Obtaining Fairness using Optimal Transport Theory

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

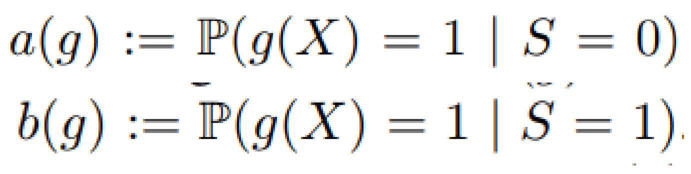
Presenter: Theo Hu

UNIVERSITY OF WATERLOO

Fairness Definition in this paper

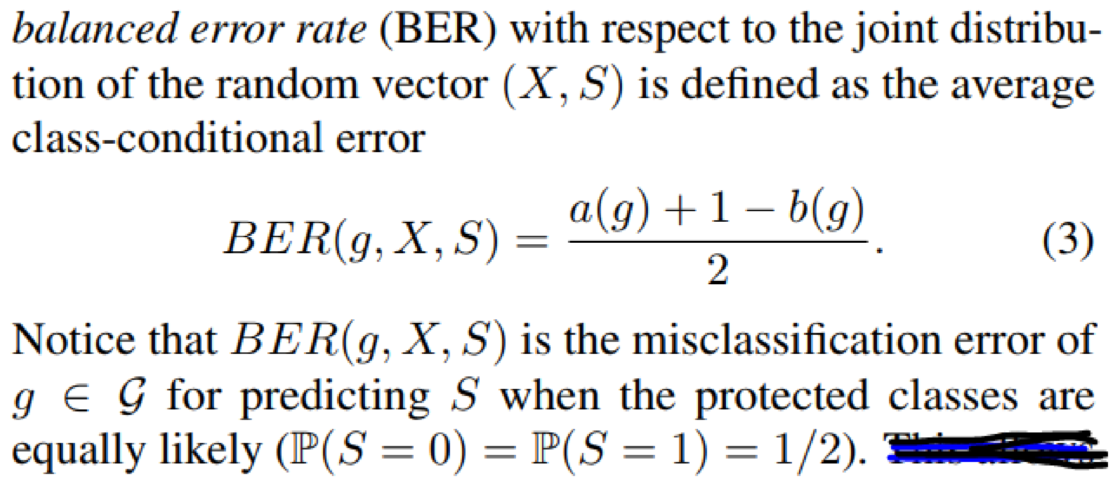
Demographic Parity:

Basically says that your prediction should be independent of sensitive attribute S



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

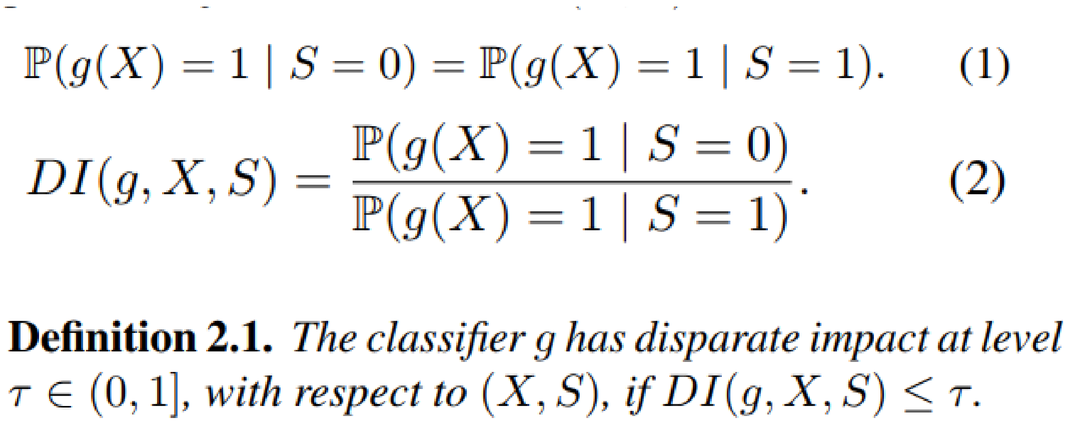
- Disparate Impact(DI)
- Balanced Error Rate(BER)



balanced error rate (BER) with respect to the joint distribution of the random vector (X, S) is defined as the average class-conditional error

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

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 ($\mathbb{P}(S = 0) = \mathbb{P}(S = 1) = 1/2$). [redacted]

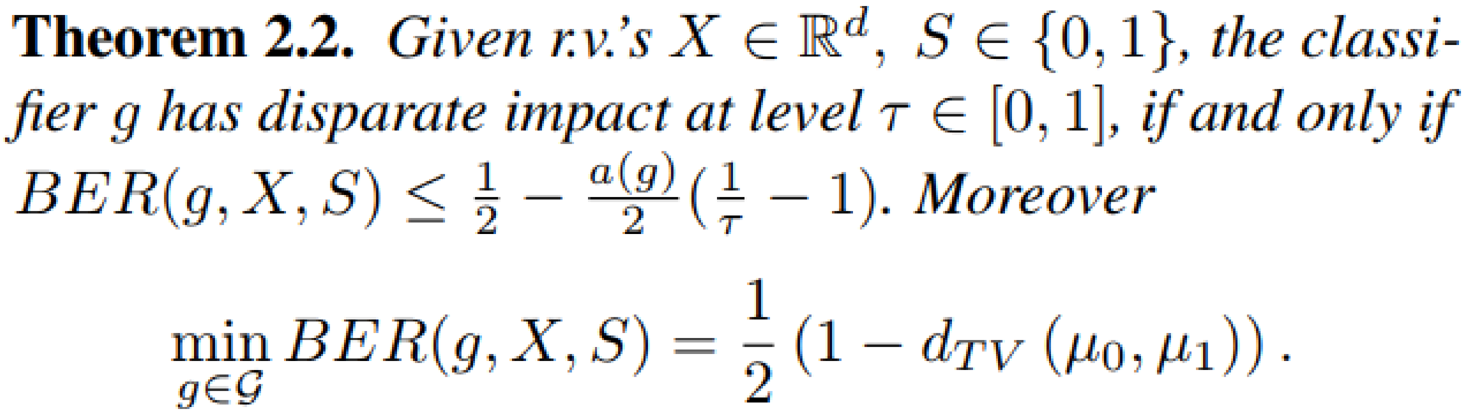


$$\mathbb{P}(g(X) = 1 \mid S = 0) = \mathbb{P}(g(X) = 1 \mid S = 1). \qquad (1)$$

