Augustin COMBES
Thibaut VALOUR

Fairness & Privacy
OBTAINING FAIRNESS USING OPTIMAL TRANSPORT THEORY

02/23/23

1 Introduction

1.1 Objective and motivations

The ever-growing use of predictive methods raises the following *Fairness* question : how to prevent a predictive method tuned from a real-world dataset from replicating unfair societal biases, while this very dataset from which it is trained proxies unfair biases in the first place ?

An elegant way of solving this problem is distorting the source (unfair) dataset in order to protect some of its variables we consider are source of unfair treatment. The naive idea of simply removing these sensible variables from the dataset is unsatisfactory, since the model is liable to reconstruct them from the proxy of the remaining variables and use it for prediction. Then, the paper [1] proposes to distort the source dataset towards one from which it is impossible to predict the sensitive variables, regardless of the predictor applied downstream.

On the other hand, we would like the classification model's performances to be as little compromised as possible. That is, we would actually want to be able to quantitatively constraint the model by the level of fairness we want to achieve.

After introducing the context and tools necessary to understand the article, we will explain each repair method introduced in the paper. We will first tackle Total repair, where all the input dataset is distorted, then two methods of partial repair : the geometric repair, in which we mix the fairly-distorted dataset and the original one using a weight λ, and finally the random repair, in which this fixed weight is replaced with a Bernoulli distribution of parameter λ.

1.2 Context of the study

We first define the formal framework in which we will be discussing.

Given a class $\mathcal{G}$ of classifiers, we perform a prediction using $g : \mathbb{R}^d \to \{0, 1\} \in \mathcal{G}$ such that $\hat{Y} = g(X)$.
This dataset can be further divided into two subpopulations by the binary random variable $S \in \{0, 1\}$.

Interpreting $Y = 1$ as *success* and $Y = 0$ as *failure*, interpreting $S = 1$ as *majority* and $S = 0$ as *oppressed minority*, it becomes obvious that when the random variable S is important for the prediction of Y, the prediction $\hat{Y}$ actually approaches S, that is, we predict *success* for the *majority* sub-sample and *failure* for the *oppressed minority* one.

When one intends to achieve fairness, it means avoiding approaching the previous case. To this end, the paper defines parity for a predictor $g \in \mathcal{G}$ as follows :

$$\mathbb{P}(g(X) = 1|S = 0) = \mathbb{P}(g(X) = 1|S = 1)$$

This is what has been called the *Equal Opportunity* definition in the course : we want both group to have the same probability of success.

1.3 Quantitative tools

The paper uses the Optimal Transport Theory as the framework to map the biased, unfair dataset X into a distorted, fair one, $\hat{X} = T_S(X)$, where T_S is an optimal transport map, that transports $\mathcal{L}(X|S)$ to $\mathcal{L}(\hat{X})$. In practice, when working towards the objective of minimizing the unfairness of the model, we will then actually be tuning T_S. We also want $\mathcal{L}(\hat{X})$, the fair-dataset underlying law, to remain close to $\mathcal{L}(X)$, for the predictive performance to remain suitable.

This calls both for a quantified definition of the fairness of a model, and a quantified definition of a distance between two laws.

1.3.1 Measure of a model's fairness

With the same notations as above, we define two ways of computing fairness.

- The model's **Disparate Impact**, motivated by the *Equal Opportunity* definition :

$$DI(g, X, S) = \frac{\mathbb{P}(g(X) = 1|S = 0)}{\mathbb{P}(g(X) = 1|S = 1)}$$

In our context, this quantity is well-defined and takes values in $]0, 1]$.
This definition induces a way of tuning our model's level of fairness. Indeed, taking $\tau \in]0, 1]$, we can argue that our model is τ-fair when $\forall g \in \mathcal{G}, DI(g, X, S) \leq \tau$. In previous literature, a (fixed) acceptable value of fairness is $\tau = \frac{4}{5}$.

