# Obtaining Fairness using Optimal Transport Theory

## 1 Supplementary material

### 1.1 Proof of Theorems

Proof of Theorem 2.2

*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 $g \in \mathcal{G}$,

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

Moreover, 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 $\mu$. In general, the misclassification error could be written as:

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

Now, for $s = 0, 1$, we fix 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 (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

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

□

**Lemma 1.1.** *Under Assumptions of Theorem 3.3, the following bound holds*

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

Proof of Lemma (1.1)

*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

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

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 $g_B \circ T_s(x) = \mathbb{1}_{\eta_s \circ T_s(x) > \frac{1}{2}}, forall x, forall s$. 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 3.3

*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

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

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

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

□

## 1.2 Application on a real dataset

To ilustrate the performance of the repairing procedures in Section 3, we consider the *Adult Income* data set (available at `https://archive.ics.uci.edu/ml/datasets/adult`). 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 Besse et al. (2018). 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. Train the classifiers based on logistic regression and random forests using the five numerical variables: *Age, Education Level, Capital Gain, Capital Loss* and *Worked hours per week*.

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 to the test sample described by the 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 **??** 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 Feldman et al. (2015), 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. Left part of Figures 1 and 2 represent the evolution of the estimated Disparate Impact with the amount of repair $0 \leq \lambda \leq 1$, while the right part show the evolution with $\lambda$ 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; and 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; and 0.2076 with the Geometric Repair

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

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

| 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 | $(0.978, 1.203)$ |

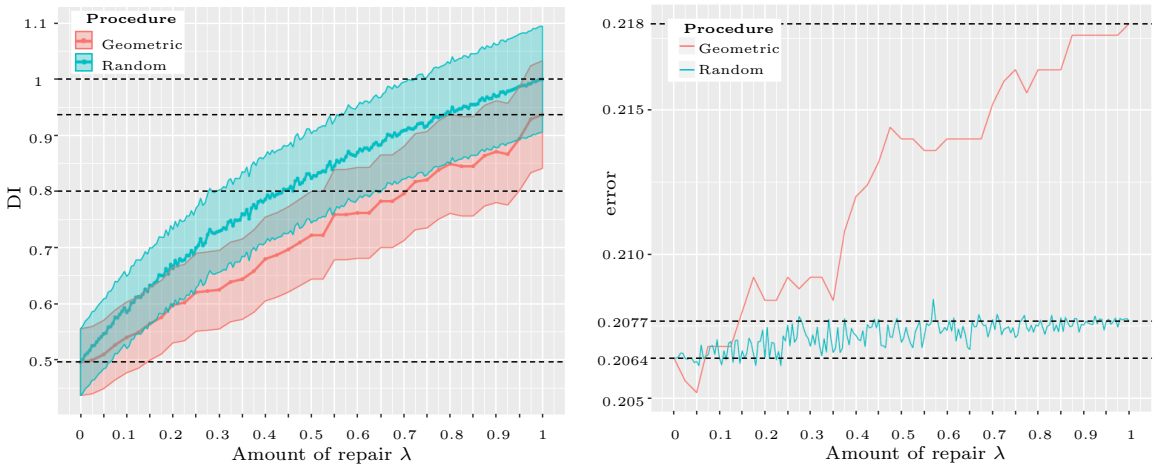


Figure 1: CI at level 95% for DI (left) and error (right) of the classifier logit with respect to Gender and the data repaired by the Geometric and Random Repair

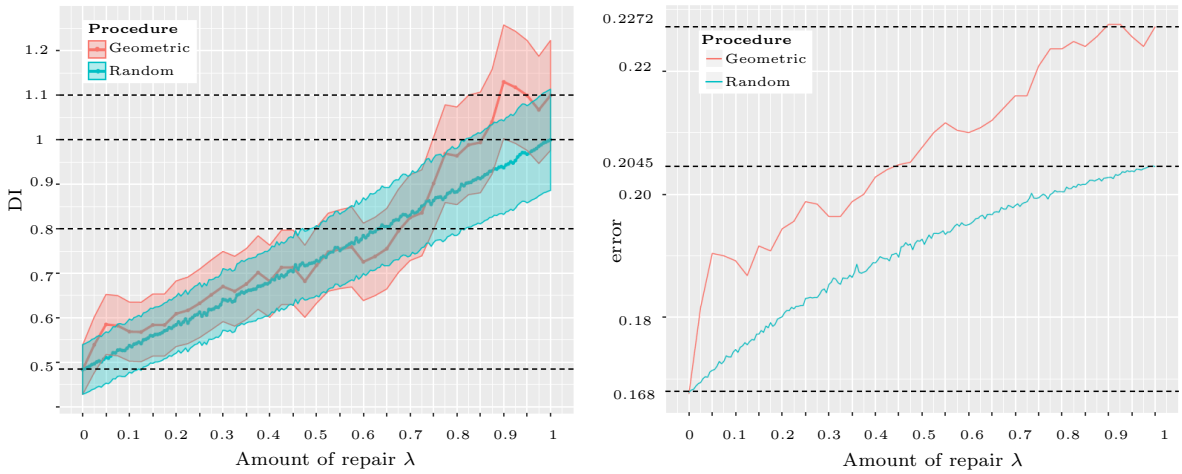


Figure 2: CI at level 95% for DI (left) and error (right) of the classifier random forests with respect to Gender and the data repaired by the Geometric and Random Repair

# References

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

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