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

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

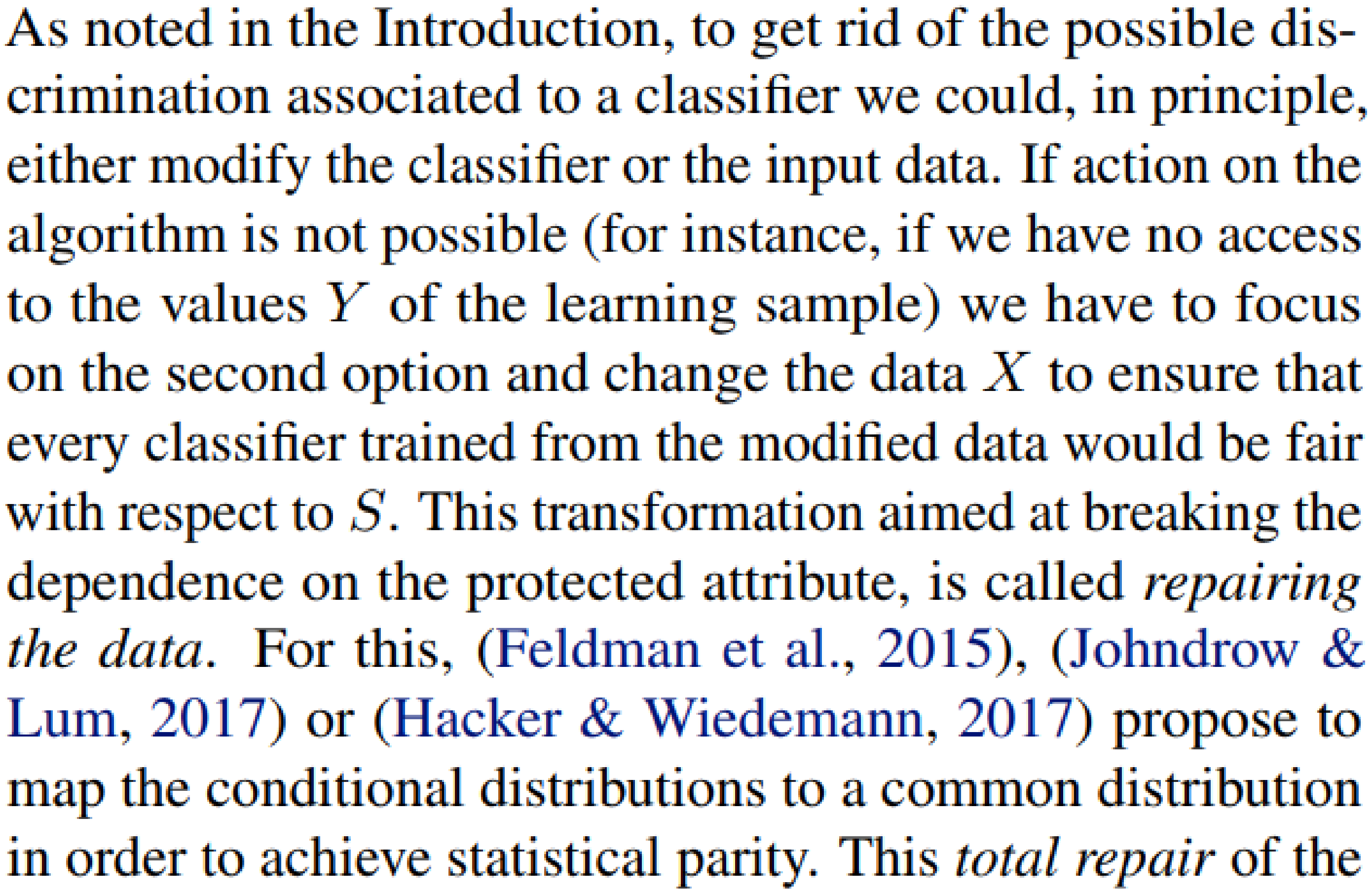
Total Variation Distance is a good way to bound unfairness



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*

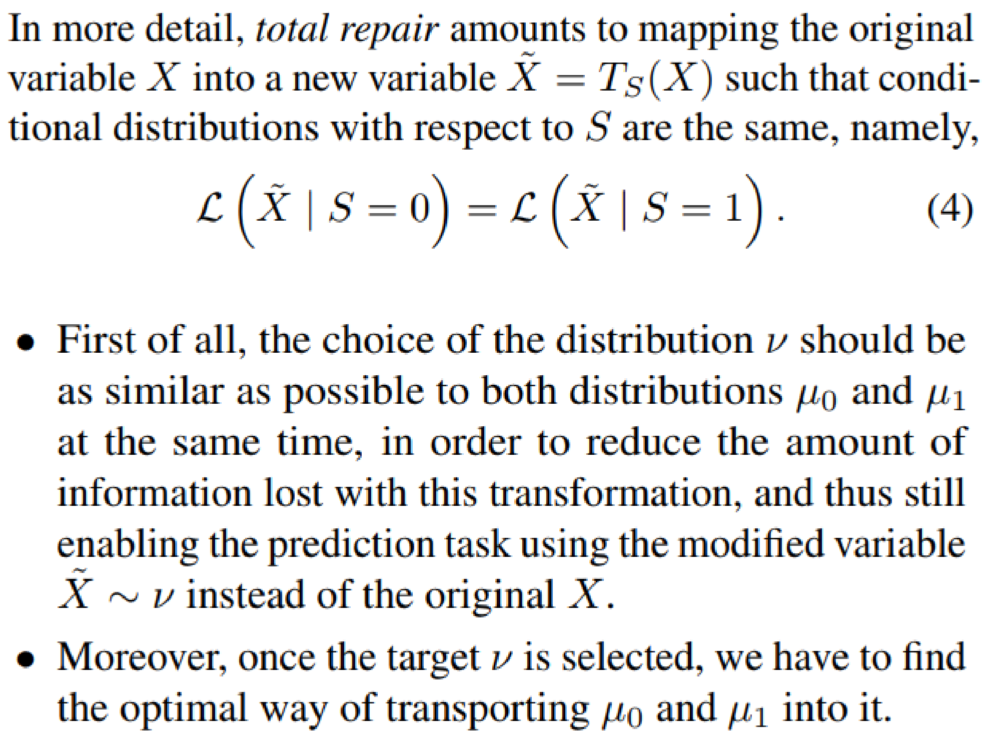
$$\min_{g \in \mathcal{G}} BER(g, X, S) = \frac{1}{2}\left(1 - d_{TV}\left(\mu_0, \mu_1\right)\right).$$

Two general ways of removing ‘unfairness’



As noted in the Introduction, to get rid of the possible discrimination 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

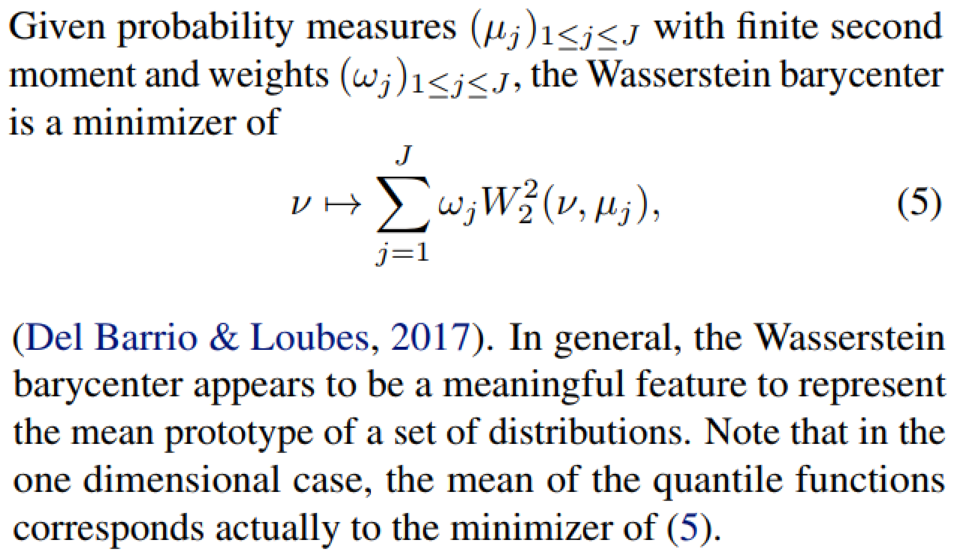
Total Repair and Wasserstein barycenter



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,

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

- 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.