- The model's **Balanced Error Rate**, average prediction error conditionally to S :

$$BER(g, X, S) = \frac{\mathbb{P}(g(X) = 1|S = 0) + 1 - \mathbb{P}(g(X) = 1|S = 1)}{2}$$

The relevance of this measure is to quantify a risk of predictability of our sensible variable S from the fairly-distorted dataset $\hat{X}$ as follows :
Let $\epsilon > 0$, S is ϵ-predictable from $\hat{X} \iff \epsilon < \min_{g \in \mathcal{G}} BER(g, \hat{X}, S)$
Furthermore, introducing the total variation distance, which represents the largest difference that the two probability distributions can assign to the same event :

$$d_{TV}(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

for P and Q two probability measures defined on $(\Omega, \mathcal{F})$, we have the following result :

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

Combining the two latter results, as the ideal is for S not to be predictable from $\hat{X}$, we shall minimize the total variation between $\mathcal{L}(\hat{X}|S = 0)$ and $\mathcal{L}(\hat{X}|S = 1)$.
Moreover, it also motives the use of the total variation distance d_{TV} to quantify the distance between the conditional distributions in our problem.

1.3.2 Wasserstein's statistical distance and barycenter

Using the Wasserstein distance W_2, widely employed in optimal transport, we can define, given μ_0 and μ_1 two probability measures defined on the same spaces and π_0 and π_1 two scalar weights, the **Wasserstein barycenter** as the following probability measure :

$$\mu_B \in \arg\min_{\nu}\{\pi_0 W_2^2(\mu_0, \nu) + \pi_1 W_2^2(\mu_1, \nu)\}$$

According to a previous work [5], μ_B exist as soon as $\forall s = 0, 1, \mu_s \ll \lambda$. We then have the following result that links an S-conditional $\mathcal{L}^2$ norm to this barycenter:

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

This measure is well known in the literature for being a suitable representative for a weighted mean of distributions. One of the goal of the paper is to justify formally the use of this particular measure for this purpose in the context of fairness.

2 Total repair

This section describes how to perform a *total repair*, that is, how to distort the source dataset X into a fair dataset $\hat{X}$ so that any classifier g tuned from the latter dataset will be fair (in the sense of Equal Opportunity). To fulfill this distortion, the *total repair* transports each of the two conditional distributions $\mathcal{L}(X|S = 0)$ and $\mathcal{L}(X|S = 1)$ into an unique distribution, $\hat{X} = T_S(X)$, that then satisfies $\mathcal{L}(\hat{X}|S = 0) = \mathcal{L}(\hat{X}|S = 1)$, hence satisfying $\mathcal{L}(g(\hat{X})|S = 0) = \mathcal{L}(g(\hat{X})|S = 1)$ for any classifier g, hence satisfying the Equal Opportunity condition.

To perform this transport, we would like to pick a common distribution $T_S(X)$ as close as the two conditional ones as possible (to maintain as much predictive information as possible), and then actually optimally transport the conditional distributions to $T_S(X)$.
As mentioned earlier, we pick the common distribution to be the Wasserstein barycenter μ_B, and for $s \in \{0,1\}$, we will be transporting $\mu_s = \mathcal{L}(X|S=s)$ to μ_B using the optimal transport map T_s.

Actually, when transporting the measure α_0 to $\alpha_1 = \alpha_0 T$: let $\lambda \in [0,1]$ be a any scalar weight, then the λ-weighted barycenter of the measures, $\lambda \mapsto \alpha_\lambda = (1{-}\lambda)\alpha_0 + \lambda\alpha_1 = \alpha_0((1-\lambda)I + \lambda T)$, is also an optimal transport map$^{[why?]}$.
Thus, computing the optimal transport map T between μ_0, μ_1 is equivalent to computing the Wasserstein barycenter of them, which further motivates the use of this barycenter as the common distribution to which we transport each conditional one.

