in the class $\mathcal{G}$. Thus, if the minimum of this quantity is achieved by a classifier g^*,

$$\min_{g \in \mathcal{G}} BER(g, X, S) = BER(g^*, X, S) = \varepsilon^*$$

then it is clear that S is not $\varepsilon-$predictable from X for all $\varepsilon \leq \varepsilon^*$.

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.

The following theorem generalizes the result in [FFM$^+$15], showing the relationship between predictability and Disparate impact.

Theorem 2.1. *Given random variables $X \in \mathbb{R}^d$, $S \in \{0, 1\}$, the classifier $g \in \mathcal{G}$ has Disparate Impact at level $\tau \in [0, 1]$, with respect to (X, S), if and only if $BER(g, X, S) \leq \frac{1}{2} - \frac{a(g)}{2}(\frac{1}{\tau} - 1)$.*

The following theorem establishes the relationship between ε^* the minimum Balance Error Rate and distance in Total Variation between the two conditional distributions $\mathcal{L}\left(X|S = 0\right)$ and $\mathcal{L}\left(X|S = 1\right)$.

Theorem 2.2. *Given the variables $X : \Omega \to \mathbb{R}^d$, $d \geqslant 1$, and $S : \Omega \to \{0, 1\}$,*

$$\min_{g \in \mathcal{G}} BER(g, X, S) = \frac{1}{2}\left(1 - d_{TV}\left(\mathcal{L}\left(X|S = 0\right), \mathcal{L}\left(X|S = 1\right)\right)\right).$$

This result shows that fairness expressed through the notion of Disparate Impact depends highly on the conditional distributions of the variables X conditionally to the protected attribute, $\mathcal{L}(X|S=0)$ and $\mathcal{L}(X|S=1)$.

Actually, Theorem 2.2 implies that S is not $\varepsilon-$predictable from X if, and only if,

$$d_{TV}\left(\mathcal{L}(X|S=0), \mathcal{L}(X|S=1)\right) < 1 - 2\varepsilon. \tag{2.5}$$

and, as a consequence of Theorem 2.1, for all $g \in \mathcal{G}$,

$$DI(g, X, S) > \frac{1}{1 + \frac{d_{TV}\left(\mathcal{L}(X|S=0), \mathcal{L}(X|S=1)\right)}{a(g)}}.$$

Hence, the smaller the Total Variation distance, the greater is the value ε that we could find satisfying Equation (2.5) and thus, the less predictable using the variables X will be S . The best case happens when $d_{TV}\left(\mathcal{L}(X|S=0), \mathcal{L}(X|S=1)\right) = 0$, which is equivalent to the equality of both conditional distributions $\mathcal{L}(X|S=0) = \mathcal{L}(X|S=1)$. In this situation, X and S are independent random variables, and we will have that S is not $\varepsilon-$predictable from X, for all $\varepsilon \leqslant \frac{1}{2}$, and $DI(g, X, S) = 1$. Note that clearly $\varepsilon = 1/2$ non predictability is the best that can be achieved.

3 Removing disparate impact using Optimal Transport

3.1 A probabilistic model for data repair

Some classification procedures exhibit a discrimination bias quantified through a potential Disparate Impact in the classification outcome $\hat{Y} = g(X)$, with respect to the joint distribution of (X, S). To get rid of the possible discrimination associated to a classifier g, two main strategies can be used, either modifying the classifiers or modifying the input data. In this work, we are facing the problem where we have no access to the values Y of the learning sample, hence we focus on the methodologies that intend to modify the data in order to achieve fairness.
The main idea is to change the data in order to break their relationship with the protected attribute. This transformation is called repairing the data. For this, [FFM$^+$15], [JL17] or [HW17] propose to map the conditional distributions to a common distribution in order to achieve statistical parity as described in (2.2). The choice of the common distribution in one dimension is described as the distribution obtained by taking the mean of the quantile functions. A total repair of the data amounts to modifying the input variables X building a repaired version, denoted by $\tilde{X}$, such that any classifier g will have Disparate Impact at level $\tau = 1$, with respect to $(\tilde{X}, S)$. This means that every classifier g used to predict the target class Y from the new variable $\tilde{X}$ will achieve *Statistical Parity* with respect to $(\tilde{X}, S)$. As a counterpart, it is clear that the choice of the distribution to whom the original variables are mapped should convey as much as information possible on the original variable, otherwise it would hamper the accuracy of the new classification. This constraint led some authors to recommend the use of the so-called Wasserstein barycenter.

We now present some statistical justifications for this choice and provide some comments on the way to repair the data to obtain fair enough classification rules without modifying too much the original data set.

Achieving Statistical Parity amounts to modifying the original data into a new random variable $\tilde{X}$ such that the conditional distribution with respect to the protected attribute S is

the same for all groups, namely

$$\mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right). \tag{3.1}$$

In this case, any classifier g built with such information will be such that

$$\mathcal{L}\left(g(\tilde{X}) \mid S=0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S=1\right),$$

which implies that $DI(g, \tilde{X}, S) = 1$ and so this transformation promotes full fairness of the classification rule.

To achieve this transformation, the solution detailed in many papers is to map both conditional distributions $\mu_0 := \mathcal{L}(X|S=0)$ and $\mu_1 := \mathcal{L}(X|S=1)$ onto a common distribution ν. Actually, the distribution of the original variables X is transformed using a map T_S which depends on the value of the protected attribute S

$$\begin{array}{rccc} T_S: & \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ & X & \longmapsto & \tilde{X} = T_S(X), \end{array}$$

and such that

$$\mathcal{L}\left(T_0(X) \mid S=0\right) = \mathcal{L}\left(T_1(X) \mid S=1\right). \tag{3.2}$$

Note that the function T_S is random because of its dependence on the binary random variable S.

In this framework, the problem of achieving Statistical Parity is the same as the problem of finding a (random) function T_S such that (3.2) holds. As it is represented in Figure 1, if we denote by $\mu_S \sim X \mid S$, our goal is to map these two distributions to a common law $\nu = \mu_S \circ T_S^{-1}$. Consequently, two different problems arise

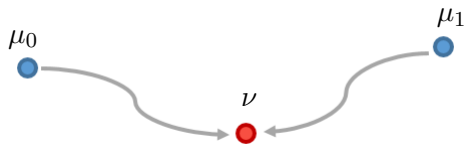

Figure 1: General repairing scheme

- 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.
- On the other hand, once we have selected the distribution ν, we have to find the optimal way of transporting μ_1 and μ_0 to this new distribution ν.

From Section 2, the natural distance related to fairness between the two conditional distributions is the total variation distance and that should be used. However, this distance is computationally difficult to handle, hence previous works promote the use of Wasserstein distance which appears as a natural distance to move distributions.

For this, we recall some results on optimal transport theory and Wasserstein metric between probability measures, which provides an appropriate tool for comparing probability distributions. In this framework, the map T_S will be a random transport plan between the distributions $\mathcal{L}(X \mid S)$ and $\mathcal{L}(\tilde{X})$. Moreover, we will first recall the definition of Wasserstein barycenters which are often chosen in the statistical literature as new distribution ν.

3.2 Wasserstein distance and Wasserstein barycenters

Consider the space $\mathcal{P}_2 \equiv \mathcal{P}_2(\mathbb{R}^d)$ of Borel probabilities on $\mathbb{R}^d$ with finite second moment. The related set $\mathcal{P}_{2,ac} \equiv \mathcal{P}_{2,ac}(\mathbb{R}^d)$ will denote the subset of $\mathcal{P}_2(\mathbb{R}^d)$ containing the probabilities that are absolutely continuous with respect to Lebesgue measure. Given $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we denote by $\Pi(\mu, \nu)$ the set of all probability measures π over the product set $\mathbb{R}^d \times \mathbb{R}^d$ with first (resp. second) marginal μ (resp. ν).

The transportation cost with quadratic cost function, or quadratic transportation cost, between these two measures μ and ν is defined as

$$\mathcal{T}_2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \, d\pi(x, y).$$

The quadratic transportation cost allows to endow the set $\mathcal{P}_2$ with a metric by defining the 2-Wasserstein distance between μ and ν as $W_2(\mu, \nu) = \mathcal{T}_2(\mu, \nu)^{1/2}$. More details on Wasserstein distances and their links with optimal transport problems can be found in [Rac84] or [Vil08] for instance.

