NEW BRIDGES BETWEEN MATHEMATICS AND DATA SCIENCE

November 8^{th}-11^{th}, Valladolid, Spain

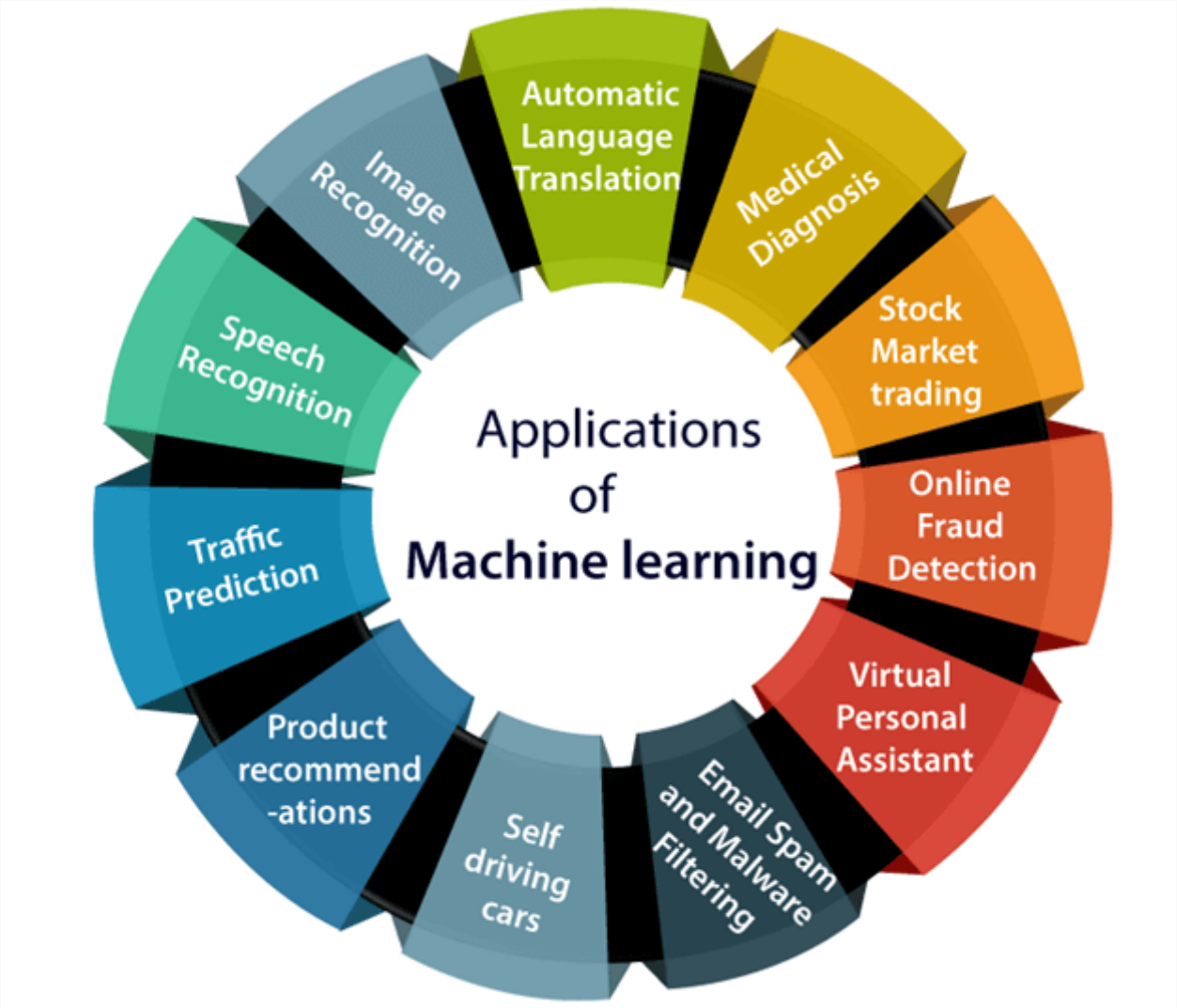
Mathematical frameworks for fair learning: review of methods and study of the price for fairness

PAULA GORDALIZA PASTOR

Joint work with

Eustasio del Barrio
Fabrice Gamboa
Jean-Michel Loubes
Philippe Besse
Laurent Risser