Quantifying the information loss when replacing X by $T_S(X)$ requires, for a given classifier c, to define two complementary risks :
- The first one, the *vanilla risk*, which denotes its risk in the context of all the source dataset available (X, S) :

$$R_1(c, X, S) = \mathbb{P}(c(X, S) \neq Y)$$

- The second one, the *fairness risk*, which denotes its risk in the context of a distorted dataset $\hat{X}$:

$$R_2(h, \hat{X}) = \mathbb{P}(c(\hat{X}) \neq Y)$$

Denoting $R_1^*(X,S)$ and $R_2^*(\hat{X})$ the minimum risks obtained for the best classifiers respectively given the source dataset (X,S) and the fairly-distorted one $\hat{X}$, we can thus quantify the information lost during the transition between them by :

$$\epsilon = \epsilon(X,S) = R_2^*(T_S(X)) - R_1^*(X,S)$$

Denoting $\eta_s(x) = \mathbb{P}(Y=1|X=x, S=s)$, let us give an expression for both of these risks :
- let c be a classifier, then

$$\mathbb{P}(c(X,S) \neq Y | X = x, S = s) = \mathbb{1}_{c(x,s)\neq 0}(1 - \eta_s(x)) + \mathbb{1}_{c(x,s)\neq 1}\eta_s(x)$$

We have :

$$\begin{aligned} R_1(c,X,S) &= \mathbb{P}(c(X,S) \neq Y) \\ &= \mathbb{P}((1 - \mathbb{1}_{c(X,S)=0})c(X,S) + \mathbb{1}_{c(X,S)=0}c(X,S) \neq Y) \\ &= \mathbb{E}[\mathbb{E}[\mathbb{P}(c(X,S) \neq Y)|X,S]] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{P}(c(X,S) \neq Y|X,S)]] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{c(X,S)\neq 0}(1 - \eta_S(X)) + \mathbb{1}_{c(X,S)\neq 1}\eta_S(X)|X,S]] \\ &= \mathbb{E}[\mathbb{E}[1 - \eta_S(X) + \mathbb{1}_{c(X,S)=0}(2\eta_S(X) - 1)|X,S]] \\ &= \mathbb{E}[\mathbb{1}_{c(X,S)=0}(2\eta_S(X) - 1] + \mathbb{E}[1 - \eta_S(X)] \end{aligned} \tag{3}$$

We then finally minimize the risk, using Bayes' rule, e.g., choosing $g^*(x,s) = \mathbb{1}_{\eta_s(x)>1/2}$:

$$R_1^*(X,S) = \min_c R_1(c,X,S) = \mathbb{E}[\mathbb{1}_{2\eta_S(X)-1<0}(2\eta_S(X) - 1)] + \mathbb{E}[1 - \eta_S(X)]$$

Similarly as (3), we can show that, with $\hat{X} = T_S(X)$ and any predictor h:

$$R_2^*(h(\hat{X})) = \mathbb{E}[\mathbb{1}_{h\circ T_S(X)=0}(2\eta_S(X) - 1)] + \mathbb{E}[1 - \eta_S(X)]$$

Finally, for $X \in \mathbb{R}^d$, $S \in \{0,1\}$ and $T_S : \mathbb{R}^d \to \mathbb{R}^d$, under the additional condition$^{[aretheystrong?]}$ that for any $s = 0,1$, $\eta_s(X)$ is K_s-lipschitzian, and denoting $K = max(K_0, K_1)$, we can control our error ϵ using the following bound:

$$\begin{aligned} \epsilon(X,S) &= R_2^*(T_S(X)) - R_1^*(X,S) \\ &\leq 2\sqrt{2}K\left(\sum_{s=0,1}\pi_s W_2^2(\mu_s, \mu_s T_s)\right)^{\frac{1}{2}} \\ &\leq \frac{K}{\sqrt{2}}W_2^2(\mu_0, \mu_1) \end{aligned} \tag{4}$$

The proof of this result is available in the Appendix section of the original paper. We can interpret these bounds as a quantitative justification for the Wasserstein barycenter to be a feasible solution to make a fair classification.

