Bias in Machine Learning: Challenges for certifiable AI

JM Loubes
Chair Fair & Robust Machine Learning (ANITI)
Forum Mobilit.AI

Institut de Mathématiques de Toulouse
& Artificial and Natural Intelligence Institute of Toulouse

based on :

- **Obtaining Fairness with Optimal Transportation (Proceedings of ICML 2019)**
- **Central limit theorems for empirical transportation cost in general dimension (The Annals of Probability 2019)**

How Machine Learning could go possibly wrong ?

Decisions are taken based on machine learning algorithms (most often black box models)

Used for recommendation systems, insurance, banks, human ressources, education, communication ... but also areas justice, medicine, police, political decisions ....

Learning Sample : $(Y_1, X_1), \ldots, (Y_n, X_n)$ with distribution $\mathbb{P}$ learnt from empirical version $\mathbb{P}_n$

Parameter of interest :

$$f^\star \in \arg\min \mathbb{E}_{\mathbb{P}}\{\ell(Y, f(X)) + \text{penalty}(f)\}$$

Decision Rule

$$\widehat{f}_n \in \arg\min \mathbb{P}_n \tilde{\ell}(Y, f(X)) = \arg\min \frac{1}{n}\sum_{i=1}^{n}\{\ell(Y_i, f(X_i)) + \text{penalty}(f)\}$$

Optimised from a mathematical point of view and **generalized** for all new observations

$$\widehat{Y} = \widehat{f}_n(X)$$

How Machine Learning could go possibly wrong ?

- AI **generalizes** the situation encountered in the learning sample to the whole population.
  It shapes the reality according to the learnt rule without questioning nor evolution.
- The density from **world of the data** may not represent **the real world**.
  The observations reflect **use** but they may be different from the **ideal model** we desire.
- Presence of Bias in the data set.

Acceptability of AI requires that the algorithm behaves in a fair way for all people.
But the learning sample may be biased or not reflect the desired behavior of the model

Examples

1. **Adult Income Data**.
   Data from a bank : Forecast from characteristics if someone has the potential to have a high income ($\geq 50k\$$) to grant a loan.
   Variables : Age, Workclass, Final weight, Education, Marital status, Occupation, Relationship, **Gender**, Race, Capital gain, Capital loss, Hours per week, Native country.
   Output $Y \in \{0, 1\}$ if predicted income is higher than the threshold or not.
   Protected Variables : Gender, Race, Native country.
   Result :
   $$\mathbb{P}(\hat{Y} = 1|S = 1) >> \mathbb{P}(\hat{Y} = 1|S = 0).$$
2. **ProPublica vs Northpoint**
3. **Bias in learning sample in image**

Examples

1. **Adult Income Data**.
2. **ProPublica vs Northpoint**
   Northpoint produces a score COMPAS to measure the probability of recidivism of offendants. This score has been designed using Machine Learning Algorithm from a learning sample to predict if someone will commit a crime when set free $Y = 0$.
   Variables : characteristics of people and their crime
   Protected Variable : Ethnic Origin $S = 0$ coding Afro-American.
   It is balanced
   $$\mathbb{P}(\hat{Y} = 1|S = 1) \sim \mathbb{P}(\hat{Y} = 1|S = 0).$$
   But the errors are different
   $$\mathbb{P}(\hat{Y} = 1|S = 1, Y = 0) >> \mathbb{P}(\hat{Y} = 1|S = 0, Y = 0).$$
3. **Bias in learning sample in image**

Examples

1. **Adult Income Data**.
2. **ProPublica vs Northpoint**
3. **Bias in learning sample in image**

S is snow in the background (using Lime package from *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* (2016))