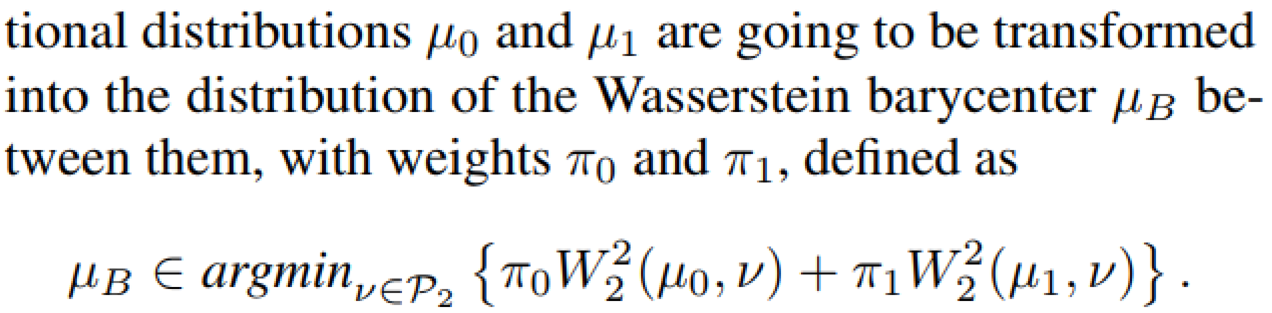


Given probability measures $(\mu_j)_{1 \leq j \leq J}$ with finite second moment and weights $(\omega_j)_{1 \leq j \leq J}$, the Wasserstein barycenter is a minimizer of

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

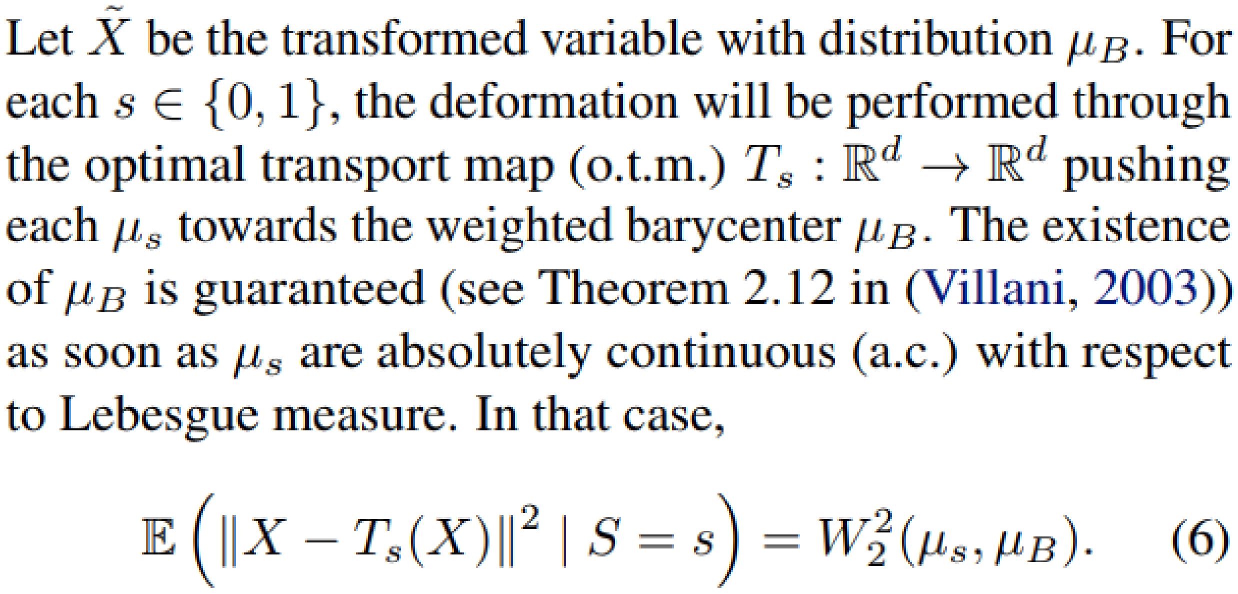
(Del Barrio & Loubes, 2017). In general, the Wasserstein barycenter appears to be a meaningful feature to represent the mean prototype of a set of distributions. Note that in the one dimensional case, the mean of the quantile functions corresponds actually to the minimizer of (5).

Total Repair and Wasserstein barycenter



tional 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

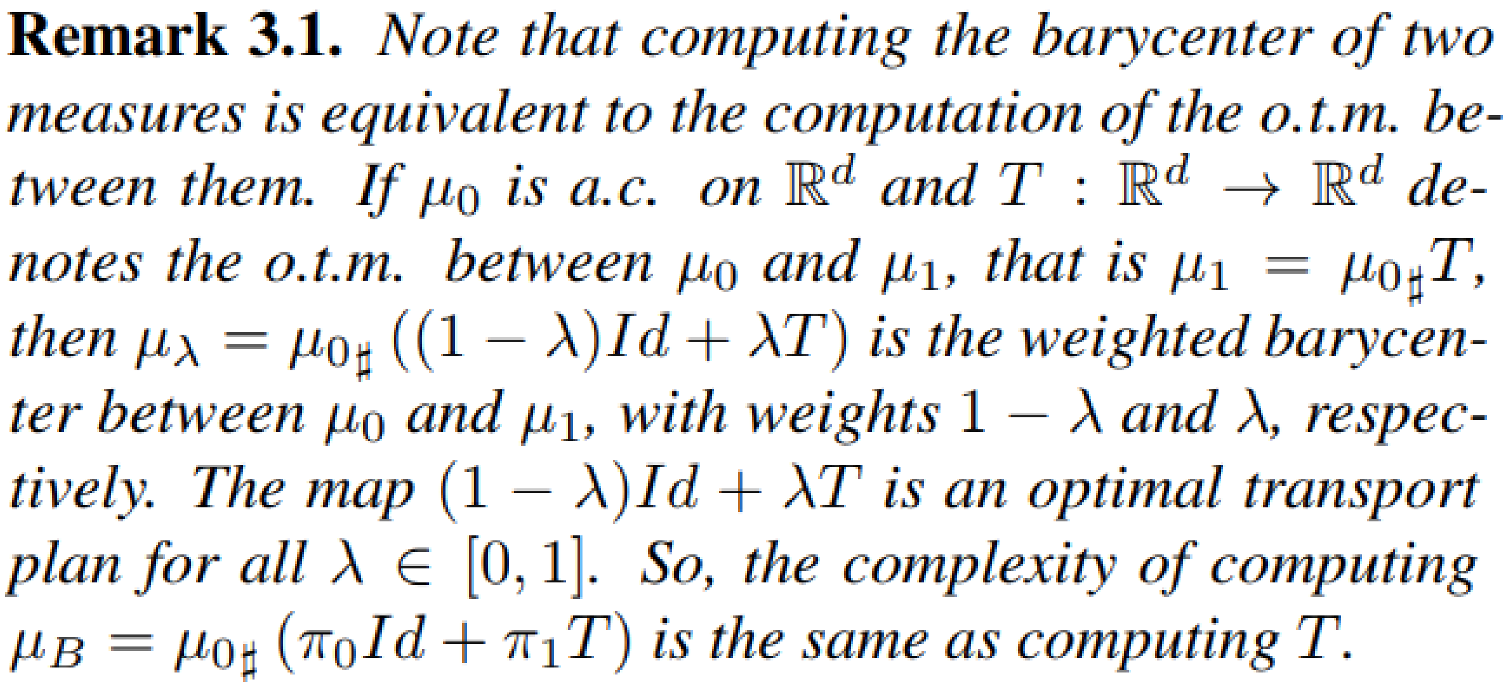
$$\mu_B \in argmin_{\nu \in \mathcal{P}_2} \left\{ \pi_0 W_2^2(\mu_0, \nu) + \pi_1 W_2^2(\mu_1, \nu) \right\} .$$