3 Partial Repair: Geometric repair

3.1 Theory

The previous discussion highlighted that the total repair process guarantees fairness, but this comes at the cost of reducing the accuracy of the classification, since we are removing too much information from the initial dataset. A potential solution is found in the geometric repair method proposed by Zafar in [4]. This approach does not move the conditional distributions all the way to the barycentre, but rather only partially along the Wasserstein's path between the two distributions. In this section, we will present this method and point out its limits.
Let $R_s = T_s^{-1}$, and Z be a target variable with the distribution μ the wasserstein barycentre.
For any $\lambda \in [0,1]$:

$$\mu_{s,\lambda} = \mathcal{L}(\lambda Z + (1-\lambda)R_s(Z)) = \mathcal{L}(\lambda Z + (1-\lambda)R_s(Z))$$

This approach seems appealing since it enables to control the closeness between the two final distributions $\mathcal{L}(T(X)|s=0)$ and $\mathcal{L}(T(X)|s=1)$. Then thanks to this method, intuitively, one could simply increase lambda until one obtains a fair criterion. For instance, the 80 % rule detailed by Feldman in [3], which says that an algorithm's discrimination is lawful if DI(g, X, S) $\leq \tau$ with $\tau = \frac{4}{5}$.

3.2 Limits

3.2.1 Theoretical limits

The article mentions the geometric repair method, almost only to quickly dismiss it. If it is possible to show a practical case where the geometrical repair fails to remove bias from some data, as we will do in the next subsection, even the theoretical concept of this repairing method has lead in the wing.
Fairness is defined with the total variation distance, as we can see in the equation (1), So while the transformation with the wassertein distance might be satisfying for this distance, there's no reason that it would be the case for the total distance.
Massart has shown in [2] that the distance between two probabilities P and Q in total variation can be computed as:

$$d_{TV}(P,Q) = \min_{\pi \in \Pi(P,Q)} \pi(x \neq y) \tag{5}$$

where $\Pi(P,Q)$ denotes the set of all joint distributions with marginals P and Q. From this, by replacing (P, Q) by $(\mu_{0,\lambda}, \mu_{1,\lambda})$, with $\lambda \in (0,1)$, we can derive the following inequality:

$$\begin{aligned} d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) &\leq P(\lambda Z + (1-\lambda)R_0(Z) \neq \lambda Z + (1-\lambda)R_1(Z)) \\ &= 1 - P(\lambda Z + (1-\lambda)R_0(Z) = \lambda Z + (1-\lambda)R_1(Z)) \\ &= 1 - P(R_0(Z) = R_1(Z)) \\ &= P(R_0(Z) \neq R_1(Z)) \end{aligned} \tag{6}$$

The previous bound indicates that the amount of repair quantified by λ has no impact on the total variation distance between the modified conditional distributions. There is therefore no assurance of the outcome of such a distribution on the disparate impact. In fact, it is even possible to find a pathological case in which the result in (6) is an equality, which means that no equity will be obtained by this transformation.

3.2.2 A Pathological example

Let's consider the following pathological case :

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

This case is quite extreme, as it assumes that the two distributions are disjointed. This means that no one in the oppressed minority has characteristics even close to those of a person in the majority. However, even if it is not representative of most situations where ML classifiers are used, it highlights the limitations of geometric repair which could be problematic in future uses.
One can compute the barycentre of the two distributions : $\mu_{0,1} = \mu_{1,1} = \mathrm{U}\left(-\frac{1}{2}, \frac{1}{2}\right)$ (Admitted)