In the one-dimensional case $W_2(\mu, \nu)$ is simply the L_2-distance between the quantile functions of μ and ν, enabling its direct computation

$$W_2^2(\mu, \nu) = \int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right|^2 dt, \; F \sim \mu, \; G \sim \nu.$$

A distribution π with marginals μ and ν which minimizes (3.2) is called an optimal coupling of μ and ν. Moreover, if μ vanishes on sets of dimension $d-1$, in particular if $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, then there exists an optimal transport map, T, transporting (pushing forward) μ to ν. The following Theorem is a convenient version that can be found in [Vil03, Theorem 2.12] .

Theorem 3.1. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and let $\pi = \mathcal{L}(X, Y)$ be the joint distribution of a pair (X, Y) of R^d-valued random vectors with probability laws $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$.*

(i) The probability distribution π is an optimal coupling of μ and ν if, and only if, there exists a convex lower semi-continuous function $\varphi \in \partial\varphi(X)$ a.s. such that π is concentrated on $\partial\varphi$, the subgradient of φ.

(ii) If we assume that μ does not give mass to sets of dimension at most $d-1$, then there is a unique optimal coupling π of μ and ν, that can be characterized as the unique solution to the Monge transportation problem - an optimal transport map - T, i.e.: $\pi = \mu \circ (Id, T)^{-1}$ (or $Y = T(X)$ a.s.), and

$$\begin{aligned} W_2^2(\mu, \nu) &= \int \|x - T(x)\|^2 \, d\mu(x) \\ &= \inf \left\{ \int \|x - S(x)\|^2 \, d\mu(x), \; \textit{where } S \textit{ satisfies } \nu = \mu \circ S^{-1} \right\} \end{aligned}$$

Such a map is characterized as $T = \nabla\varphi$ $\mu - a.s.$, the $\mu - a.s.$ unique function that maps μ to ν and that is the gradient of a lower semi-continuous funtion φ. In the following we will use the notation

$$\nu = T_\sharp \mu = \mu \circ T^{-1}.$$

We point out that the existence of the o.t map is commonly referred to as Brenier's theorem and originated from Y. Brenier's work in the analysis and mechanics literature. However, it is worthwile pointing out that a similar statement was established earlier independently in a

probabilistic framework by J.A. Cuesta-Albertos and C. Matrán [CM89] : they show existence of an optimal transport map for quadratic cost over Euclidean and Hilbert spaces, and prove monotonicity of the optimal map in some sense (Zarantarello monotonicity).

When dealing with a collection of distributions $\mu_1, \dots, \mu_J$, we can define a notion of variation of these distributions. For any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, set

$$V_2(\nu) = \sum_{j=1}^{J} \omega_j W_2^2(\nu, \mu_j)$$

where $\omega_1, \dots, \omega_J$ are positive real numbers such that $\sum_{j=1}^{J} \omega_j = 1$. This quantity provides a global measure of separation between the probabilities μ_j, $j = 1, \dots, J$, with respect to fixed weights and has been received recently. Finding the distribution minimizing the variance of the distributions has been tackled when defining the notion of barycenter of distributions with respect to Wasserstein's distance in the seminal work of [AC11]. More precisely, given $p \geq 1$, they provide conditions to ensure existence and uniqueness of the barycenter of the probability measures $(\mu_j)_{1 \leq j \leq J}$ with weights $(\omega_j)_{1 \leq j \leq J}$, i.e. a minimizer of the following criterion

$$\nu \mapsto \sum_{j=1}^{J} \omega_j W_2^2(\nu, \mu_j). \tag{3.3}$$

Such a minimizer, μ_B, is called a barycenter or Fréchet mean of $\mu_1, \dots, \mu_J$, with respect to the weights $\omega_1, \dots, \omega_J$. Empirical versions of the barycenter and their properties are analyzed in [BLGL$^+$15] or [LGL17]. Similar ideas have also been developed in [CD14] or [BK12]. Hence the Wasserstein barycenter distribution appears to be a meaningful feature to represent the mean variations of a set of distributions.

We point out that its computation is a difficult issue for the general case. Yet, in this work, we only consider barycenter between two probabilities $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$. For the one dimensional case, the solution proposed in [FFM$^+$15] to repair the data is to map these distributions to a distribution whose quantile function is defined by taking the mean of the quantile functions of μ_0 and μ_1. This corresponds actually to the minimizer of (3.3) for distributions on the real line denoted by $\mathcal{P}_{2,ac}(\mathbb{R})$. In the following, we present in Section 5 how to compute a barycenter between two distributions in higher dimensions and propose in Section 4 a justification for using the Wasserstein barycenter to repair the data.

4 Full and Partial Repair with Wasserstein Barycenter

In our particular problem, where we have $J = 2$, the two conditional distributions of the random variable X by the protected attribute S are going to be transformed into the distribution of the Wasserstein barycenter μ_B between μ_0 and μ_1, with weights π_0 and π_1, defined as

$$\mu_B \in \mathit{argmin}_{\nu \in \mathcal{P}_2} V_2^2(\mu_0, \mu_1; \pi_0, \pi_1) = \mathit{argmin}_{\nu \in \mathcal{P}_2} \left\{ \pi_0 W_2^2(\mu_0, \nu) + \pi_1 W_2^2(\mu_1, \nu) \right\}.$$

Let $\tilde{X}$ be the transformed variable with distribution μ_B. For each $S = s$, the deformation will be performed through the optimal transport map $T_s : \mathbb{R}^d \to \mathbb{R}^d$ pushing each μ_s towards the weighted barycenter μ_B, whose existence is guaranteed as soon as μ_s are absolutely continuous with respect to Lebesgue measure using Theorem 3.1, which satisfies

$$\mathbb{E}\left(\|X - T_s(X)\|^2 \mid S = s \right) = W_2^2(\mu_s, \mu_B). \tag{4.1}$$

Remark 4.1. *Note first that in the particular setting of two distributions, the computation of the barycenter of the two measures is equivalent to the computation of the optimal transport map between them. More precisely, if we assume that* $\mu_0 \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ *and denote by* $T : \mathbb{R}^d \to \mathbb{R}^d$ *the optimal transport map between* μ_0 *and* μ_1*, that is* $\mu_1 = \mu_{0\sharp}T$*, then we can write*

$$\mu_\lambda = \mu_0\sharp\left((1-\lambda)Id + \lambda T\right),$$

where the map $(1-\lambda)Id + \lambda T$ *is an optimal transport plan, for all* $\lambda \in [0,1]$*. We have that the measure* μ_λ *is the weighted barycenter between* μ_0 *and* μ_1*, with weights* $1-\lambda$ *and* λ*, respectively. So, the complexity of computing* $\mu_B = \mu_0\sharp\left(\pi_0 Id + \pi_1 T\right)$ *is the same as the complexity of computing* T*.*

Remark 4.2. *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 (4.1). Given* $S = s$ *and* $X \in \mathbb{R}$*, let* $F_s : \mathbb{R} \to [0,1]$ *denote the cumulative distribution function of* X *given that* $S = s$*, and* $F_s^{-1} : [0,1] \to \mathbb{R}$ *its quantile associated function. The weighted Wasserstein barycenter* μ_B *of the two distributions* μ_0 *and* μ_1 *is the unique minimizer of the functional (3.3) 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, we note that

$$F_s\left(X \mid S = s\right) \sim \mathcal{U}(0,1), \ s = 0, 1,$$

and the optimal transport map solution to (4.1) is $T_s = F_B^{-1} \circ F_s$*.*

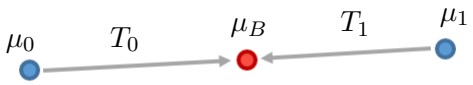

Figure 2: Repairing scheme towards the Wassersein barycenter

4.1 Total repair

To understand the use of the Wasserstein barycenter distribution as the target distribution for μ_0 and μ_1, we quantify the amount of information lost by replacing the distribution of X by the distribution of $\tilde{X}$ obtained by transporting these two distributions. Set the random transport plan $T_S : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, and the modified variable $\tilde{X} = T_S(X)$. We point out that choosing the distribution of $\tilde{X}$ amounts to choose the transportation plans T_0 and T_1.

We are facing learning problems in two different settings.