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,

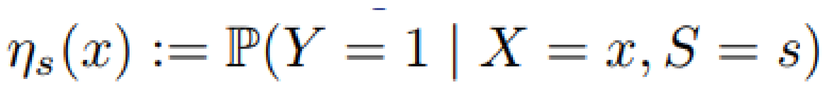
$$\mathbb{E}\left(\|X - T_s(X)\|^2 \mid S = s \right) = W_2^2(\mu_s, \mu_B). \qquad (6)$$

Computing Barycenter

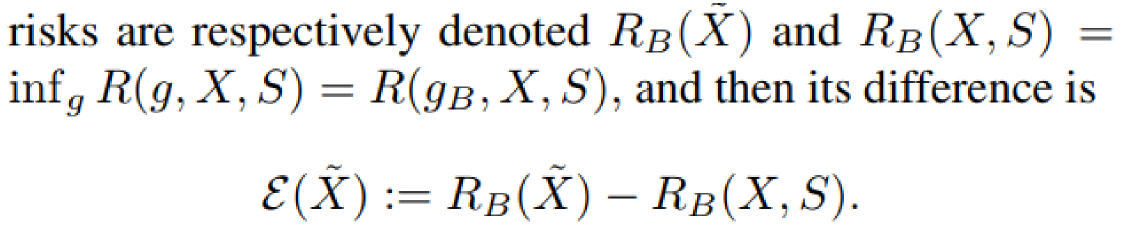


Remark 3.1. *Note that computing the barycenter of two measures is equivalent to the computation of the o.t.m. between them. If μ_0 is a.c. on $\mathbb{R}^d$ and $T : \mathbb{R}^d \to \mathbb{R}^d$ denotes the o.t.m. between μ_0 and μ_1, that is $\mu_1 = \mu_{0\sharp}T$, then $\mu_\lambda = \mu_{0\sharp}\left((1-\lambda)Id + \lambda T\right)$ 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 $\mu_B = \mu_{0\sharp}\left(\pi_0 Id + \pi_1 T\right)$ is the same as computing T.*

Bound of Utility Lost

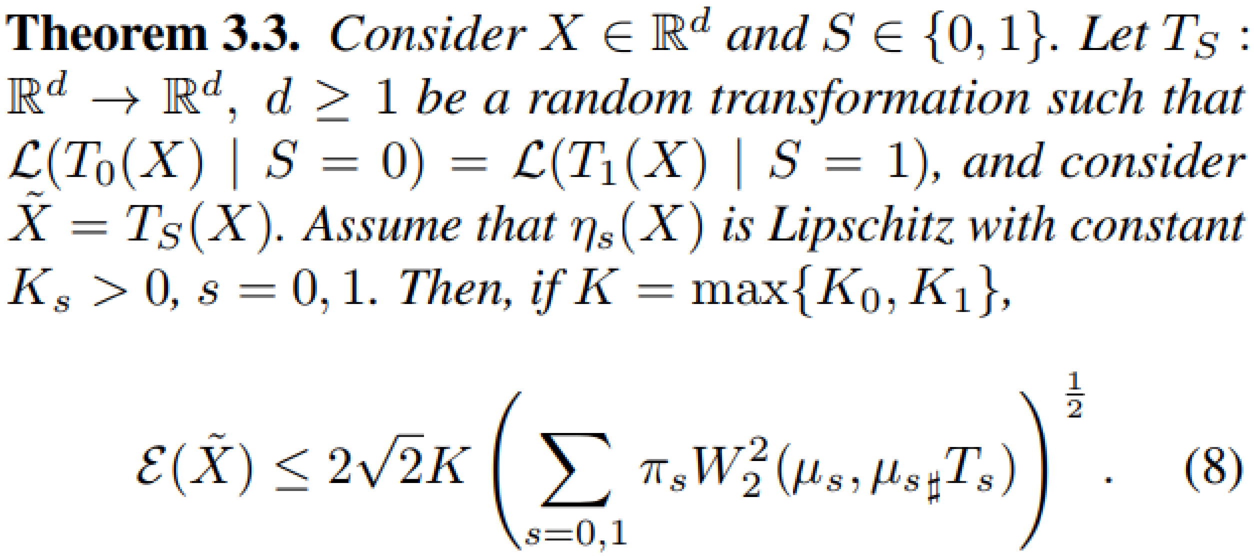


$$\eta_s(x) := \mathbb{P}(Y = 1 \mid X = x, S = s)$$



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

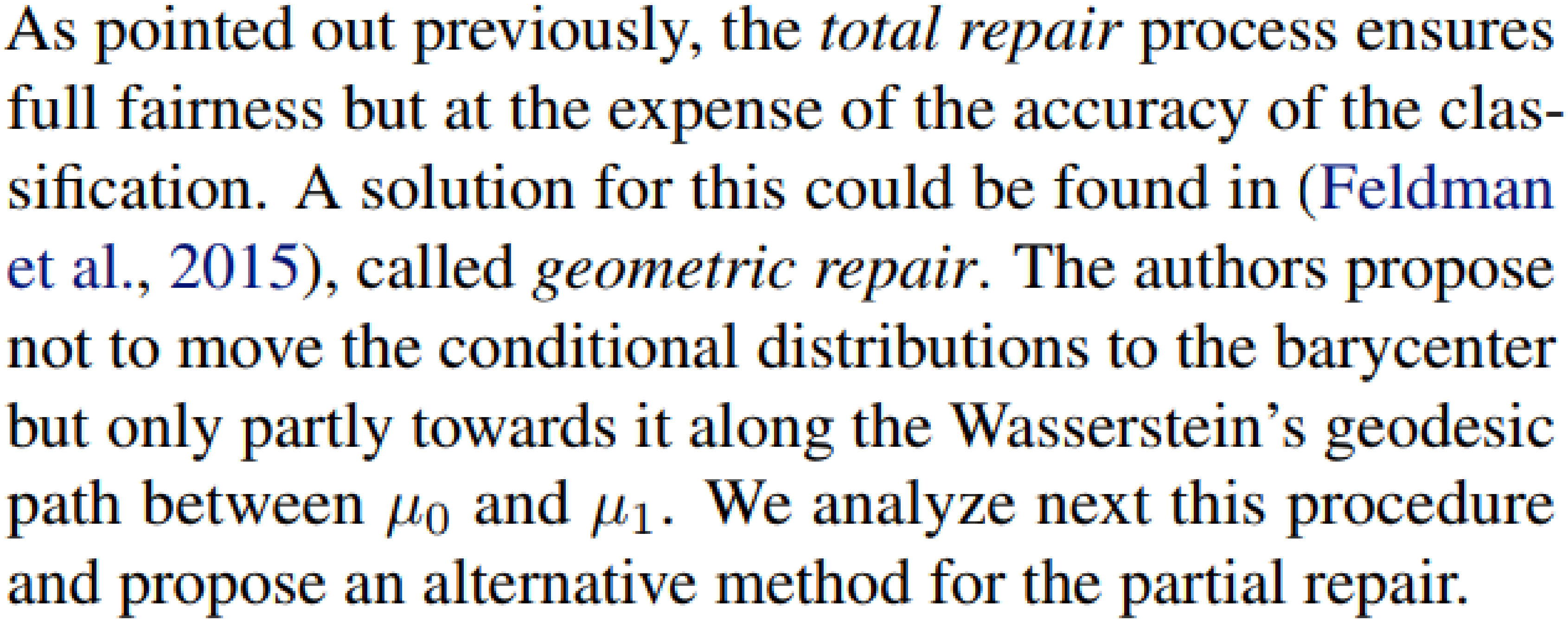
$$\mathcal{E}(\tilde{X}) := R_B(\tilde{X}) - R_B(X, S).$$