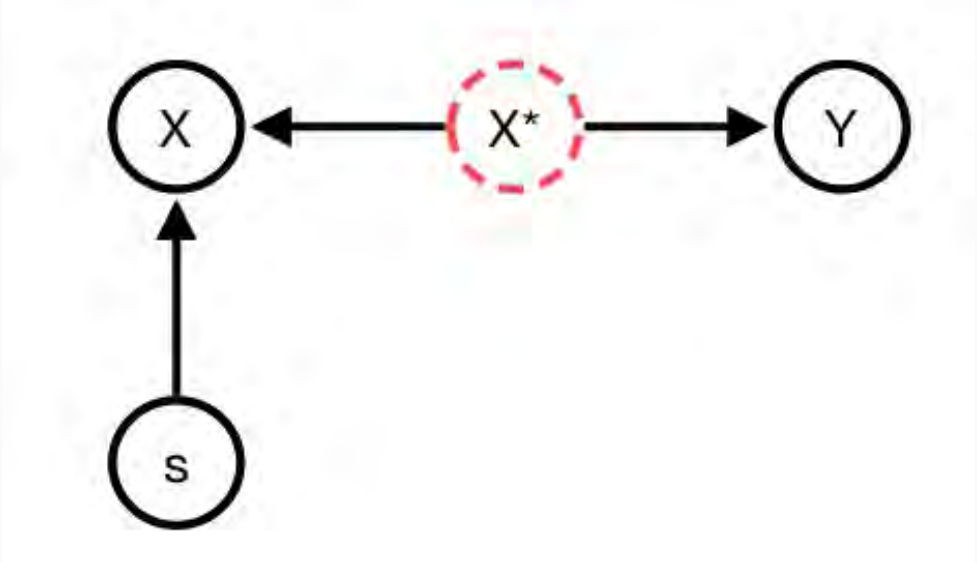
S is the color of colorized Mnist

Examples

1. **Adult Income Data**.
2. **ProPublica vs Northpoint**
3. **Bias in learning sample in image**

Removing the S variable is not enough since the X are highly correlated with S. Machine Learning Algorithm **amplifies the bias** and transforms the correlation into a causal relationship.

Mathematical Model for Fairness



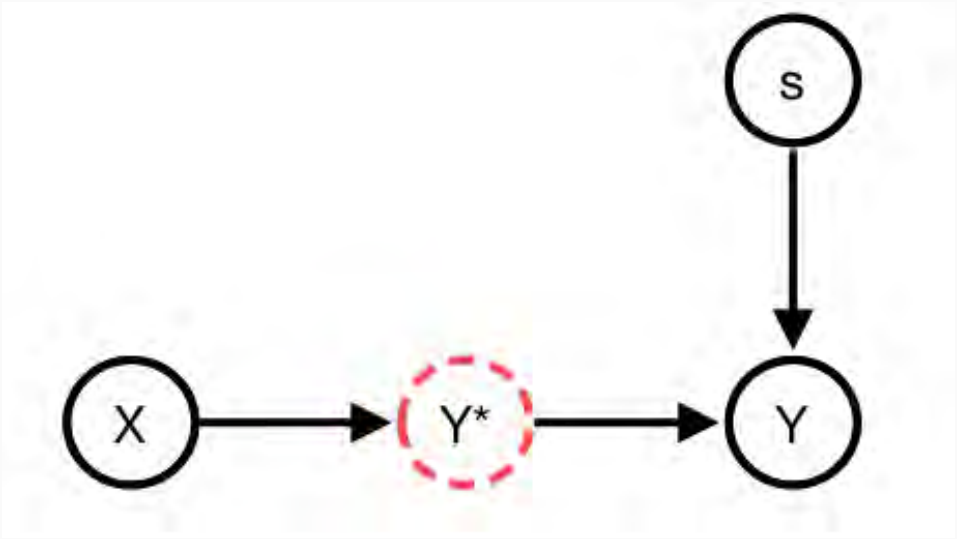
- Y **target**
- $X : \Omega \to \mathbf{R}^d,\ d \geqslant 1$, **visible attributes**
- $S : \Omega \to \{0, 1\}$ which induces a bias **protected attribute**

$$S = \begin{cases} 0 & \textit{minority} \quad (\textit{unfavored}) \\ 1 & \textit{majority} \quad (\textit{favored}) \end{cases}$$

(pictures from Loubes, Pauwels, Serrurier (2019))

Fairness deals with the relationships between Y, $\hat{Y}$ and S

Mathematical Model for Fairness



- Y **target**
- $X : \Omega \to \mathbf{R}^d,\ d \geqslant 1$, **visible attributes**
- $S : \Omega \to \{0, 1\}$ which induces a bias **protected attribute**

$$S = \begin{cases} 0 & \textit{minority} \quad (\textit{unfavored}) \\ 1 & \textit{majority} \quad (\textit{favored}) \end{cases}$$

(pictures from Loubes, Pauwels, Serrurier (2019))