- On the one hand, the full information available is the input variables X and also the protected variables S which play an important role in the classification, since the classifier has a different behavior according to the class $S = 0$ 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 of the input data is at hand, $\tilde{X}$. Hence learning a classifier amounts to minimizing

$$R(h, \tilde{X}) := \mathbb{P}(h(\tilde{X}) \neq Y).$$

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 [BBL04] for instance. We obtain

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

Denoting the conditional expectation as

$$\eta_s(x) = \mathbb{P}(Y = 1 \mid X = x, S = s), \tag{4.2}$$

we can write that

$$\begin{aligned}
\mathbb{P}(g(X, S) \neq Y \mid X = x, S = s) &= (1 - \mathbb{1}_{g(x,s)=0})(1 - \eta_s(x)) + \mathbb{1}_{g(x,s)=0}\eta_s(x) \\
&= \mathbb{1}_{g(x,s)=0}(2\eta_s(x) - 1) + 1 - \eta_s(x).
\end{aligned}$$

Finally, we get

$$R(g, X, S) = \mathbb{E}\left[\mathbb{1}_{g(X,S)=0}(2\eta_S(X) - 1)\right] + \mathbb{E}\left[1 - \eta_S(X)\right]. \tag{4.3}$$

The minimum risk is thus obtained using the Bayes' rule $g_B(x, s) = \mathbb{1}_{\eta_s(x)>1/2}$, leading to

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

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

$$R(h, \tilde{X}) = R(h, T_S(X)) = \mathbb{E}\left[\mathbb{1}_{h\circ T_S(X)=0}(2\eta_S(X) - 1)\right] + \mathbb{E}\left[1 - \eta_S(X)\right]. \tag{4.4}$$

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

Theorem 4.3. *Consider $X \in \mathbb{R}^d$ and $S \in \{0, 1\}$. Let $T_S : \mathbb{R}^d \to \mathbb{R}^d$, $d \geqslant 1$ be a random transformation of X such that $\mathcal{L}(T_0(X) \mid S = 0) = \mathcal{L}(T_1(X) \mid S = 1)$, and consider the transformed version $\tilde{X} = T_S(X)$. For each $s \in \{0, 1\}$, assume that the function $\eta_s(X)$ defined in (4.2) is Lipschitz with constant $K_s > 0$ Then, if $K = \max\{K_0, K_1\}$,*

$$\mathcal{E}(\tilde{X}) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s W_2^2(\mu_s, \mu_{s\sharp}T_s)\right)^{\frac{1}{2}}. \tag{4.5}$$

The proof of this theorem which relies on the following lemma is postponed to the Appendix.

Lemma 4.4. *Under Assumptions of Theorem 4.3, the following bound holds*

$$R(g_B \circ T_S, X) - R(g_B, X, S) \leq 2\mathbb{E}\left[|\eta_S(X) - \eta_S \circ T_S(X)|\right].$$

Hence, Theorem 4.3 provides some justification to the use of the Wasserstein barycenter as the distribution of the modified variable. Actually, minimizing the upper bound in (4.5) with respect to the function $T_S : \mathbb{R}^d \to \mathbb{R}^d,\ d \geqslant 1$, leads to consider the transport plan carrying the conditional distributions $\mu_s,\ s = 0, 1$, 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 [FFM+15]. This leads to the following bound

$$\begin{aligned}
\inf_{T_S}\{R(g_B \circ T_S, X) - R(g_B, X, S)\} &\leq 2\sqrt{2}K \inf_{T_S}\left(\sum_{s=0,1}\pi_s W_2^2(\mu_s, \mu_{s\sharp}T_s)\right)^{\frac{1}{2}} \\
&= \sqrt{2}K\left(\sum_{s=0,1}\pi_s W_2^2(\mu_s, \mu_B)\right)^{\frac{1}{2}} \\
&= \sqrt{2}K\pi_0\pi_1 W_2^2(\mu_0, \mu_1)
\end{aligned}$$

Yet this bound is only an upper bound which only provides some guidelines on the choice of the choice of the distribution to which the conditional distribution have to be mapped. Nevertheless, choosing the Wasserstein barycenter provides a simply a reasonable and, more important, feasible solution for fairness, to achieve statistical parity.

4.2 Partial repair

As pointed out in previous section, 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 [FFM+15], called *Geometric Repair*. The authors propose not to move all conditional distributions to the barycenter but only towards the barycenter on the Wasserstein's geodesic path between μ_0 and μ_1. We analyze next this procedure and propose an alternative to this choice.

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 optimal transport map pushing each μ_s towards the target μ. In the literature, μ is chosen as the barycenter μ_B and the *Partially Repaired* conditional distributions for $s \in \{0, 1\}$ are defined as

$$\mu_{s,\lambda} = \mathcal{L}(\lambda Z + (1-\lambda)R_s(Z)) = \mathcal{L}(\lambda T_s(X) + (1-\lambda)X \mid S = s).$$

Figure 3: Original distributions (blue) and their partially repaired versions (green) towards the barycenter (red)

This procedure is represented in Figure 3. 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. Choosing the parameter λ should be a trade-off between, on the one hand, accuracy of the classification error that leads to little change in the

initial distribution, and, on the other hand, non predictability of the protected variable which implies that the two conditional distribution should be close with respect to the total variation distance.

Arguing among the lines of previous section to obtain an upper bound for the classification risk using the two distributions $\mu_{s,\lambda},\ s \in \{0,1\}$, does not lead to a satisfying result. This comes from the fact that we move distributions according to Wasserstein distance, while fairness is measured using the total variation distance and they are of different nature. In fact, the distance in total variation between two probabilities P and Q can be computed as

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

see, e.g. [Mas07].

So if $\lambda \in (0,1)$, this implies that

$$d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) \leq \mathbb{P}(\lambda Z + (1-\lambda)R_0(Z) \neq \lambda Z + (1-\lambda)R_1(Z)) = \mathbb{P}(R_0(Z) \neq R_1(Z)). \quad (4.6)$$

Previous bound means that the amount of repair quantified by the parameter λ does not affect the distance in Total Variation between the modified conditional distributions. Moreover, in some situations, (4.6) turns out to be an equality. Consider, for instance,

$$\mu_{0,0} = U(K, K+1)$$

$$\mu_{1,0} = U(-K-1, -K)$$

as the distributions of X in each class. Then, the barycenter is $\mu_{0,1} = \mu_{1,1} = U(-1/2, 1/2)$ and the partially repaired distributions are

$$\mu_{0,\lambda} = U(-\lambda/2 + (1-\lambda)K, -\lambda/2 + (1-\lambda)K + 1)$$

$$\mu_{1,\lambda} = U(-\lambda/2 - (1-\lambda)(K+1), -\lambda/2 - (1-\lambda)(K+1) + 1).$$

In this particular case, the distance in total variation can be easily computed as

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

As a consequence, $d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) = 1$, 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, this upper bound provides some argument against the use of this kind of repair since the reparation should favor small distance between these two 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 following approach called *Random Repair*, that enables a better control of their Total Variation distance.

Let Z be a target variable with general distribution μ and let B be a Bernoulli variable with parameter λ, $B \sim \mathcal{B}(\lambda)$, independent of (X, S, Y). Note that $R_0(Z)$ and $R_1(Z)$ follow the original conditional distributions μ_0 and μ_1. Let us consider the following repair procedure which consists in randomly changing the original distribution of the variables X by either selecting the target distribution μ or the original conditional distributions. The choice between both is governed by the Bernoulli parameter λ. Define for $s \in \{0,1\}$, the repaired distributions

$$\tilde{\mu}_{s,\lambda} = \mathcal{L}(BZ + (1-B)R_s(Z)) = \mathcal{L}(BT_s(X) + (1-B)X \mid S = s). \quad (4.7)$$

Note that, similarly as in the Geometric Repair, for $\lambda = 0$ $\tilde{\mu}_{s,0} = \mathcal{L}(X \mid S = s)$ and for $\lambda = 1$ $\tilde{\mu}_{s,1} = \mathcal{L}(Z) = \mu$. Unlike the previous procedure, in this setting, the parameter λ does play a role in controlling the distance between the repaired distributions

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

Hence, this bound suggests that λ should be close to 1 to ensure non predictability of the protected attribute.