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 $K = \max\{K_0, K_1\}$,*

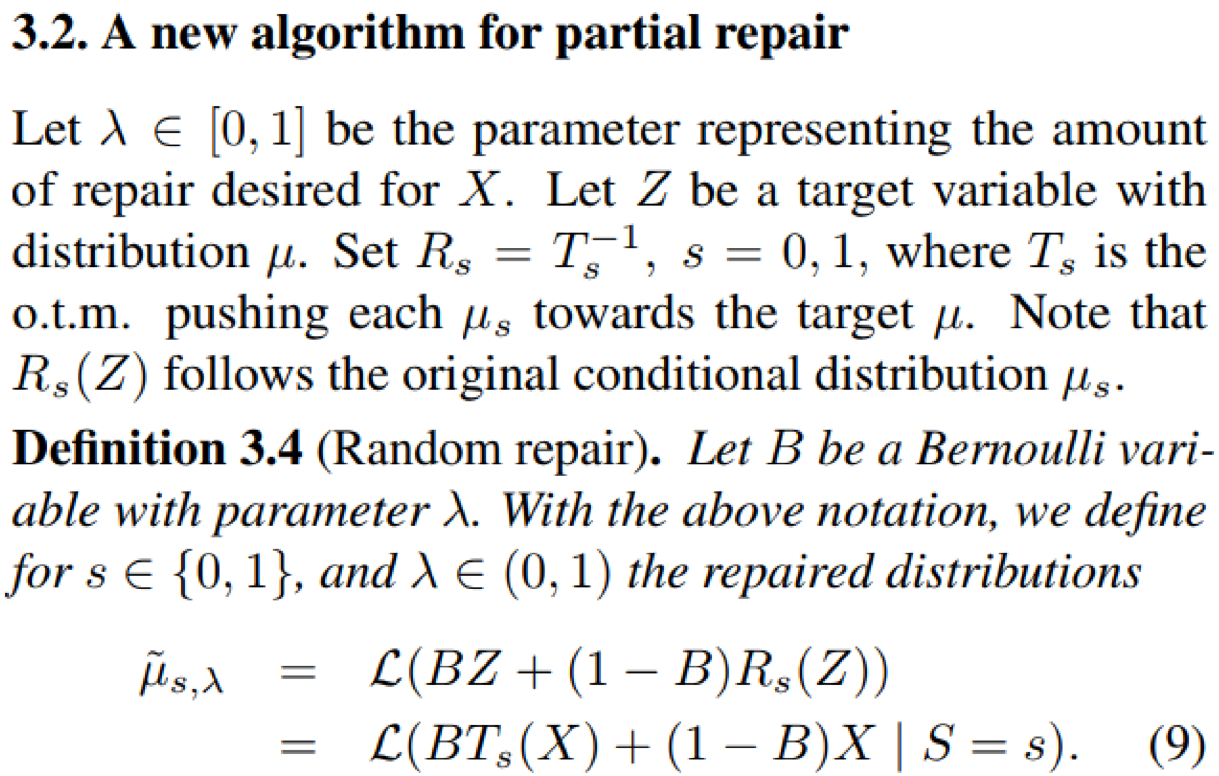
$$\mathcal{E}(\tilde{X}) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s W_2^2(\mu_s, \mu_{s\sharp}T_s) \right)^{\frac{1}{2}}. \quad (8)$$

Total vs. Partial Repair



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.

Random Repair and its guarantee

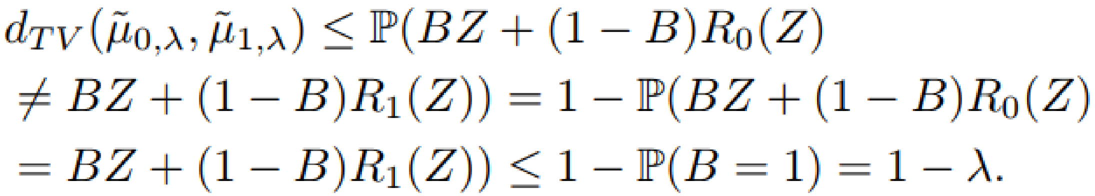


3.2. A new algorithm for partial repair

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.

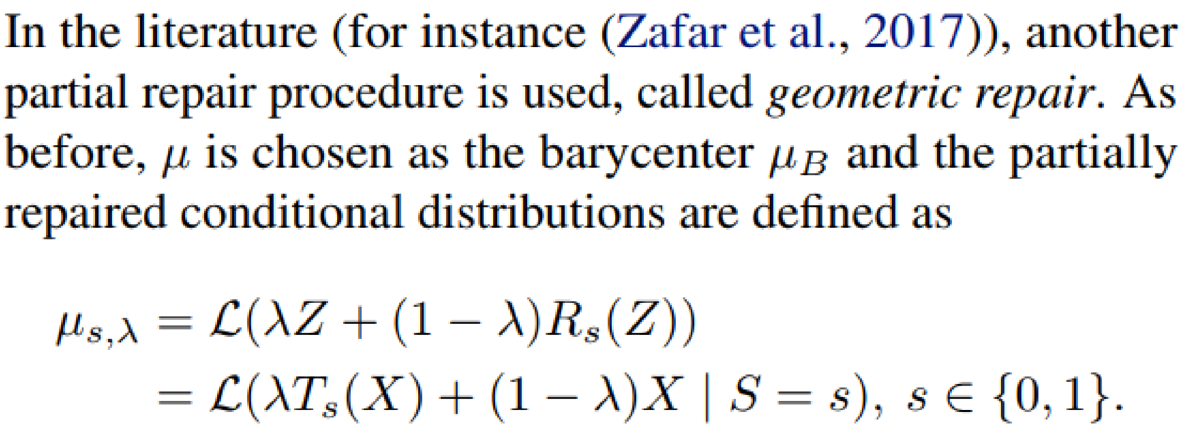
Definition 3.4 (Random repair)**.** *Let B be a Bernoulli variable with parameter λ. With the above notation, we define for $s \in \{0,1\}$, and $\lambda \in (0,1)$ the repaired distributions*