Fairness deals with the relationships between Y, $\hat{Y}$ and S

Measuring Bias in Machine Learning

- $Y = \begin{cases} 0 & \textit{failure} \\ 1 & \textit{success} \end{cases}$ target class

Criteria of bias or unfairness

- Disparate Impact

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

$\rightarrow$ g is said not to have Disparate Impact at level $\tau \in (0, 1]$ if $DI(g, X, S) > \tau$

- Balanced Error Rate

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

$\rightarrow$ Given $\varepsilon > 0, S$ is not $\varepsilon-$predictable from X if $BER(g, X, S) > \varepsilon$

- New Criterion based on Distance between distributions of each class driven by S. **Wasserstein distance and Optimal Transport**

Monge-Kantorovich a.k.a Wasserstein distance

Consider the set $\mathcal{W}_2(\mathbf{R}^d)$ of probabilities with finite second moment
For any $\mu, \nu \in \mathcal{W}_2(\mathbf{R}^d)$, $\Pi(\mu, \nu)$ the set of all probability measures π over the product set $\mathbf{R}^d \times \mathbf{R}^d$ with first (resp. second) marginal μ (resp. ν)

- The Wasserstein distance of second order is defined as

$$W_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \, d\pi(x, y) = \min_{X \sim \mu, \ Y \sim \nu} \mathbb{E}\left(\|X - Y\|^2\right).$$

- The **Wasserstein Variation** with respect to weights $\omega_j, j = 1, \ldots, J$:

$$V_2\left(\mu_1, \ldots, \mu_J\right) = \inf_{\eta \in \mathcal{W}_2\left(\mathbb{R}^d\right)} \left(\sum_{j=1}^{J} \omega_j W_2^2(\mu_j, \eta)\right)^{1/2}$$
$$= \left(\sum_{j=1}^{J} \omega_j W_2^2(\mu_j, \mu_B)\right)^{1/2}$$

μ_B : **Wasserstein barycenter** (Agueh & Carlier (2015), Le Gouic & Loubes (2017), Del Barrio & Loubes (2018))

Application to a real data example: Adult Income data

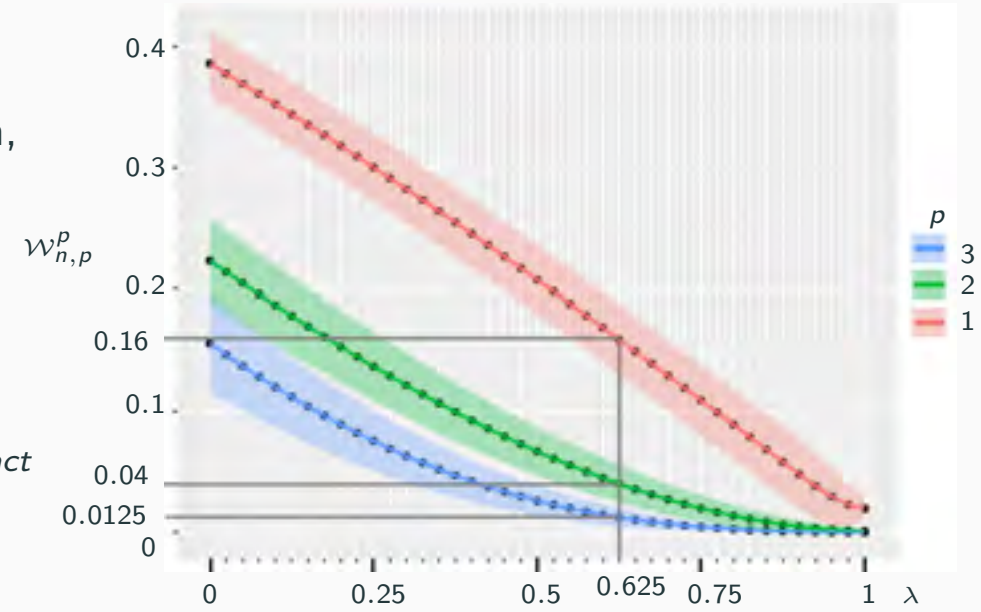
$$Y = \begin{cases} 1 & \text{income exceeds \$ 50.000/}\textit{year} \\ 0 & \text{otherwise} \end{cases}$$