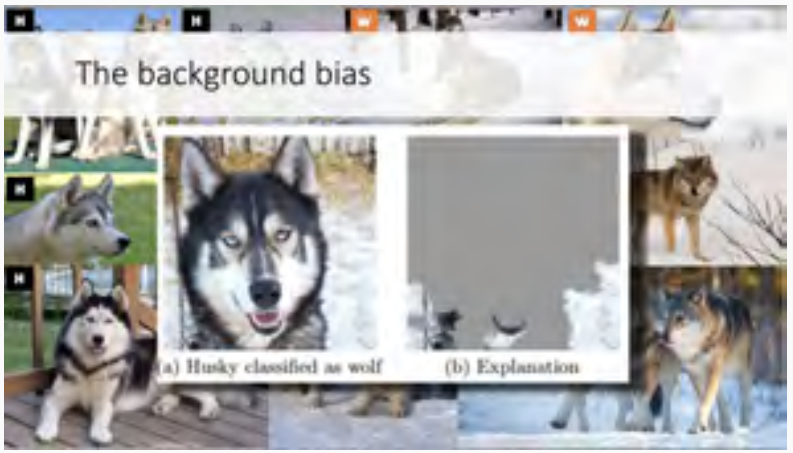
The generalization of applications based on ML models in the everyday life and the professional world has been accompanied by concerns about the ethical issues that may arise from the adoption of these technologies



AI technologies make life easier, but they are not absolutely objective...

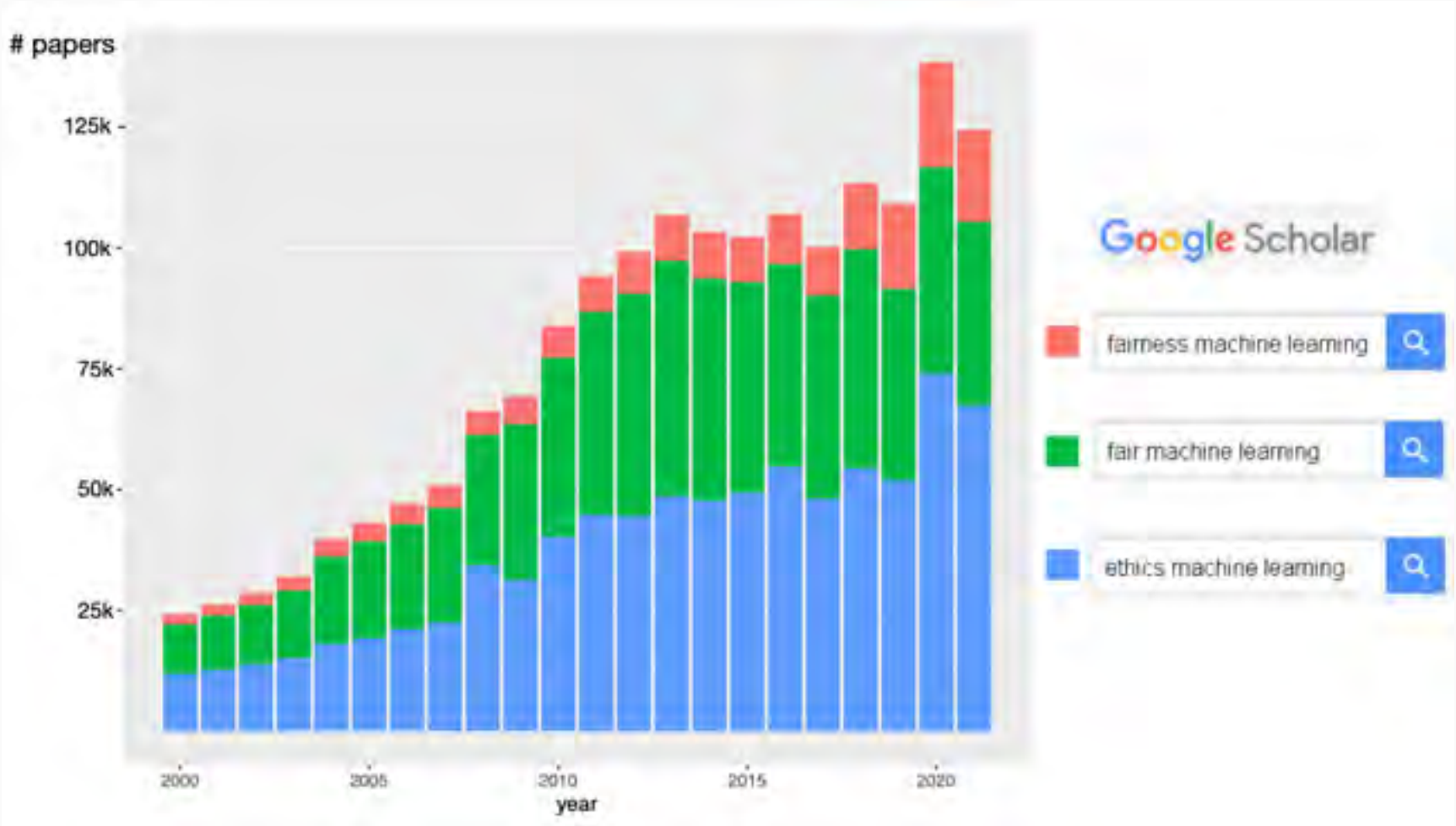
ML algorithms that are meant to automatically take accurate and efficient decisions that mimic, and even sometimes outmatch human expertise, rely heavily on potentially biased data