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. Thus the use of the Wasserstein barycenter $Z \sim \mu_B$ is still justified.

$$\begin{aligned} &R(g, \tilde{X}_\lambda) = \mathbb{P}(g(\tilde{X}_\lambda) \neq Y) \\ &= (1-\lambda)\mathbb{P}(g(BT_S(X) + (1-B)X) \neq Y \mid B = 0) + \lambda\mathbb{P}(g(BT_S(X) + (1-B)X) \neq Y \mid B = 1) \\ &= (1-\lambda)\mathbb{P}(g(X) \neq Y \mid B = 0) + \lambda\mathbb{P}(g(T_S(X)) \neq Y \mid B = 1) \\ &= (1-\lambda)\mathbb{P}(g(X) \neq Y) + \lambda\mathbb{P}(g(T_S(X)) \neq Y). \end{aligned}$$

Therefore, in the following we promote the use of Random Repair to enhance Disparate Impact while not hampering too much the efficiency of the classification. This will be studied in the following section.

5 Numerical Analysis of Fair Correction of a database

As the distributions at hand are empirical, the existence of an optimal transport map is not guaranteed and the repair procedure in section 4 that blurs the protected variable in the original data must be adapted. In this section, we propose a new algorithm to carry this out, which in practice, achieves total fairness in contrast with the existing in the literature.

5.1 Computational aspects

Let $\{(X_i, S_i, Y_i), i = 1, \ldots, N\}$ be an observed sample of (X, S, Y), and denote by n_0 and n_1 the number of instances in each protected class. For ease of exposition and without loss of generality, suppose that the observations are ordered by the value of S, so we can write

$$\begin{aligned} x_{0,i} &:= X_i, \qquad \text{if } s_i = 0,\ i = 1, \ldots, n_0, \\ x_{1,j-n_0} &:= X_j, \quad \text{if } s_j = 1,\ j = n_0 + 1, \ldots, N = n_0 + n_1. \end{aligned}$$

Generally, the sizes n_0 and n_1 of the samples $\mathcal{X}_0 = \{x_{0,1}, \ldots, x_{0,n_0}\}$ and $\mathcal{X}_1 = \{x_{1,1}, \ldots, x_{1,n_1}\}$ 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 $c_{ij} = \|x_{0,i} - x_{1,j}\|^2, 1 \leq i \leq n_0,\ 1 \leq j \leq n_1$. 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

$$\begin{cases} \min_\gamma \sum_{\substack{1\leq i\leq n_0 \\ 1\leq j\leq n_1}} c_{ij}\gamma_{ij}, & \\ \text{subject to} & \gamma_{ij} \geqslant 0, \\ & \sum_{i=1}^{n_0} \gamma_{ij} = \frac{1}{n_1}, \text{ for all } j, \\ & \sum_{j=1}^{n_1} \gamma_{ij} = \frac{1}{n_0}, \text{ for all } i. \end{cases} \tag{5.1}$$

If $\hat{\gamma}$ is a solution to the linear program (5.1) then, acordingly to Remark 4.1, the distribution

$$\mu_{B,n} = \sum_{\substack{1\leq i\leq n_0 \\ 1\leq j\leq n_1}} \hat{\gamma}_{ij}\delta_{\{\pi_0 x_{0,i}+\pi_1 x_{1,j}\}}$$

is a barycenter of $\mu_{0,n}$ and $\mu_{1,n}$ with respect to weights π_0 and π_1. See [CD14] for details on the discrete Wasserstein and Optimal Transport computation.

5.1.1 Total repair

In practice, the implementation of the repair scheme in section 4 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 $\tilde{\mathcal{X}}$. In the following, we detail two different methods, of which the first one is similar to some existing in the literature and does not achieve total fairness in the practical framework, while the second one is a novelty and does guarantee this property for the new data $\tilde{\mathcal{X}}$.

(A) On the one hand, as depicted in Figure 4 (left), each original point in $\mathcal{X}_0, \mathcal{X}_1$ is changed by a unique point given by

$$\tilde{x}_{0,i} = \frac{\sum_{j=1}^{n_1} \gamma_{ij}\left(\pi_0 x_{0,i} + \pi_1 x_{1,j}\right)}{\sum_{j=1}^{n_1} \gamma_{ij}} = \pi_0 x_{0,i} + n_0\pi_1 \sum_{j=1}^{n_1} \gamma_{ij} x_{1,j},\ 1 \leq i \leq n_0,$$

$$\tilde{x}_{1,j} = \frac{\sum_{i=1}^{n_0} \gamma_{ij}\left(\pi_0 x_{0,i} + \pi_1 x_{1,j}\right)}{\sum_{i=1}^{n_0} \gamma_{ij}} = n_1\pi_0 \sum_{i=1}^{n_0} \gamma_{ij} x_{0,i} + \pi_1 x_{1,j},\ 1 \leq j \leq n_1.$$

Doing this, the set $\tilde{\mathcal{X}}$ will be a collection of exactly $n_0 + n_1$ points. This approach generalizes to higher dimensions the idea of previous works [FFM$^+$15] and [JL17], who just considere the unidimensional case, where the transport is written in terms of the distribution funtions. However, in practise it generates two sets $\tilde{\mathcal{X}}_0 = \{x_{0,i},\ 1 \leqslant i \leqslant n_0\}$ and $\tilde{\mathcal{X}}_1 = \{x_{1,j},\ 1 \leqslant j \leqslant n_1\}$ that are not the same and do not reach (3.1).

(B) To ensure total fairness, each point cannot be changed by a unique repaired point. Hence, each point will split its mass to be transported into several modified versions, generating an extended set $\tilde{\mathcal{X}} = \tilde{\mathcal{X}}_0 \cup \tilde{\mathcal{X}}_1$, which is formed by the complete distribution $\mu_{B,n}$. More precisely, as represented in Figure 4 (right), for every $1 \leq i \leq n_0,\ 1 \leq j \leq n_1$, if $\hat{\gamma}_{ij} > 0$ we define two points

$$\tilde{x}_{0,i,j} = \tilde{x}_{1,j,i} := \pi_0 x_{0,i} + \pi_1 x_{1,j}, \tag{5.2}$$

and the sets

$$\tilde{\mathcal{X}}_0 := \bigcup_{1 \leq i \leq n_0} \{\tilde{x}_{0,i,j} \ / \ \hat{\gamma}_{ij} > 0, 1 \leq j \leq n_1\}$$
$$\tilde{\mathcal{X}}_1 := \bigcup_{1 \leq j \leq n_1} \{\tilde{x}_{1,j,i} \ / \ \hat{\gamma}_{ij} > 0, 1 \leq i \leq n_0\}.$$

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 5.1. We have simulated two samples $\mathcal{X}_0$ and $\mathcal{X}_1$ of points in $\mathbb{R}$ of sizes $n_0 = 4$ and $n_1 = 7$, respectively,. The optimal matrix solution to the problem (5.1) 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 $\mathcal{L}(X \mid S = 0)$ and $\mathcal{L}(X \mid S = 1)$, respectively, then the left part of Figure 4 represents the blurring 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, the scheme on the right carries out the procedure **(B)**, and we note that $\tilde{\mathcal{X}}_0 = \tilde{\mathcal{X}}_1$.

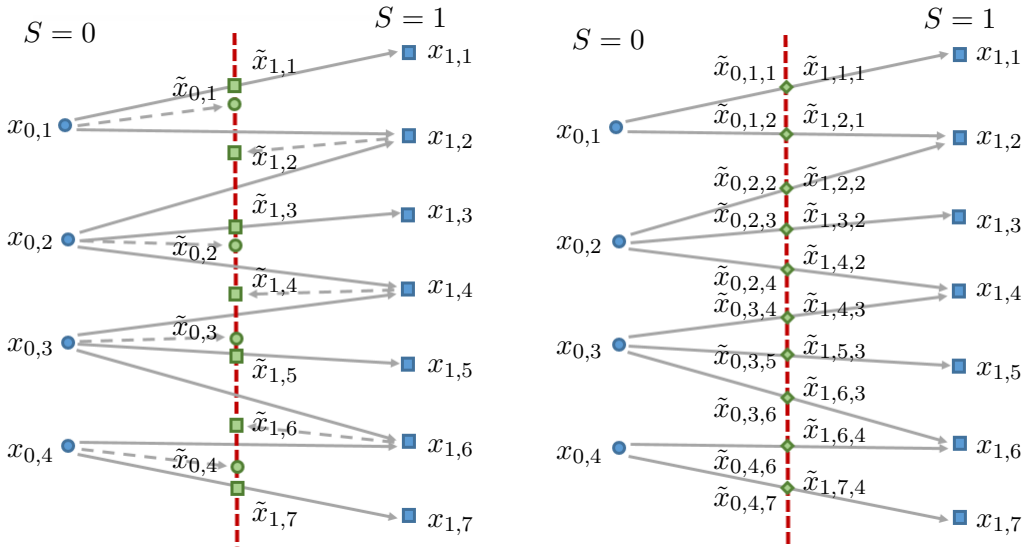

Figure 4: Example of the two repairing processes when $n_0 = 4$ and $n_1 = 7$.