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



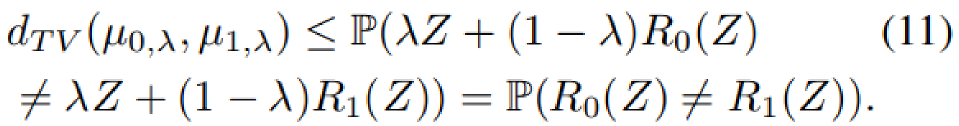
$$\begin{aligned}&d_{TV}(\tilde{\mu}_{0,\lambda}, \tilde{\mu}_{1,\lambda}) \leq \mathbb{P}(BZ + (1-B)R_0(Z) \\ &\neq BZ + (1-B)R_1(Z)) = 1 - \mathbb{P}(BZ + (1-B)R_0(Z) \\ &= BZ + (1-B)R_1(Z)) \leq 1 - \mathbb{P}(B = 1) = 1 - \lambda.\end{aligned}$$

Comparison to Geometric Repair (Previous work)

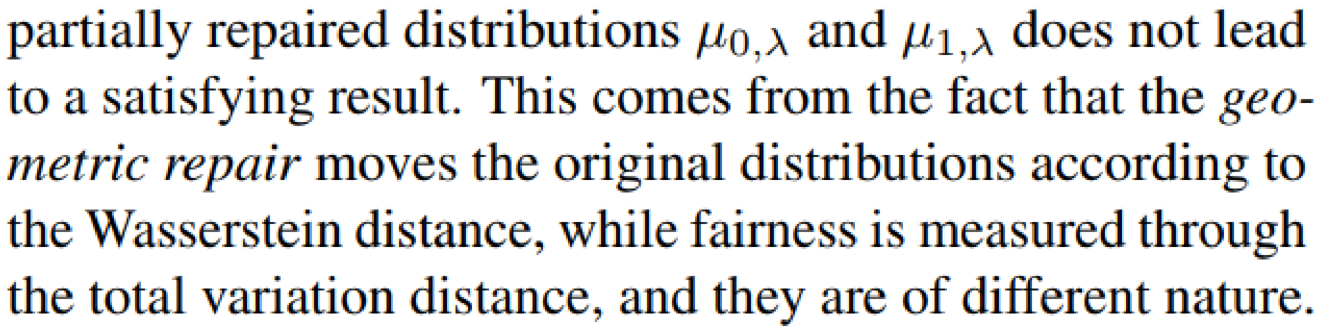


In the literature (for instance (Zafar et al., 2017)), another partial repair procedure is used, called *geometric repair*. As before, μ is chosen as the barycenter μ_B and the partially repaired conditional distributions are defined as

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

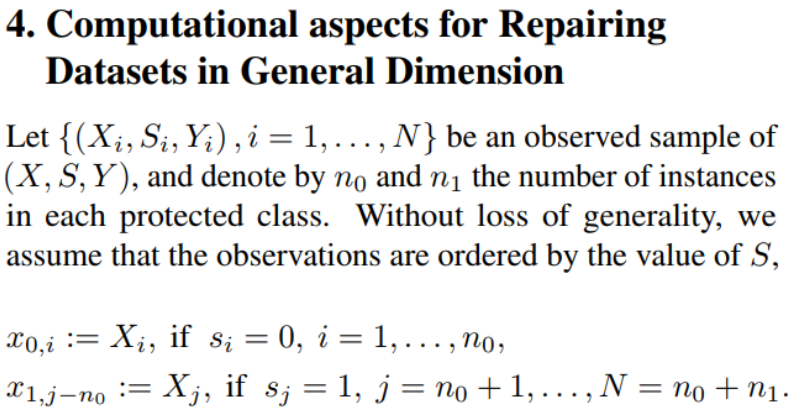


$$\begin{aligned}d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) \leq \mathbb{P}(\lambda Z + (1-\lambda)R_0(Z) \\ \neq \lambda Z + (1-\lambda)R_1(Z)) = \mathbb{P}(R_0(Z) \neq R_1(Z)).\end{aligned} \tag{11}$$



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.

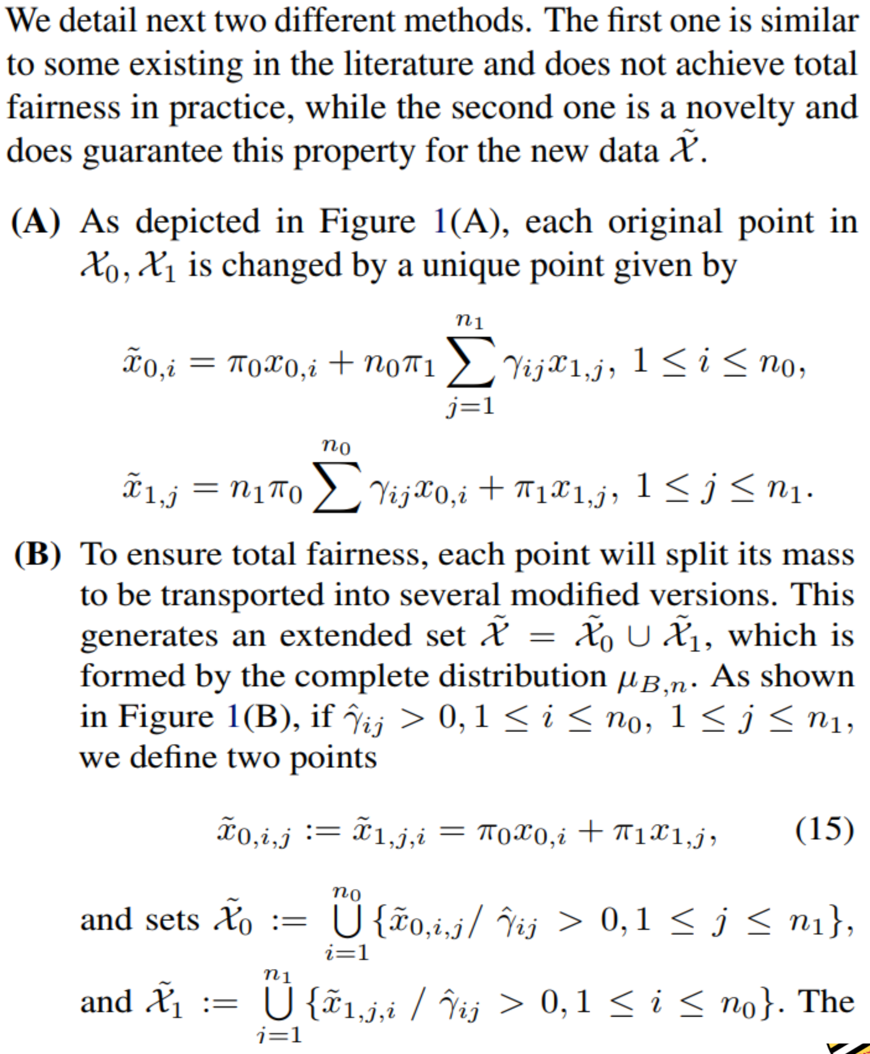
Computational Aspects



4. Computational aspects for Repairing Datasets in General Dimension

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. Without loss of generality, we assume that the observations are ordered by the value of S,