Inherent social bias existing in the population that is used to generate the training set



Bias without social unfairness

Fairness has become one of the most popular topics in ML over the last years and the research community is investing a **large amount of effort** in this area.



COMPAS recidivism black bias

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Results from job platform XING

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

TABLE II: Top k results on www.xing.com (Jan 2017) for the job serach query "Brand Strategist".

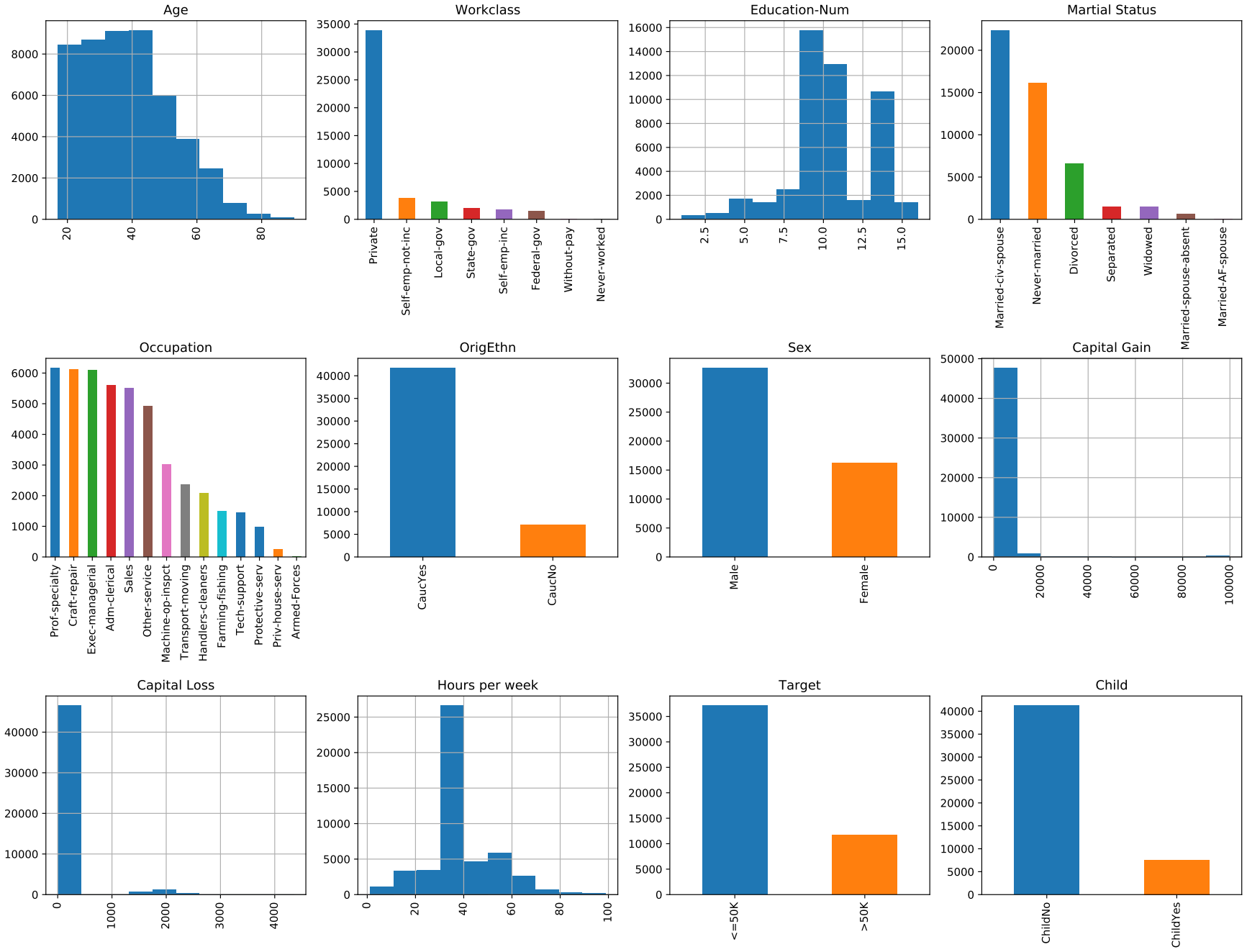
Less qualified male candidates were highly ranked