Remark 5.1. *In the special situation when the two samples $\mathcal{X}_0$ and $\mathcal{X}_1$ have equal size n and all weights are uniform, that is $\gamma_{ij} = \frac{1}{n},\ 1 \leq i, j \leq n$, the mass conservation constraint implies that γ is a bijection and the Monge problem is equivalent to the optimal matching problem*

$$\min_{\sigma \in Perm(n)} \frac{1}{n} \sum_{i=1}^{n} c_{i,\sigma(i)}.$$

*Every point in each original set will be modified by a unique point as depicted in Figure 5. In this case, both repairing procedures **(A)** and **(B)** perfom in the same way, and total fairness is always achieved:*

$$\tilde{x}_{0,i} = \tilde{x}_{1,i} = \frac{1}{2}\left(x_{0,i} + x_{1,i}\right),\ 1 \leq i \leq n.$$

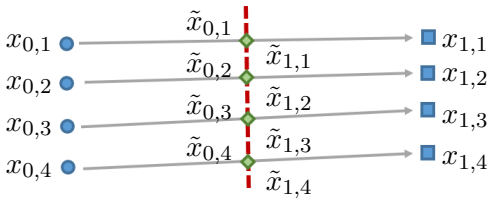

Figure 5: Repairing process when both protected groups have the same number of instances.

5.1.2 Random repair

As previously noted, trying to build the set $\tilde{\mathcal{X}}$ satisfying the goal (3.1) may compromise too much the accuracy of the classification with these new data. In this sense, the Random Repair procedure proposed in section 4.2 aims at setting a tradeoff between fairness and accuracy through the parameter λ, that models the amount of repair desired.

In this section, we detail how to compute the randomly repaired set denoted by $\tilde{\mathcal{X}}_\lambda$, with respect to parameter $\lambda \in [0, 1]$. According to (4.7), we will randomly select either the points in the original samples $\mathcal{X}_0$ and $\mathcal{X}_1$ or their repaired sequels generated with procedure **(B)** in Figure 4 (right). More precisely, consider a sample $\{b_l\}_{l=1,\ldots,n_0+n_1} \sim B(\lambda),\ \lambda \in [0, 1]$, and define

$$\tilde{\mathcal{X}}_{0,\lambda} := \bigcup_{1 \leq i \leq n_0} R_{0,i,\lambda} \tag{5.3}$$

$$\tilde{\mathcal{X}}_{1,\lambda} := \bigcup_{1 \leq j \leq n_1} R_{1,j,\lambda}, \tag{5.4}$$

where $R_{0,i,\lambda}$ and $R_{1,j,\lambda}$ are the repaired sets of 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 Equation (5.2), with weights $\hat{\gamma}_{i,j}$.

Example 5.2. Consider again the situation in Example 5.1. Figure 6 represents the Random Repair procedure for $\lambda = \frac{1}{2}$. We have simulated values $b_l \sim \mathcal{B}(\frac{1}{2}),\ l = 1, \ldots, n_0 + n_1 = 11$, and

the resulting sets $R_{0,i,\lambda}, 1 \leq i \leq 4$, and $R_{1,j,\lambda}, 1 \leq j \leq 7$. Finally, from (5.3) and (5.4) we have the randomly repaired sets

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

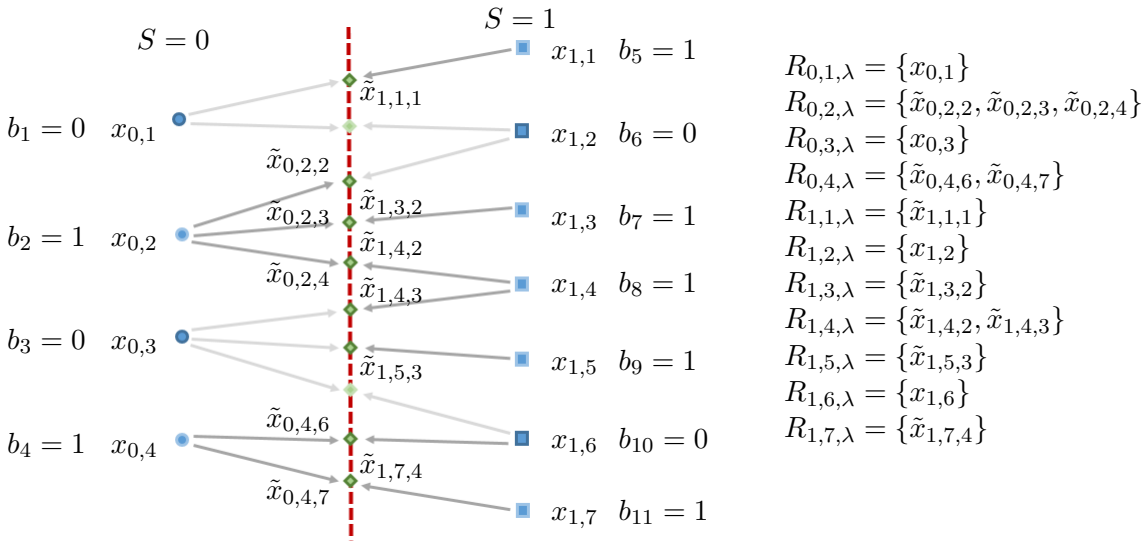

Figure 6: Example of the Random Repair algorithm with $\lambda = \frac{1}{2}$.

5.2 Application to a real example

To ilustrate the performance of the proposed repairing procedures in section 4 we consider the *Adult Income* data set. It contains 29.825 instances consisting in the values of 14 attributes, 6 numeric and 8 categorical, and a categorization of each person as having an income of more or less than $50,000$\$ per year. This attribute will be the target variable in the study. In the following, we estimate the Disparate Impact using its empirical counterpart and provide a confidence interval which was established in [BDBGL18]. Among the rest of the categorical attributes, we focus on the sensitive attribute *Gender* ("male"or "female") to be the potentially protected. As the repairing procedures work only with the numerical attributes, to check their effectiveness we will follow the next steps:

1. Split the data set into the test and the learning sample using the ratio 2.500 / 27.325.
2. Adjust a statistical model with the learning sample to predict the target attribute using the five numerical variables: *Age, Education Level, Capital Gain, Capital Loss* and *Worked hours per week*. We have trained the classifiers based on logistic regression and random forests.
3. Predict the target for the test sample with the built model and compute the misclassification error of each rule.
4. Apply the repair procedure in $\mathbb{R}^5$ to the test sample described by these numerical variables.
5. Predict the target for the repaired data set with the built model and compute the misclassification error again.

In Table 1 a summary of the performance of the two classification rules considered is presented. With a confidence of 95%, we can say that the logit classifier has Disparate Impact at level 0.555 and the Random Forests at 0.54, with respect to Gender. Hence, both rules are committing discrimination with respect to this sensitive variable. Now we will see how the repairing procedures studied in section 4 help in blurring the protected variable.

In Table 2 we can see 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 in such a way that both classification rules attain Statistical Parity. Moreover, the error in the classification done with the repaired data sets is smaller when using procedure **(B)** in the two cases. In [FFM$^+$15], they propose a generalization to higher dimension by computing the repairing procedure for each attribute. This procedure is denoted in the table with the letter **(C)**. We see that the error is smaller than with **(A)** but still much bigger than with **(B)**. Moreover, the estimated level of Disparate Impact is not 1 but it is closer to the Statistical Parity than with procedure **(A)**.