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



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

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

(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

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

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

Computational Aspects- Total Repair

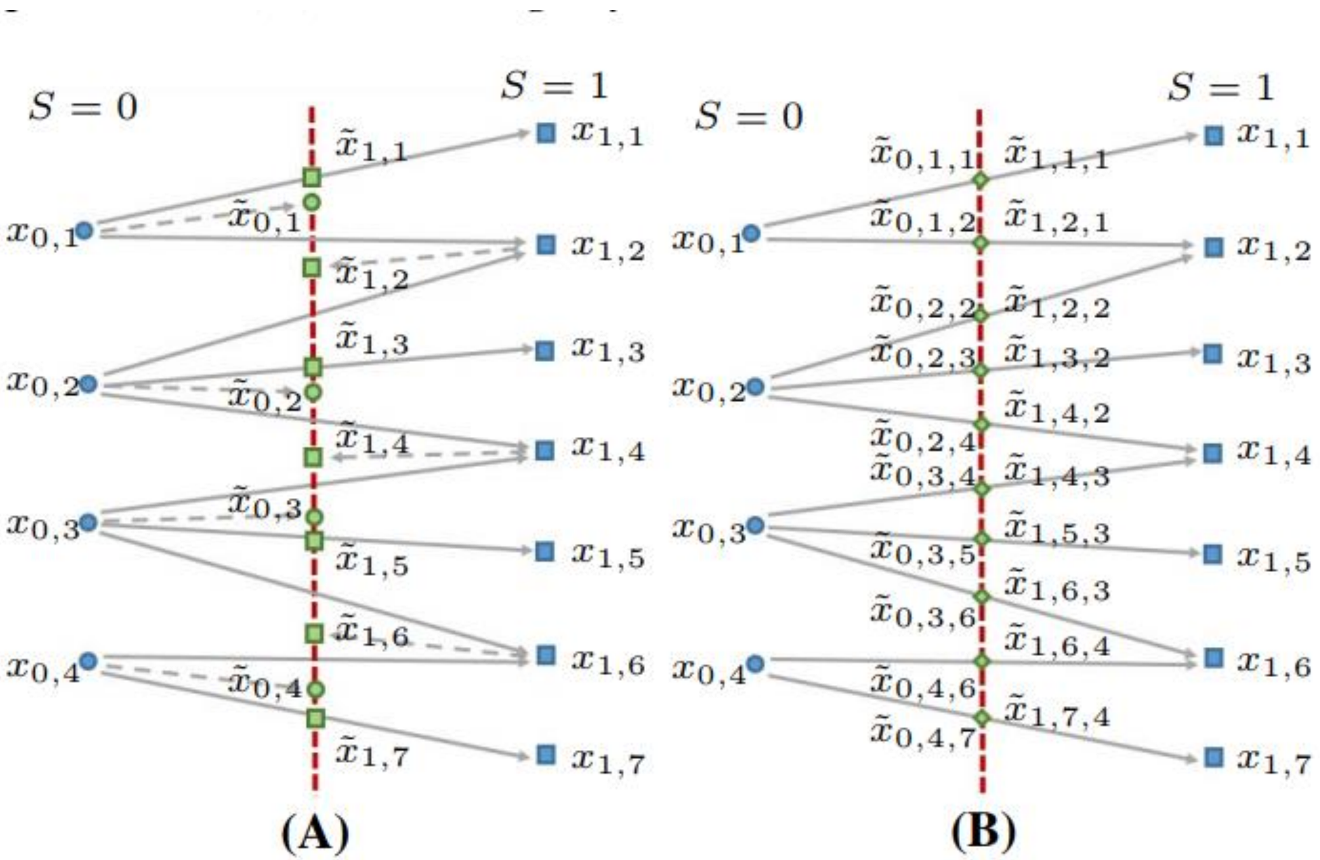


Figure 1. Example of the performance of procedures (A) and (B)

Computational Aspects- Random Repair

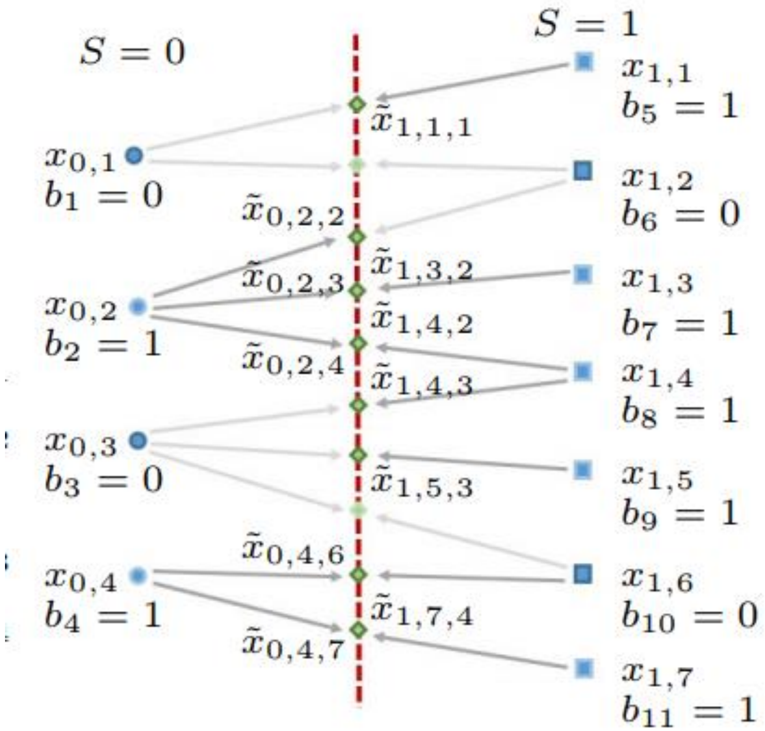
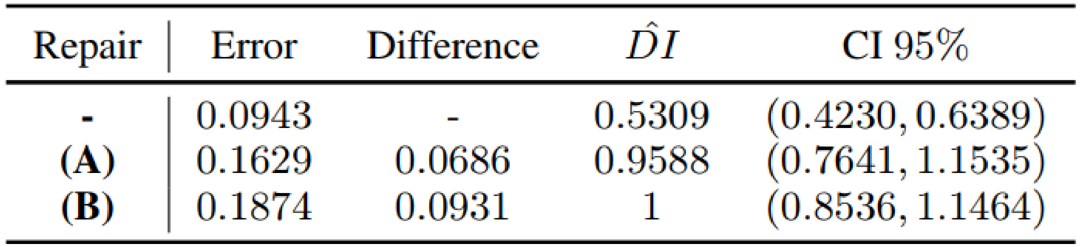


Figure 3. Example of the *random repair* with $\lambda = \frac{1}{2}$.