ML algorithms in banking industry: The Adult Data Set

Adult Data Set

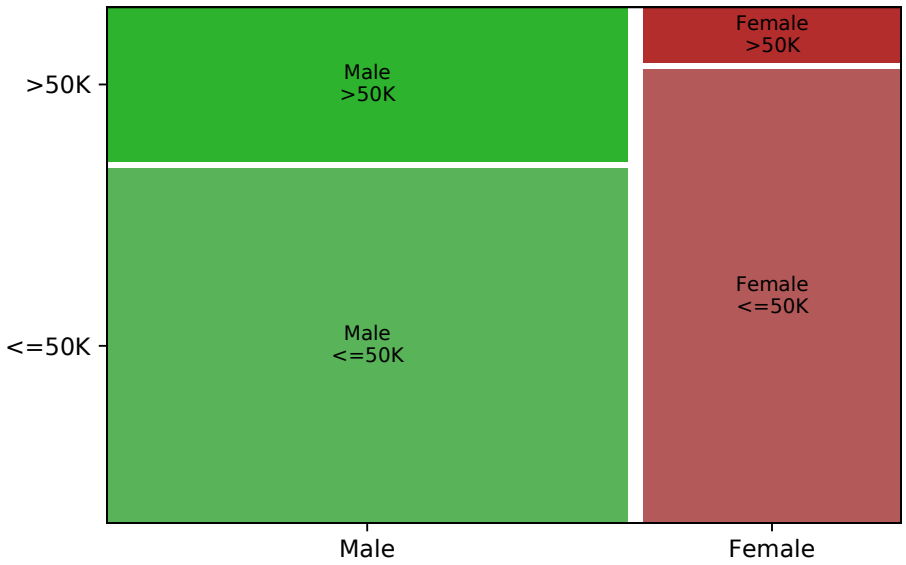
UCI Machine Learning Repository

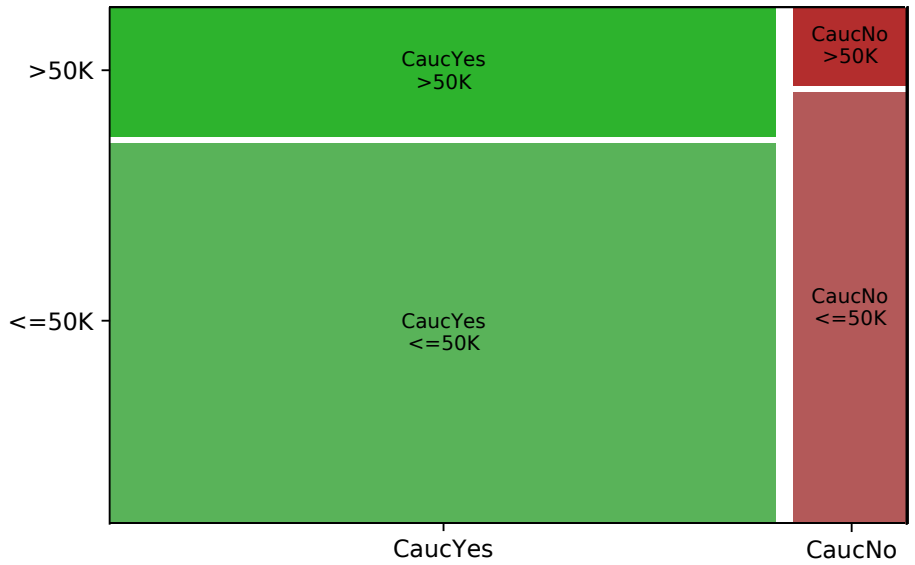
Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.



Goal : Forecast solvability (salary $> 50k\$$) to minimize the risk of granting a loan

Problem : The learning set is biased with respect to **Gender** and **Ethnic Origin**





What information would be learnt from a biased dataset?