Thus,

$$\mu_{0,\lambda} = U(-\frac{\lambda}{2} + (1-\lambda)K, -\frac{\lambda}{2} + (1-\lambda)K + 1)$$
$$\mu_{1,\lambda} = U(-\frac{\lambda}{2} - (1-\lambda)(K+1), -\frac{\lambda}{2} - (1-\lambda)(K+1) + 1)$$
$$d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) = \sup_{A \in \mathcal{F}} |\mu_{0,\lambda}(A) - \mu_{1,\lambda}(A)|$$

By taking both extremes of uniform distributions and also consider that the difference is smaller than 1 since $\mu_{s,\lambda}$ are probability measures:

$$\begin{aligned} d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) &= min(1, -\frac{\lambda}{2} + (1-\lambda)K + 1) - (-\frac{\lambda}{2} - (1-\lambda)(K+1))) \\ &= min(1, (1-\lambda)(2K+1)) \\ &= 1 \text{ if } \lambda \leq \frac{2K}{2K+1} \end{aligned}$$

This means that for any value $\lambda \in (0,1)$, as close to 1 as we want, there is a practical example in which it's perfectly possible to predict the protected attribute from the repaired data, that will then be used by the classifier.

In this section, we have highlighted important limitations of the geometric repair proposed by Zafar in [4]. The article [1] suggests a second method which is very much inspired by this one, but allows overcoming the presented limitations.

4 Partial Repair: Random repair

4.1 Theory

To define random repair, we use the same definition as partial repair, but with a small change. Instead of using λ as a weighting parameter, which we have shown in previous section does not necessarily remove bias, we will use a random variable B with the following distribution:

$$B \sim \text{Bernoulli}(\lambda)$$

We can then define the new repaired distribution $\mu_{s,\lambda}$ with

$$\mu_{s,\lambda} = \mathcal{L}(BT_s(X) + (1-B)X) = \mathcal{L}(BZ + (1-B)R_s(Z))$$

Using the same demonstration as in equation (6), one can show that

$$\begin{aligned} d_{TV}(\tilde{\mu}0, \lambda, \tilde{\mu}1, \lambda) &\leq P(BZ + (1-B)R_0(Z) \neq BZ + (1-B)R_1(Z)) \\ &= 1 - P(BZ + (1-B)R_0(Z) = BZ + (1-B)R_1(Z)) \\ &\leq 1 - P(B=1) \\ &= 1 - \lambda \end{aligned} \tag{7}$$

This bound suggests that with a λ close to 1 we ensure that S is unpredictable.

4.2 Implementation

For the 5D case discussed in section 5 of the reference article [1], we applied both random and geometric repair methods to various values of λ ranging from 0 to 1. We then evaluated the resulting Disparate Impact with its confidence interval (as shown in figure 2) and the accuracy of a logit classifier using the repaired data (as depicted in figure 3). This allowed us to compare the effectiveness of the two repair methods.
Our code is available on this repository `https://github.com/AugustinCombes/optimal-transport-fairness`