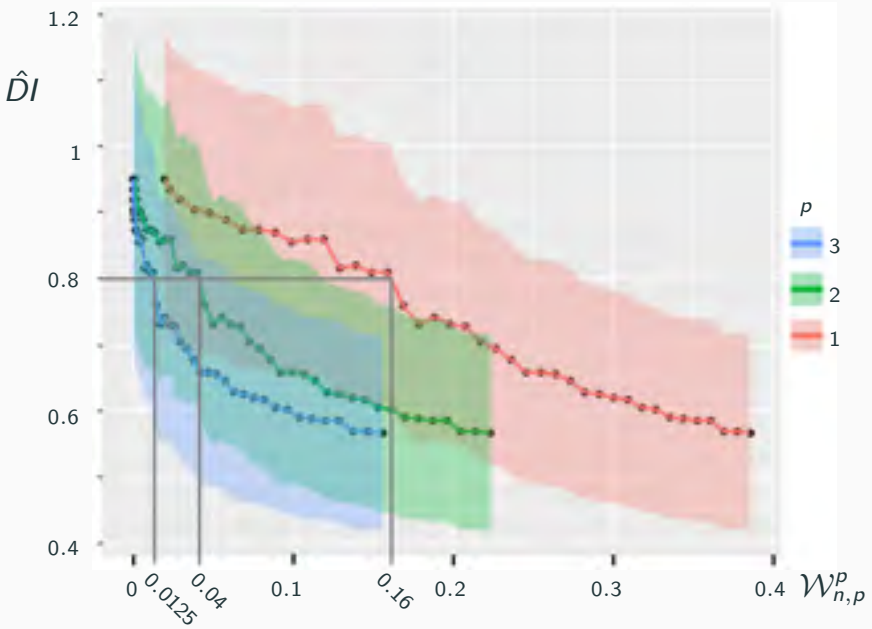
X =(age, education number, capital gain, capital loss, worked hours/ week)

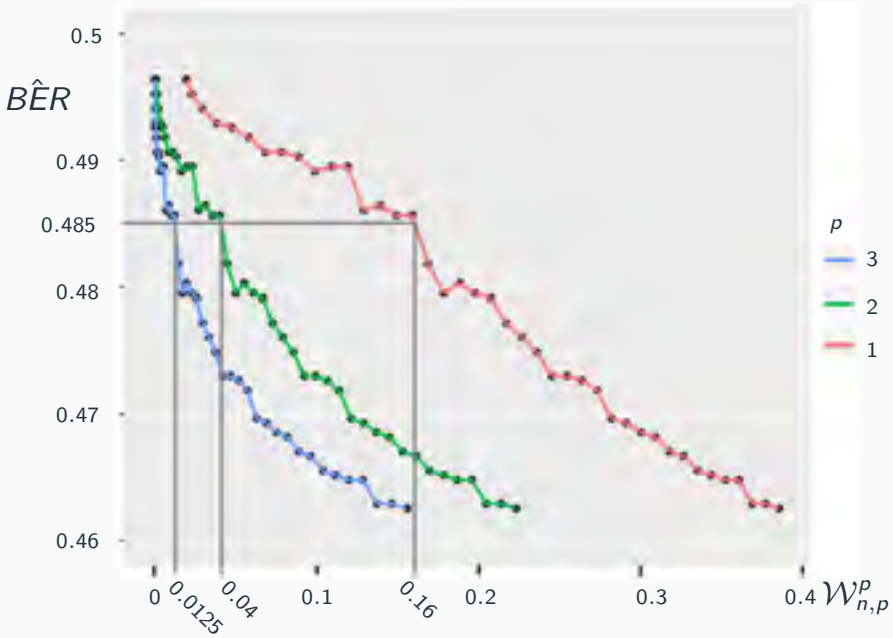
$$S = \text{gender} \begin{cases} 0 & \textit{female} \\ 1 & \textit{male} \end{cases}$$

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

↓







Achieving Fairness

Different strategies to ensure Fairness : independency w.r.t to S

- Finding Classifiers g such that $\mu_0(g) := \mathcal{L}(g(X)|S = 0)$ **is close to** $\mu_1(g) := \mathcal{L}(g(X)|S = 1)$ by adding **a penalty**
- Modify the input data $\Rightarrow$ to break the relationship with the protected attribute
  Changing the data X into $\tilde{X}$
  such that $\mu_0 := \mathcal{L}(\tilde{X}|S = 0)$ **is close to** $\mu_1 := \mathcal{L}(\tilde{X}|S = 1)$ to gain fairness of all possible classifiers constructed using $\tilde{X}$.