Consider $\left(\Omega \subset \mathbb{R}^d, \mathcal{B}, \mathbb{P}\right)$, $\mathcal{B}$ Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geqslant 1$

Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \mathbb{R}^d$	$\hat{Y} = f(X, S),\ f \in \mathcal{F}$

Definition of fairness as independence criterion

Perfect fairness requires that S does not play any role in the forecast $\hat{Y}$

(I) **Statistical Parity** : $\hat{Y} \perp\!\!\!\perp S$

(II) **Equality of Odds** : $\hat{Y} \perp\!\!\!\perp S \mid Y$

Consider $\left(\Omega \subset \mathbb{R}^d, \mathcal{B}, \mathbb{P}\right)$, $\mathcal{B}$ Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geqslant 1$

Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S} = \{0, 1\}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \{0, 1\}$	$\hat{Y} = g(X, S),\ g \in \mathcal{G}$
$\begin{cases} 0 & unfavored \\ 1 & favored \end{cases}$		$\begin{cases} 0 & failure \\ 1 & success \end{cases}$	$g : \mathbb{R}^d \to \{0, 1\}$

Definition of fairness as independence criterion

Perfect fairness requires that S does not play any role in the forecast $\hat{Y}$

(I) **Statistical Parity** (SP) (Dwork et al., 2012): $\hat{Y} \perp\!\!\!\perp S$

$$\mathbb{P}(\hat{Y} = 1 \mid S = 0) = \mathbb{P}(\hat{Y} = 1 \mid S = 1)$$

(II) **Equality of Odds** (EO) (Hardt et al., 2016): $\hat{Y} \perp\!\!\!\perp S \mid Y$

$$\mathbb{P}(\hat{Y} = i \mid Y = i, S = 0) = \mathbb{P}(\hat{Y} = i \mid Y = i, S = 1),\ i = 0, 1$$

Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S} = \{0,1\}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \{0,1\}$	$\hat{Y} = g(X,S),\ g \in \mathcal{G}$
$\begin{cases} 0 & unfavored \\ 1 & favored \end{cases}$		$\begin{cases} 0 & failure \\ 1 & success \end{cases}$	$g : \mathbb{R}^d \to \{0,1\}$

The **Disparate Impact** of the classifier $g \in \mathcal{G}$, with respect to (X, S) is defined as

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

- Ideal scenario: g achieves Statistical Parity $\Leftrightarrow DI(g, X, S) = 1$

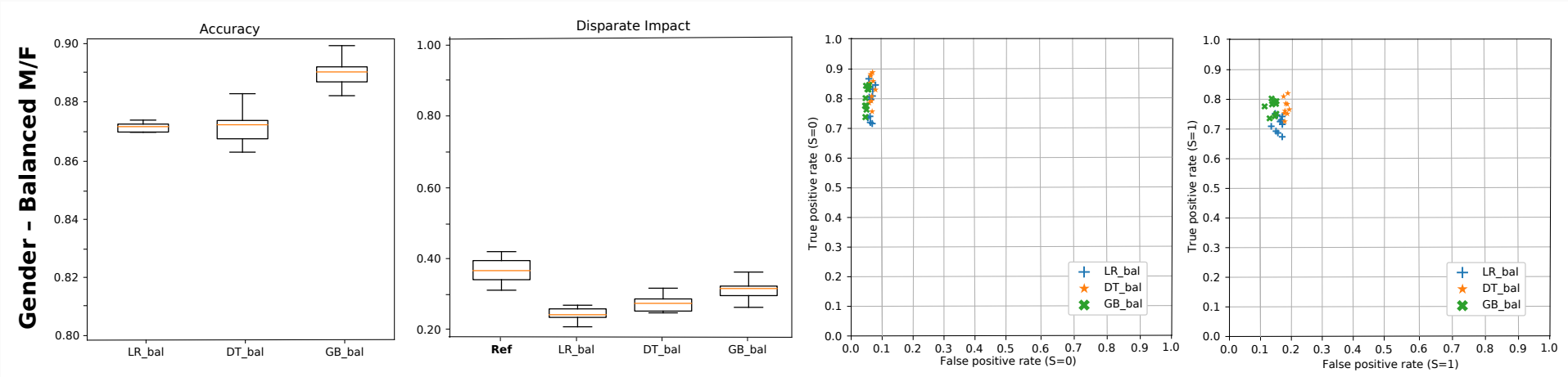
Definition

A classifier $g : \mathbb{R}^d \to \{0,1\}$ is said not to have **Disparate Impact at level** $\tau \in [0,1]$, with respect to (X, S), if $DI(g, X, S) > \tau$.

- $\tau_0 = 4/5$ --→ 80% rule (*1971, State of California Fair Employment Comission*)

Some ineffective standard procedures...

1. The problem cannot be solved by simply having a balanced amount of observations with $S = 0$ and $S = 1$



Some ineffective standard procedures...