Experiments

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	$(0.8536, 1.1464)$

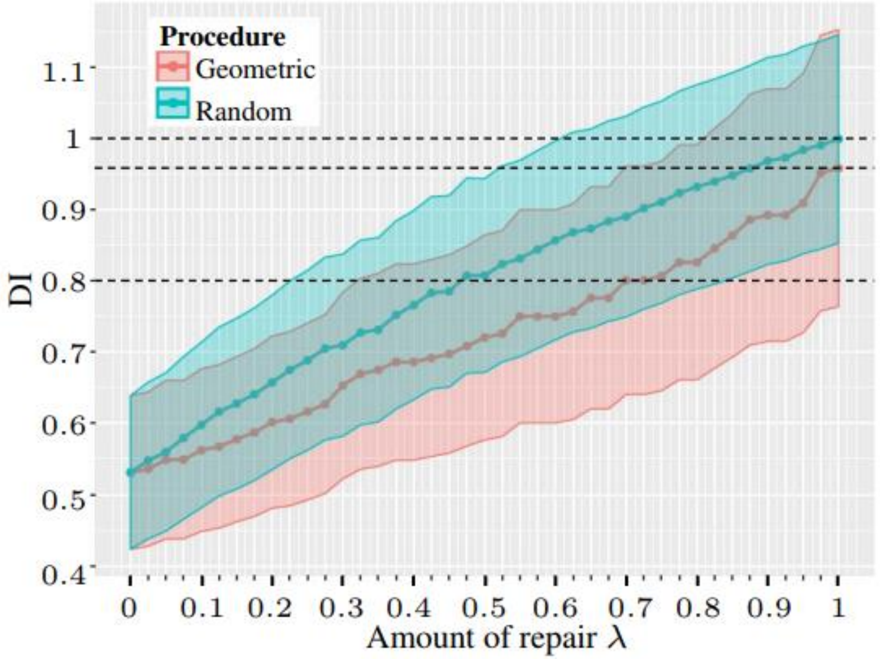
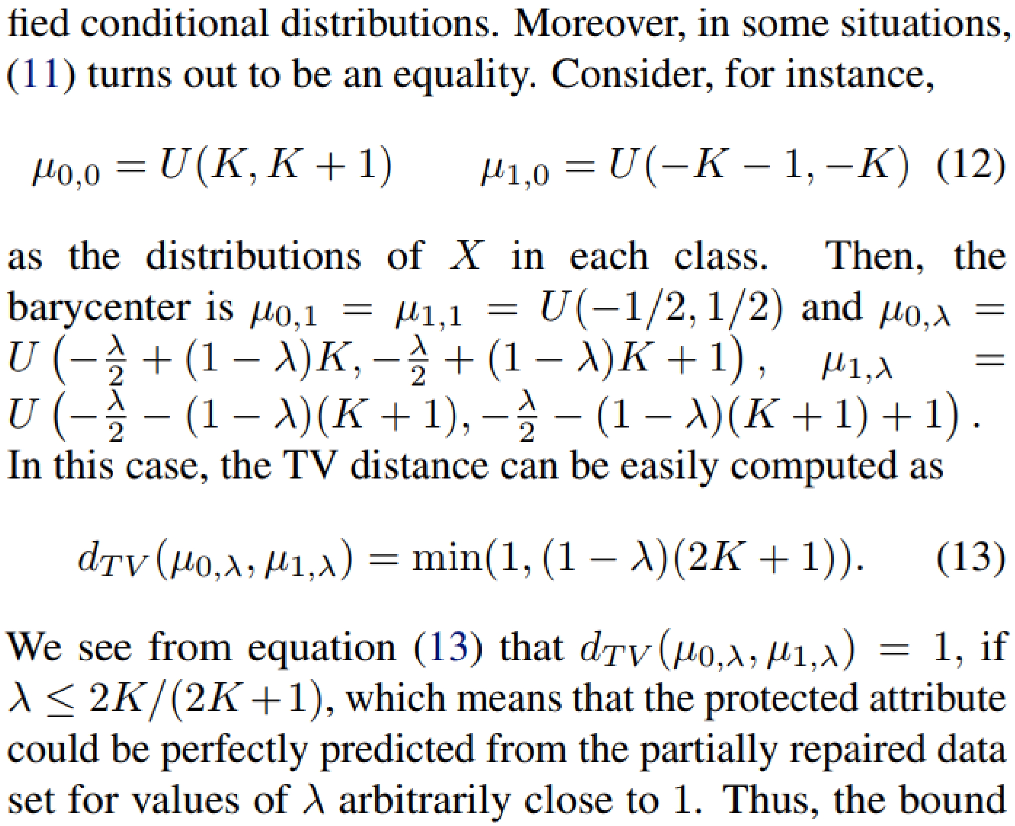


Figure 4. CI at level 95% for DI of the logit

Case that Random Repair beats Geometric Repair



fied 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

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

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

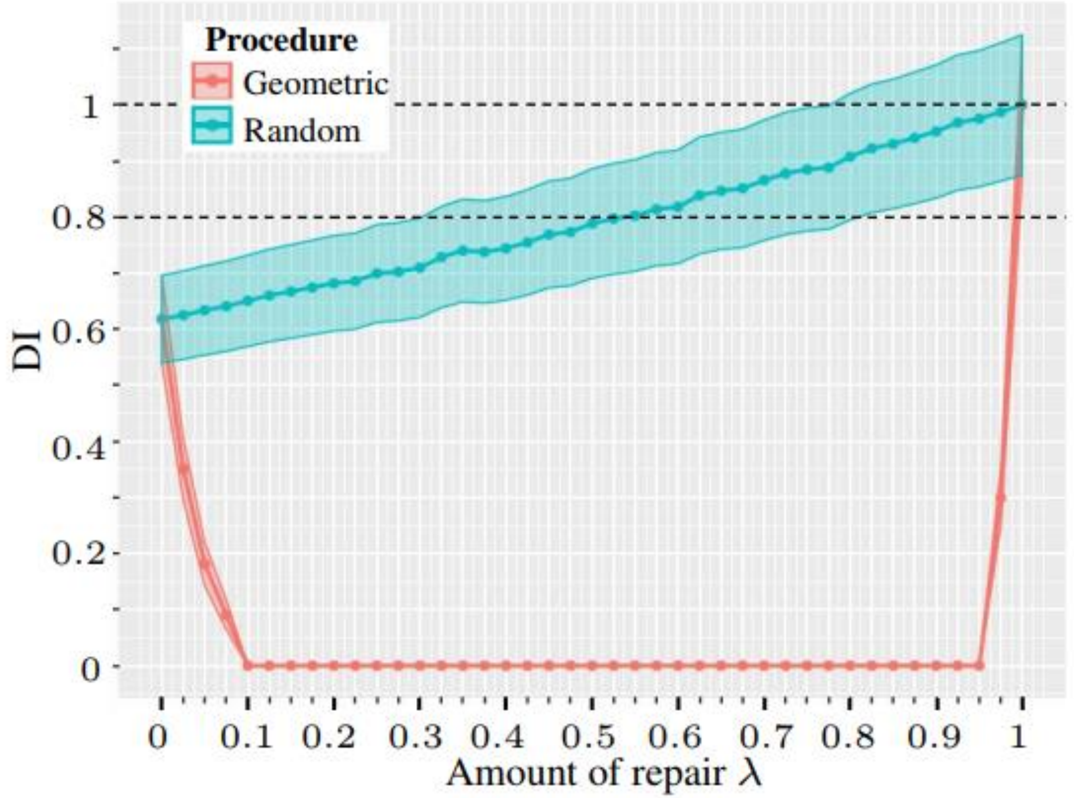


Figure 6. CI at level 95% for DI of the random forest classifier