Finally, we present some results of the performance of the Geometric and Random Repairs. Figures 7 and 9 represent the evolution of the estimated Disparate Impact with the amount of repair $0 \leq \lambda \leq 1$. The Figures 8 and 10 show the evolution with λ of the error in the classification done from the modified data set. For the experiments concerning the Random Repair procedure (denoted RR in the figures) we have repeated it 100 times, and then we have computed the mean of the simulations. Clearly, the level of DI reached is higher with the Random Repair for the logit rule. For the random forest procedure since the rule is not linear, the difference is not as high and Disparate Impacts have similar behaviors. Yet for larger amount of repair the gap between the two different kinds of repair increases at the advantage of the Geometric Repair. Moreover, the error in the prediction from the new data modified with this procedure is smaller than with the Geometric Repair. We note that the amount of repair necessary to achieve a confidence interval for DI at level 0.8 for the logit rule is

- 0.3 with the Random Repair, which entails an error of 0.2068
- 0.55 with the Geometric Repair, which entails an error of 0.2136.

In the case of the random forests rule, this value is 0.5 for both but the error is

- 0.1927 with the Random Repair
- 0.2076 with the Geometric Repair

Statistical Model	Error	$\hat{DI}$	CI 95%
Logit	0.2064	0.496	$(0.437, 0.555)$
Random Forests	0.168	0.484	$(0.429, 0.54)$

Table 1: Performance and Disparate Impact with respect to the protected variable Gender.

Statistical Model	Repair	Error	Difference	$\hat{DI}$	CI 95%
Logit	**(A)**	0.218	0.0116	0.937	$(0.841, 1.033)$
Logit	**(B)**	0.2077	0.00128	1	$(0.905, 1.095)$
Logit	**(C)**	0.2132	0.0068	0.94	$(0.842, 1.038)$
Random Forests	**(A)**	0.2272	0.0592	1.1	$(0.976, 1.223)$
Random Forests	**(B)**	0.2045	0.0365	1	$(0.886, 1.114)$
Random Forests	**(C)**	0.2152	0.0472	1.091	$(0.978, 1.203)$

Table 2: Repairing procedures and Disparate impact of the rules with the modified dataset

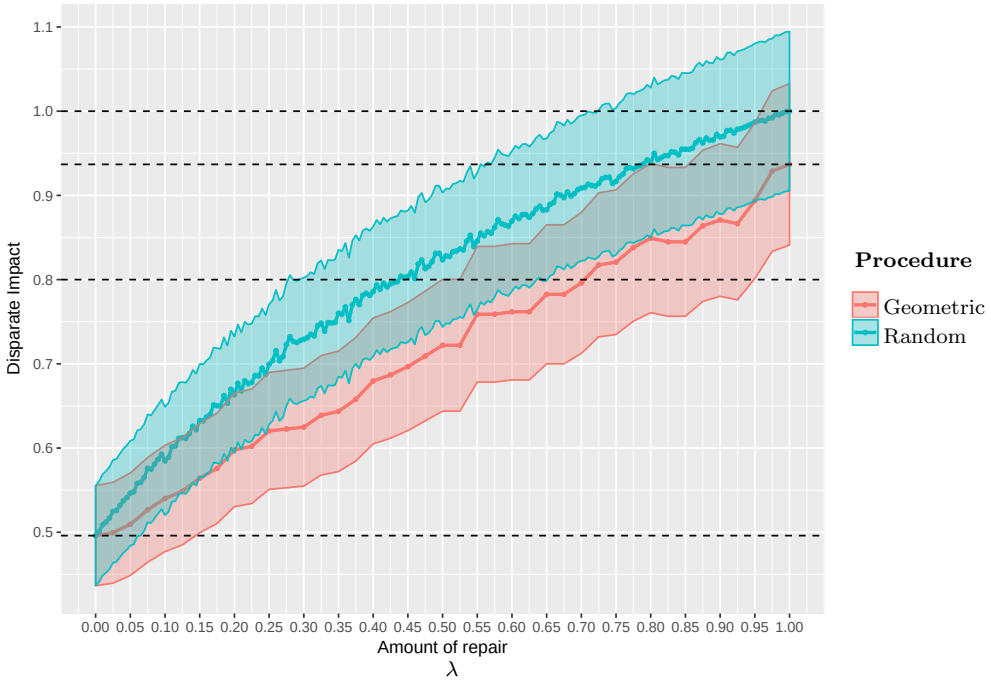

Figure 7: Confidence interval at level 95% for DI of the classifier logit with respect to Gender and the data repaired by the Geometric (red) and Random (blue) Repairs

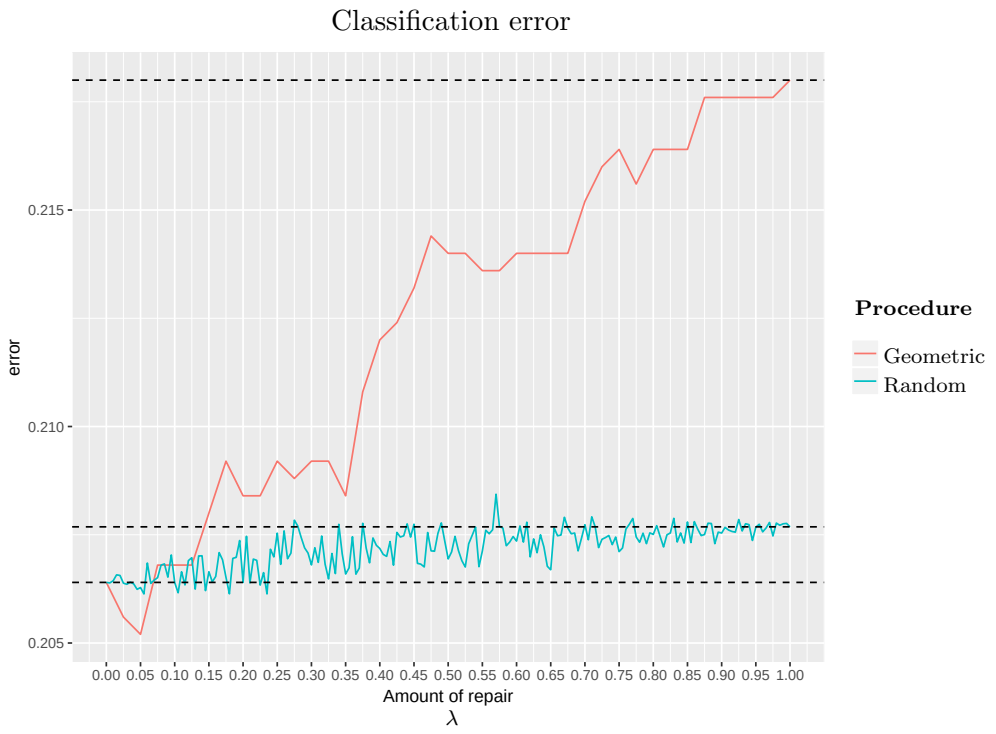

Figure 8: Misclassification error in the prediction with the classifier logit and the data repaired by the Geometric (red) and Random (blue) Repairs

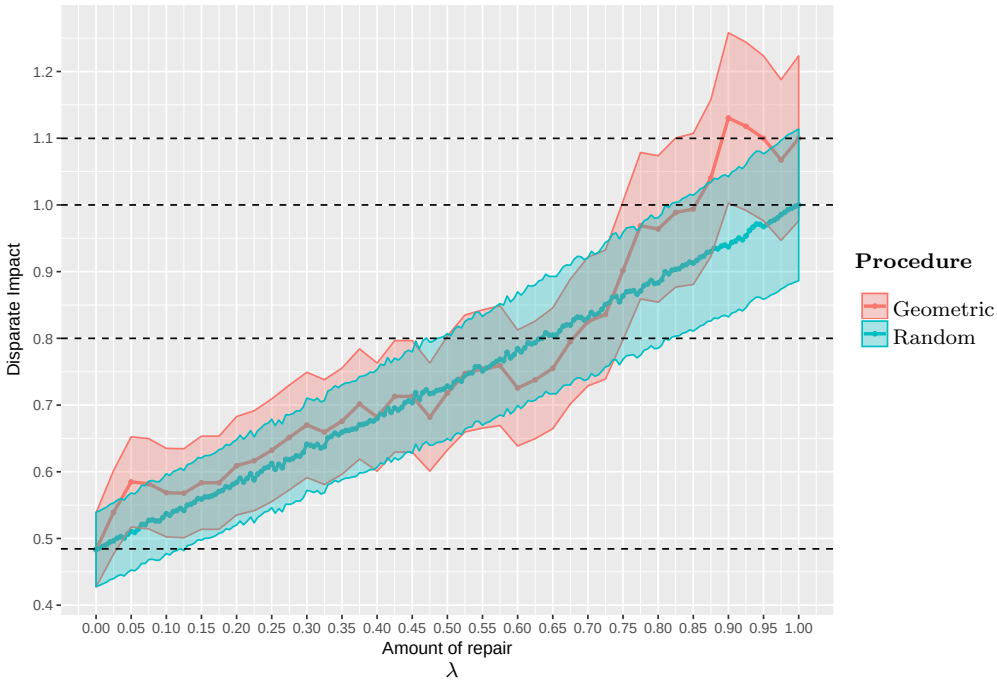

Figure 9: Confidence interval at level 95% for DI of the classifier Random Forests with respect to Gender and the data repaired by the Geometric (red) and Random (blue) Repairs