Quantify accurately the modification of the distributions : trade-off between fairness and accuracy to the observations in order to provide a **certification for generalization of the model**

Achieving Statistical Parity (Proceedings ICML 2019)

- **Goal:**

$$X \longrightarrow \tilde{X} \text{ such that } \mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right)$$

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

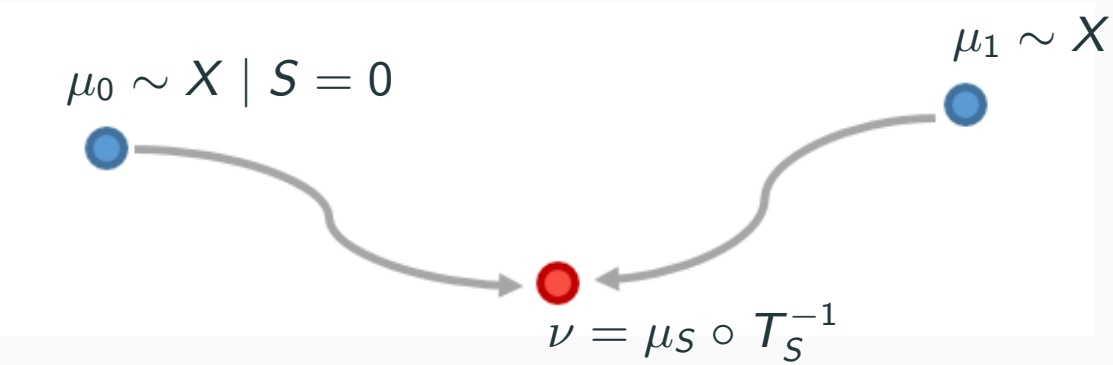
$$\Rightarrow DI(g, \tilde{X}, S) = 1$$

- **Methodology:**

$$\begin{array}{rccc} T_S: & \mathbf{R}^d & \longrightarrow & \mathbf{R}^d \\ & X & \longmapsto & \tilde{X} = T_S(X) \end{array} \quad \text{s.t.}$$

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

- T_S depends on the binary random variable S



1. **Best choice for the distribution ν of the repaired variable?**
2. **Optimal way of transporting μ_1 and μ_0 to this new distribution ν?**

Achieving Statistical Parity (Proceedings ICML 2019)

- **Goal:**

$$X \longrightarrow \tilde{X} \text{ such that } \mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right)$$

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

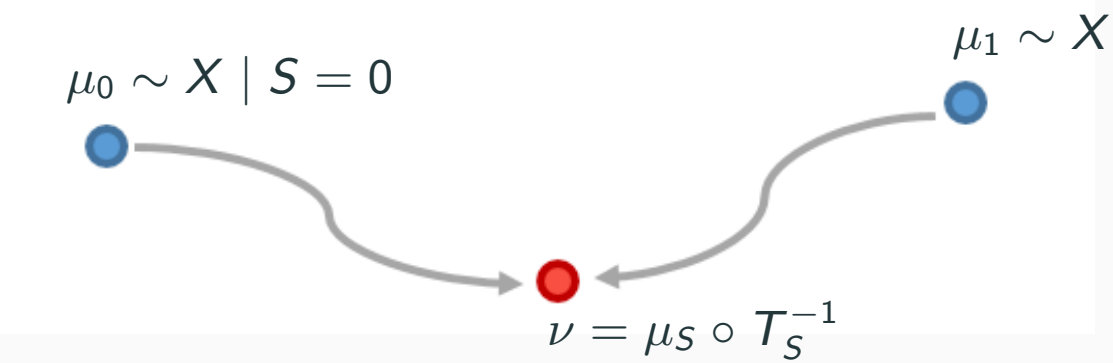
$$\Rightarrow DI(g, \tilde{X}, S) = 1$$

- **Methodology:**

$$\begin{array}{rccc} T_S: & \mathbf{R}^d & \longrightarrow & \mathbf{R}^d \\ & X & \longmapsto & \tilde{X} = T_S(X) \end{array} \quad \text{s.t.}$$