2. What if the sensitive variable S is removed?

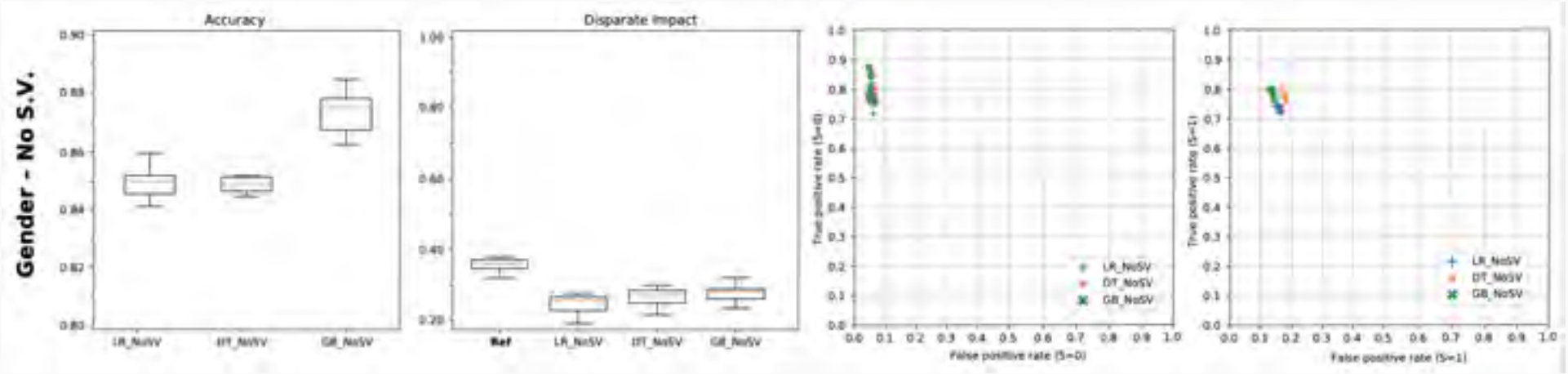
- European laws (GDPR) have prohibited recording sensitive data such as gender or ethnic origin for reasons of confidentiality so far.
- The paradoxical consequence is the impossibility of directly measuring possible discrimination.

  Proposal for a Regulation laying down harmonised rules on artificial intelligence (European Comission, 2020) authorizing, subject to privacy, the consideration of statistics of sensitive variables.

Some ineffective standard procedures...

2. What if the sensitive variable S is removed?

Performance of the ML models LR, DT and GB when removing the *Gender* variable



Some ineffective standard procedures...

3. Testing procedures (since 1939)

- French justice has taken them as a proof of biased treatment since 2006 (sociological studies of “Observatoire des discriminations”)
- Testing procedures are often used as a legal proof for discrimination:

$$(x, s) \rightarrow \text{artificial individual } (x, s')$$

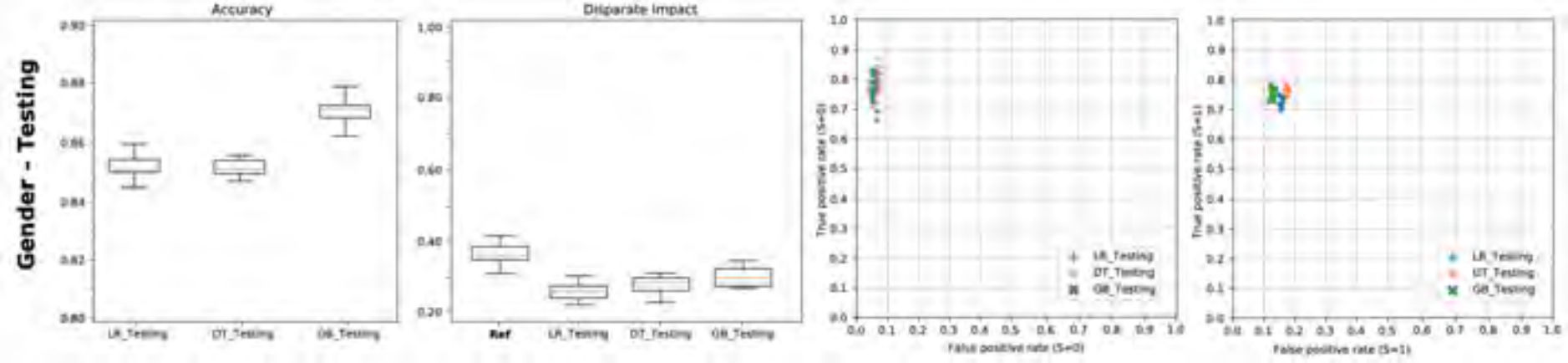
$$\text{If } \hat{y}' = f(x, s') \neq \hat{y} = f(x, s) \Rightarrow \text{legal proof for discrimination}$$

- Algorithmic solution to bypass this testing procedure:
  1.- Train a classifier $f(X, S)$ using all available information X, S
  2.- Build $\tilde{f}$ *testing compliant* version of f:
  $\tilde{f}(x, s) =$ best decision obtained on actual individual $f(x, s)$ and artificial individual $f(x, s')$

Some ineffective standard procedures...