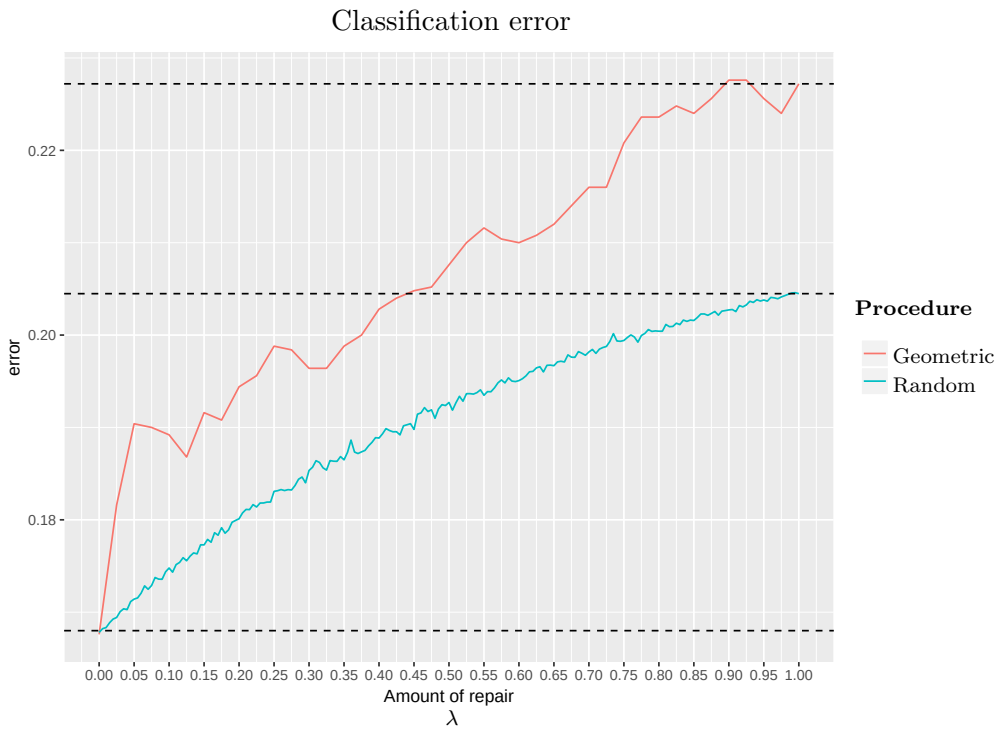

Figure 10: Misclassification error in the prediction with the classifier Random Forests and the data repaired by the Geometric (red) and Random (blue) Repairs

6 Appendix

Proof of Theorem 2.1.

Proof. We will show that the conditions $DI(g, X, S) \leq \tau$ and $BER(g, X, S) \leq \frac{1}{2} - \frac{a(g)}{2}(\frac{1}{\tau} - 1)$ are equivalent, for all $g \in \mathcal{G}$. Indeed, given $g \in \mathcal{G}$,

$$
\begin{aligned}
BER(g, X, S) \leq \frac{1}{2} - \frac{a(g)}{2}\left(\frac{1}{\tau} - 1\right) = \frac{1}{2} - \frac{(\frac{1}{\tau} - 1)}{2}\mathbb{P}(g(X) = 1 \mid S = 0) \\
\Leftrightarrow \mathbb{P}\left(g(X) = 0 \mid S = 1\right) + \mathbb{P}\left(g(X) = 1 \mid S = 0\right) \leq 1 - \left(\frac{1}{\tau} - 1\right)\mathbb{P}(g(X) = 1 \mid S = 0) \\
\Leftrightarrow \left(1 + \left(\frac{1}{\tau} - 1\right)\right)\mathbb{P}\left(g(X) = 1 \mid S = 0\right) + \mathbb{P}\left(g(X) = 0 \mid S = 1\right) \leq 1 \\
\Leftrightarrow \frac{1}{\tau}\mathbb{P}\left(g(X) = 1 \mid S = 0\right) \leq 1 - \mathbb{P}\left(g(X) = 0 \mid S = 1\right) = \mathbb{P}\left(g(X) = 1 \mid S = 1\right) \\
\Leftrightarrow DI(g, X, S) = \frac{\mathbb{P}\left(g(X) = 1 \mid S = 0\right)}{\mathbb{P}\left(g(X) = 1 \mid S = 1\right)} \leq \tau.
\end{aligned}
$$

□

Proof of Theorem 2.2

Proof. For this, we denote by $f_i, i = 0, 1$, the density functions of the conditioned variables $X/S = i$, respectively, whose corresponding probability measures are both supposed to be, without loss of generality, absolute continuous with respect to a measure μ. In general, the

misclassification error could be written as:

$$\mathbb{P}(g(X) \neq S) = \mathbb{P}(S=0)\mathbb{P}\left(g(X)=1 \mid S=0\right) + \mathbb{P}(S=1)\mathbb{P}\left(g(X)=0 \mid S=1\right) = \\ \mathbb{P}(S=0)\int_{g(X)=1} f_0(x)d\mu(x) + \mathbb{P}(S=1)\int_{g(X)=0} f_1(x)d\mu(x). \quad (6.1)$$

Now, for $s=0,1$, we fixe the value of $\pi_s = \mathbb{P}(S=s)$, and from the Bayes' Formula, we know that

$$\mathbb{P}\left(S=s|X\right) = \frac{\pi_s f_s(X)}{\pi_0 f_0(X) + \pi_1 f_1(X)}.$$

Hence,

$$\left\{\mathbb{P}\left(S=0|X\right) > \mathbb{P}\left(S=1|X\right)\right\} = \left\{\pi_0 f_0(X) > \pi_1 f_1(X)\right\},\ \mu - a.s.$$

Thus, we can deduce that the classifier that minimizes the missclassification error rate is

$$g^*(x) = \begin{cases} 1 & if \quad \pi_0 f_0(x) \leq \pi_1 f_1(x) \\ 0 & if \quad \pi_0 f_0(x) > \pi_1 f_1(x) \end{cases},$$

and from equation (6.1),

$$\min_{g\in\mathcal{G}} \mathbb{P}(g(X) \neq S) = \int_{\{\pi_0 f_0(x) \leq \pi_1 f_1(x)\}} \pi_0 f_0(x)d\mu(x) + \int_{\{\pi_0 f_0(x) > \pi_1 f_1(x)\}} \pi_1 f_1(x)d\mu(x).$$

In our particular case, $BER(g,X,S) = \mathbb{P}(g(X)\neq S)$ when considering $\pi_0 = \pi_1 = \frac{1}{2}$, so we have that

$$g^*(x) = \begin{cases} 1 & if \quad f_0(x) \leq f_1(x) \\ 0 & if \quad f_0(x) > f_1(x) \end{cases}$$

and

$$\min_{g\in\mathcal{G}} BER(g,X,S) = BER(g^*,X,S) = \frac{1}{2}\left[\int_{f_0(x)\leq f_1(x)} f_0(x)d\mu(x) + \int_{f_0(x)>f_1(x)} f_1(x)d\mu(x)\right] \\ = \frac{1}{2}\int (f_0 \wedge f_1)(x)d\mu(x).$$

This concludes the proof since by definition

$$d_{TV}(\mu_0,\mu_1) = \frac{1}{2}\int |f_0 - f_1|\, d\mu = 1 - \int (f_0 \wedge f_1)(x)d\mu(x).$$