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

- T_S depends on the binary random variable S



1. **Best choice for the distribution ν of the repaired variable?**

$\Rightarrow$ **Wasserstein barycenter** proposed in Fair Learning litterature

2. **Optimal way of transporting μ_1 and μ_0 to this new distribution ν?**

Achieving Statistical Parity (Proceedings ICML 2019)

- **Goal:**

$$X \longrightarrow \tilde{X} \text{ such that } \mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right)$$

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

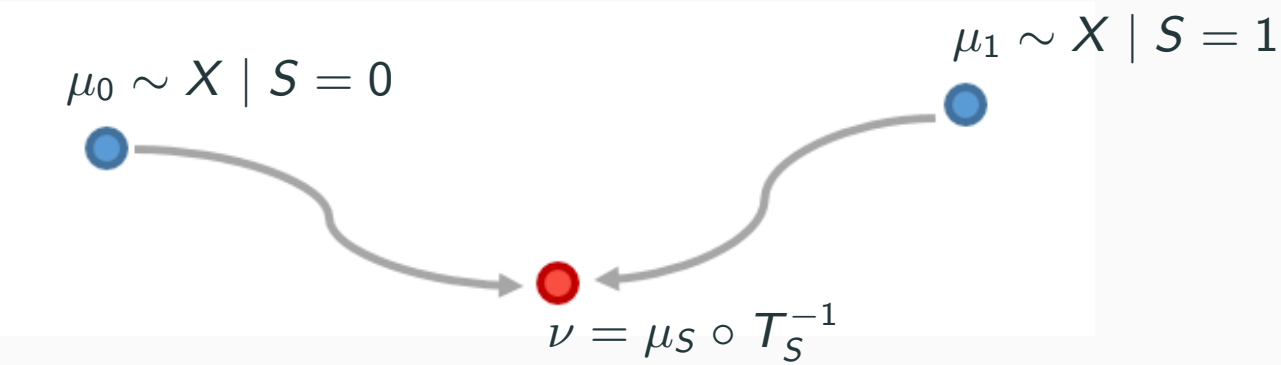
$$\Rightarrow DI(g, \tilde{X}, S) = 1$$

- **Methodology:**

$$\begin{array}{rccc} T_S: & \mathbf{R}^d & \longrightarrow & \mathbf{R}^d \\ & X & \longmapsto & \tilde{X} = T_S(X) \end{array} \quad \text{s.t.}$$

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

- T_S depends on the binary random variable S



1. **Best choice for the distribution ν of the repaired variable?**

$\Rightarrow$ **Wasserstein barycenter** proposed in Fair Learning litterature

2. **Optimal way of transporting μ_1 and μ_0 to this new distribution ν?**

$\Rightarrow$ **Optimal Transport Maps**

Controlling the loss of information in the new classification

- **Amount of information lost when replacing X by $\tilde{X}$?**
- Risk when the full data (X, S) is available

$$R(g, X, S) := \mathbb{P}(g(X, S) \neq Y) \dashrightarrow R_B(X, S) = \inf_g R(g, X, S) = R(g_B, X, S)$$

- In the repaired data $\tilde{X} = T_S(X)$

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

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

Theorem (Upper bound for cost for fairness)

For each $s \in \{0, 1\}$, assume that the function $\eta_s(x) = \mathbb{P}(Y = 1 \mid X = x, S = s)$ is Lipschitz with constant $K_s > 0$. Then, if $K = \max\{K_0, K_1\}$,

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

Conclusion(s)

- Fairness constraints enable either to increase the accuracy of the forecast by removing learning sample unwanted bias
- ... or shape a *fair* reality

And provides a control on the generalization power of Machine Learning algorithms (certification).

- Challenge : definition of bias in sample
- Automatic detection of unknown bias and detection of areas with bias

- New methods with DEEL partners : fair clustering, ressource allocation, PCA, auto-encoders, GAN, online algorithms ...
- Related Fields : Transfert Learning and Domain Generalization

E. del Barrio, P. Gordaliza and J.-M. Loubes. (2019)
A central limit theorem on the real line with application to fairness assessment in machine learning.
Information and Inference.

E. del Barrio, P. Gordaliza and J.-M. Loubes. (2019)
Obtaining Fairness with Optimal Transportation
Proceedings of ICML

E. del Barrio and J.-M. Loubes. (2019)
Central limit theorems for empirical transportation cost in general dimension
The Annals of Probability