3. Testing procedures

Performance of the ML models LR, DT and GB when using a testing procedure



Obtaining fairness is a far more complicated task that needs mathematical models

The performance of an algorithm $f \in \mathcal{F}$ is measured through its risk defined by

$$R(f) = \mathbb{E}(\ell(Y, f(X, S)))$$

with $\ell : \ (Y, \hat{Y}) \mapsto \ell(Y, \hat{Y}) \in \mathbb{R}^+$ a loss function

An **optimal fair model** can be achieved by restricting the risk minimization to a fair class of models $\mathcal{F}_{\text{Fair}}$

$$\inf_{f \in \mathcal{F}_{\text{Fair}}} R(f)$$

The **price for fairness** of the class $\mathcal{F}_{\text{Fair}}$ is

$$\mathcal{E}(\mathcal{F}_{\text{Fair}}) := \inf_{f \in \mathcal{F}_{\text{Fair}}} R(f) - \inf_{f \in \mathcal{F}} R(f)$$

$\rightarrow$ **Bayes estimator** $\eta(X, S) :=$ minimizer of $\inf_{f \in \mathcal{F}} R(f)$

Statistical parity $\quad \mathcal{F}_{\text{Fair}} := \mathcal{F}_{SP} = \{f(X, S) \in \mathcal{F} \quad \text{s.t.} \quad \hat{Y} \perp\!\!\!\perp S\}$

Equality of odds $\quad \mathcal{F}_{\text{Fair}} := \mathcal{F}_{EO} = \{f(X, S) \in \mathcal{F} \quad \text{s.t.} \quad \hat{Y} | Y \perp\!\!\!\perp S\}$

Price for statistical parity in regression

Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \mathbb{R}^d$	$\hat{Y} = f(X, S),\ f \in \mathcal{F}$

$$R(f) = \mathbb{E}\|Y - f(X,S)\|^2$$

minimizer of $\inf_{f\in\mathcal{F}} R(f) \to \eta(X,S) := \mathbb{E}[Y|(X,S)]$ (Bayes estimator)

Lower bound for the price for fairness (Le Gouic et al. (2020)) If $d = 1$, $S \in \{1, ..., k\}$, $\zeta_S := \mathcal{L}(\eta(X,S) \mid S)$ and $\nu_S(g) := \mathcal{L}(g(X,S) \mid S)$. Then

$$\mathcal{E}(\mathcal{F}_{\text{Fair}}) \geq \inf_{g\in\mathcal{F}} \mathbb{E}\mathcal{W}_2^2(\zeta_S, \nu_S(g)).$$

Equality holds for $\mathcal{F}_{\text{SP}}$: if ζ_s has density w.r.t. Lebesgue measure for almost every s,

$$\mathcal{E}(\mathcal{F}_{\text{SP}}) = \inf_{g\in\mathcal{F}_{SP}} \mathbb{E}_S\mathcal{W}_2^2(\zeta_S, \nu_S(g)) = \inf_{\nu(g)} \mathbb{E}_S\mathcal{W}_2^2(\zeta_S, \nu(g)) = \inf_{\nu} \mathbb{E}_S\mathcal{W}_2^2(\zeta_S, \nu)$$

if $\mathcal{S} = \{0,1\}$ $\quad = \pi_0\mathcal{W}_2^2(\zeta_0, \zeta_B) + (1-\pi_0)\mathcal{W}_2^2(\zeta_1, \zeta_B)$,

ζ_B Wasserstein barycenter of ζ_0 and ζ_1 w.r.t. $\pi_s = \mathbb{P}(S = s)$

- Similar results in Chzhen et al. (2020)
- Extension to general dimension in Le Gouic and Loubes (2020)