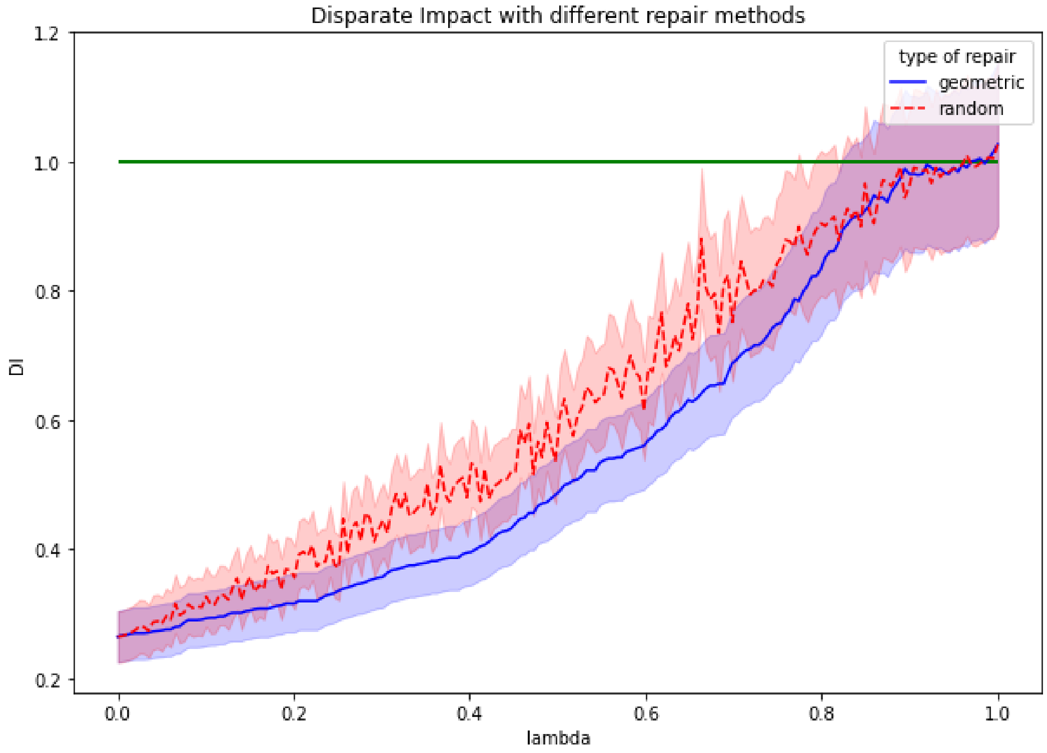

Figure 1: DI by lambda

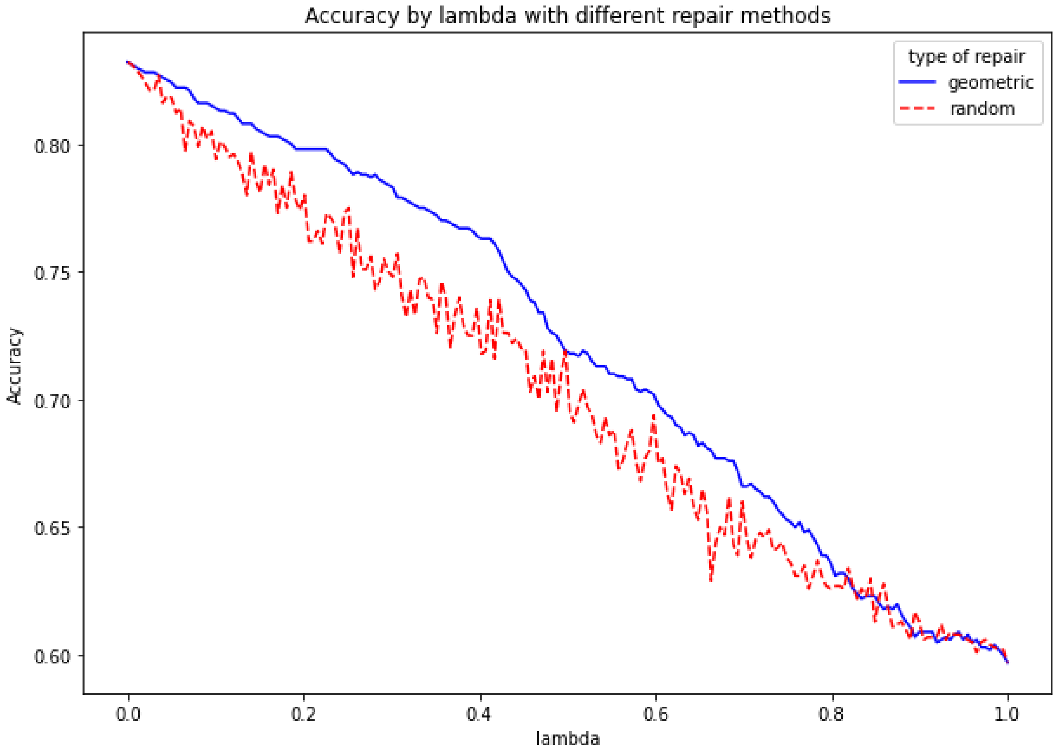

Figure 2: Accuracy by lambda

It is evident that the random repair method outperforms the geometric repair method in terms of disparate impact, regardless of the value of λ that is selected. Additionally, the accuracy of the classifier is generally higher after undergoing a random repair, compared to the geometric repair, for the majority of λ values. Overall, the random repair method appears to be significantly more effective than the geometric repair method.

4.3 Limits

The main limitation of this approach presented in this article is that it only allows for one definition of fairness, equal opportunity. This is a limitation, but all solutions will have this kind of limitation, since we proved during the course with the Impossibility Theorem that it is impossible to create a method that respects all definitions of fairness, except by removing all precision from the classifier.

We discovered another limitation in the random repair approach: When $\lambda \neq 1$, which is the case when we are not doing total repair, then some kind of unfairness between people of the same class S might be introduced by the model. Let's consider the following example:

We are working on a one-variable problem $X \in \mathbb{R}^1$, and X is the number of years of study.
Suppose we have two individuals, $X^1_{s=0}$ and $X^2_{s=0}$, from the minority class, with $X^1_{s=0} = 1$ and $X^2_{s=0} = 3$. Let $T_0(X^1) = 3.5$ and $T_0(X^2) = 5$, which may occur if the majority class has a very high average number of years of study.
If we choose a value of λ in the range $(0, 1)$, it is possible that the sampling function $B(\lambda)$ takes the value 1 for the first person and 0 for the second. This would result in the following equations:

$$\tilde{X}^1_{s=0} = (BT_0(X^1) + (1 - B)X^1) = T_0(X^1) = 3.5$$
$$\tilde{X}^2_{s=0} = (BT_0(X^2) + (1 - B)X^2) = X^2 = 3$$

Thus, the second person is now less likely to be considered by a classifier as a high performer than the first, even if he or she has more years of education. This means that while we may have introduced global fairness by improving the Disparate Impact, we may have also created an unfair treatment among individuals within the same class.l