□

Proof of Lemma (4.4)

Proof. We want to be able to control the difference $\inf_{h\in\mathcal{G}} R(h,\tilde{X}) - \inf_{g\in\mathcal{G}} R(g,X,S)$. To do this, observe that

$$R_B(\tilde{X}) - R_B(X,S) := \inf_{h\in\mathcal{G}} R(h,\tilde{X}) - \inf_{g\in\mathcal{G}} R(g,X,S) \\ \leq R(g_B\circ T_S, X) - R(g_B,X,S) = E\left[(2\eta_S(X)-1)(\mathbb{1}_{g_B\circ T_S(X)=0} - \mathbb{1}_{g_B(X,S)=0})\right] \\ = \mathbb{E}\left[(2\eta_S(X)-1)\mathbb{1}_{g\circ T_S(X)\neq g_B(X,S)}(\mathbb{1}_{g_B\circ T_S(X)\neq 1} - \mathbb{1}_{g_B(X,S)\neq 1})\right],$$

where the last equality holds because $\left(\mathbb{1}_{g_B\circ T_S(X)\neq 1}\right) - \left(\mathbb{1}_{g_B(X,S)\neq 1}\right) = 0$ if, and only if, both classifiers have the same response $g_B\circ T_S(X) = g_B(X,S)$.

Consider $X=x$ and $S=s$,

- if $g_B(x,s)=1$, $2\eta_s(x)-1 \geqslant 0$ and $\mathbb{1}_{g_B(x,s)\neq 1}=0$. In this situation, we deduce that

$$\mathbb{1}_{g_B\circ T_s(x)\neq g_B(x,s)} = 1 \Leftrightarrow g_B\circ T_s(x) = 0,$$

and

$$\mathbb{1}_{g_B\circ T_s(x)\neq 1} - \mathbb{1}_{g_B(x,s)\neq 1} = 1.$$

- if $g_B(x,s)=0$, $2\eta_s(x)-1 < 0$ and $\mathbb{1}_{g_B(x,s)\neq 1}=1$. We deduce that

$$\mathbb{1}_{g_B\circ T_s(x)\neq g_B(x,s)} = 1 \Leftrightarrow g_B\circ T_s(x) = 1,$$

and

$$\mathbb{1}_{g_B\circ T_s(x)\neq 1} - \mathbb{1}_{g_B(x,s)\neq 1} = -1.$$

In any case, the random variable $(2\eta_S(X)-1)\mathbb{1}_{g\circ T_S(X)\neq g_B(X,S)}(\mathbb{1}_{g_B\circ T_S(X)\neq 1} - \mathbb{1}_{g_B(X,S)\neq 1})$ is positive and so it is its expectation

$$R(g_B\circ T_S, X) - R(g_B, X, S) = \mathbb{E}\left[|2\eta_S(X)-1|\,\mathbb{1}_{g\circ T_S(X)\neq g_B(X,S)}\right] \geqslant 0.$$

Moreover, notice that $g_B\circ T_s(x) = \mathbb{1}_{\eta_s\circ T_s(x)>\frac{1}{2}}$, for all x, for all s. Hence, $g_B\circ T_s(x)\neq g_B(x,s)$ if, and only if, either $\eta_s(x)>\frac{1}{2}$ and $\eta_s\circ T_s(x)<\frac{1}{2}$ or $\eta_s(x)<\frac{1}{2}$ and $\eta_s\circ T_s(x)>\frac{1}{2}$. In both cases,

$$|\eta_s(x)-\eta_s\circ T_s(x)| = \left|\eta_s(x)-\frac{1}{2}+\frac{1}{2}-\eta_s\circ T_s(x)\right| = \left|\eta_s(x)-\frac{1}{2}\right| + \left|\frac{1}{2}-\eta_s\circ T_s(x)\right|,$$

and then it is clear that

$$\left|\eta_s(x)-\frac{1}{2}\right| \leq |\eta_s(x)-\eta_s\circ T_s(x)|, \text{ for all } x, \text{for all } s.$$

In conclusion, the difference between the risk using the Bayes' classifier with the original variable X, S and the modified version $\tilde{X}=T_S(X)$ can be bounded as follows

$$R(g_B\circ T_S, X) - R(g_B, X, S) \leq 2\mathbb{E}\left[|\eta_S(X)-\eta_S\circ T_S(X)|\right].$$

□

Proof of Theorem 4.3

Proof. First, note that $R(h,\tilde{X}) = R(h, T_S(X)) \leq R(g_B, T_S(X)) = R(g_B\circ T_S, X)$. Thus, it suffices bounding the difference between the minimal risks obtained for the best classifier with input data (X,S), called g_B, and the risk obtained with this classification rule using the input data $\tilde{X}$

$$\begin{aligned} &R(g_B\circ T_S, X) - R(g_B, X, S) \leq 2\mathbb{E}_{(X,S)}\left[|\eta_S(X)-\eta_S\circ T_S(X)|\right] \\ &= 2\left[\mathbb{P}(S=0)\mathbb{E}_X\left[|\eta_0(X)-\eta_0\circ T_0(X)| \mid S=0\right] + \mathbb{P}(S=1)\mathbb{E}_X\left[|\eta_1(X)-\eta_1\circ T_1(X)| \mid S=1\right]\right] \\ &= 2\sum_{s=0,1}\pi_s\mathbb{E}_X\left[|\eta_s(X)-\eta_s\circ T_s(X)| \mid S=s\right]. \end{aligned}$$

Moreover, by the Lipschitz condition and noting that $a+b \leq 2^{\frac{1}{2}}(a^2+b^2)^{\frac{1}{2}},\ forall\ a,b \in \mathbb{R}$, we can write

$$R(g_B \circ T_S, X) - R(g_B, X, S) \leq 2 \sum_{s=0,1} \pi_s K_s \mathbb{E}_X \left[\|X - T_s(X)\| \mid S = s \right]$$
$$\leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s^2 \left(\mathbb{E}_X \left[\|X - T_s(X)\| \mid S = s \right] \right)^2 \right)^{\frac{1}{2}},$$

where $K = \max\{K_0, K_1\}$. Finally, the Cauchy-Schwarz inequality gives

$$R(g_B \circ T_S, X) - R(g_B, X, S) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s^2 \mathbb{E}_X \left[\|X - T_s(X)\|^2 \mid S = s \right] \right)^{\frac{1}{2}}$$
$$= 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s^2 W_2^2(\mu_s, \mu_{s\sharp} T_s) \right)^{\frac{1}{2}} \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s W_2^2(\mu_s, \mu_{s\sharp} T_s) \right)^{\frac{1}{2}}.$$

□

References

[AC11] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[BBL04] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[BDBGL18] Philippe Besse, Eustasio Del Barrio, Paula Gordaliza, and Jean-Michel Loubes. Statistical tests of unfairness in algorithmic decisions. *working paper*, 2018.

[BHJ+17] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

[BK12] Jérémie Bigot and Thierry Klein. Characterization of barycenters in the wasserstein space by averaging optimal transport maps. *arXiv preprint arXiv:1212.2562*, 2012.

[BL17] Y. Bechavod and K. Ligett. Penalizing Unfairness in Binary Classification. *ArXiv e-prints*, June 2017.

[BLGL+15] Emmanuel Boissard, Thibaut Le Gouic, Jean-Michel Loubes, et al. Distributions template estimate with wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[CD14] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[CM89] Juan Antonio Cuesta and Carlos Matrán. Notes on the wasserstein metric in hilbert spaces. *The Annals of Probability*, pages 1264–1276, 1989.

[DOB+18] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical Risk Minimization under Fairness Constraints. *ArXiv e-prints*, February 2018.

[FFM+15] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[FSV+18] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. *ArXiv e-prints*, February 2018.

[HW17] Philipp Hacker and Emil Wiedemann. A continuous framework for fairness. *CoRR*, abs/1712.07924, 2017.

[JL17] James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*, 2017.

[LGL17] Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

[Mas07] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[PRT12] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 126–131. ACM, 2012.

[Rac84] S. T. Rachev. The Monge-Kantorovich problem on mass transfer and its applications in stochastics. *Teor. Veroyatnost. i Primenen.*, 29(4):625–653, 1984.

[Vil03] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[Vil08] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[ZVGRG17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.