5 Conclusion

It is important to consider the fairness of machine learning algorithms, as they can perpetuate and amplify existing biases in society. In this article, we have presented three methods for repairing biased datasets in order to make their classification more fair. The first method, total repair, involves transporting the conditional distributions of the sensitive attribute to a common distribution, the Wasserstein barycentre, in order to achieve equal opportunity for all groups. However, this method can result in a significant loss of information and accuracy in the classification.
The second method, geometric repair, involves partially transporting the conditional distributions along the Wasserstein path between them in order to control the amount of repair. While this method may appear appealing at first, it has been shown to have theoretical limitations and may not necessarily remove bias.
The third method, random repair, involves randomly applying either the original data or the total repair transformation with a certain probability, controlled by the parameter λ. This method has been shown to be more effective in reducing bias and maintaining classification accuracy compared to the geometric repair method.
Finally, it should be noted that all of these repair methods are limited to addressing equal opportunity fairness and may not necessarily address other definitions of fairness.

References

[1] Paula Gordaliza Jean-Michel Loubes Eustasio del Barrio, Fabrice Gamboa. Obtaining fairness using optimal transport theory. 2018.

[2] Massart. Concentration inequalities and model selection, volume 6. springer. 2007.

[3] John Moeller Carlos Scheidegger-Suresh Venkatasubramanian Michael Feldman, Sorelle Friedler. Certifying and removing disparate impact. 2015.

[4] Manuel Gomez Rodriguez-Krishna P. Gummadi Muhammad Bilal Zafar, Isabel Valera. Fairness beyond disparate treatment disparate impact: Learning classification without disparate mistreatment. 2017.

[5] Villani. Optimal transport: old and new. 2009.