Price for statistical parity in classification

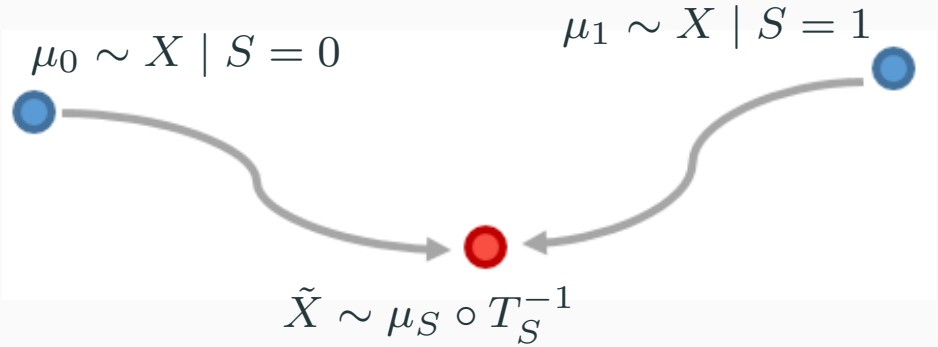
Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S} = \{0, 1\}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \{0, 1\}$	$\hat{Y} = g(X, S),\ g \in \mathcal{G}$ $g : \mathbb{R}^d \to \{0, 1\}$

Find $\tilde{X} = T_S(X)$ such that $\mathcal{L}\left(T_0(X) \mid S = 0\right) = \mathcal{L}\left(T_1(X) \mid S = 1\right)$

$$\Downarrow$$

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

Fair classifier: $g \circ T_S \in \mathcal{F}_{SP}$, for all $g \in \mathcal{G}$



Questions:

a) Best choice for the distribution $\tilde{X} \sim \nu$?

b) Optimal way of transporting μ_0, μ_1 to ν?

Upper bound for the price for fairness (Gordaliza et al., 2019) If $\eta_s(x) = \mathbb{P}(Y = 1 \mid X = x, S = s), s \in \{0, 1\}$, is Lipschitz with constant $K_s > 0$ and $K = \max\{K_0, K_1\}$,

$$\mathcal{E}(T_S) := \inf_{g \in \mathcal{G}} \mathbb{P}(g(T_S(X)) \neq Y) - \inf_{g \in \mathcal{G}} R(g) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, {\mu_s}_\sharp T_s) \right)^{\frac{1}{2}}.$$

Upper bound for the price for statistical parity

$$\mathcal{E}(\mathcal{F}_{SP}) \leq \inf_{T_S} \mathcal{E}(T_S) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_B) \right)^{\frac{1}{2}}$$

Reasonable and feasible solutions:

a) Wasserstein barycenter μ_B with weights $\pi_0 = P(S = 0)$ and $\pi_1 = P(S = 1)$

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

b) T_S optimal transport map carrying μ_S towards μ_B

$${\mu_S}_\sharp T_S = \mu_B$$

Price for equality of odds in regression

Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \mathbb{R}^d$	$\hat{Y} = f(X, S),\ f \in \mathcal{F}$

Related work Woodworth et al. (2017)

Normal model: $(X_1, S_1, Y_1), \ldots, (X_n, S_n, Y_n)$ i.i.d. observed from (X, S, Y) and $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, 1)$

$$Y = f_{\beta_0,\beta}(X, S) + \varepsilon, \text{ where}$$

$$f_{\beta_0,\beta}(X, S) = \beta_0 S + \beta^T X,\ \beta_0 \in \mathbb{R},\ \beta \in \mathbb{R}^{p\times 1} \tag{5.1}$$

$$\text{and} \quad \mathbb{E}(\varepsilon \mid (X, S)) = 0$$

$$(X, S, Y) \sim \mathcal{N}_{p+2}\left(\begin{bmatrix} \mu_X \\ \mu_S \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_X & \Sigma_{XS} & \Sigma_{XY} \\ \Sigma_{XS}^T & \Sigma_S & \Sigma_{SY} \\ \Sigma_{XY}^T & \Sigma_{SY}^T & \Sigma_Y \end{bmatrix}\right)$$

Fair linear predictor:

$$f_{\beta_0,\beta}(X, S) \in \mathcal{F}_{EO} \Leftrightarrow f_{\beta_0,\beta}(X, S) \perp\!\!\!\perp S \mid Y \Leftrightarrow Cov(f_{\beta_0,\beta}(X, S), S \mid Y) = 0$$

Proposition (del Barrio et al., 2020)

Under the normal model, the ***optimal fair linear predictor*** *of the form* (5.1) *is given as the solution to the following optimization problem*

$$\left(\hat{\beta}_{0,fair}, \hat{\beta}_{fair}\right) := \textit{argmin}_{(\beta_0,\beta)\in\mathcal{F}_{EO}} \mathbb{E}\left[(Y - f_{\beta_0,\beta}(X,S))^2\right]$$

$$\mathcal{F}_{EO} = \{(\beta_0,\beta) \in \mathbb{R}\times\mathbb{R}^p \textit{ s.t. } \beta^T\left(\Sigma_{XS}\Sigma_Y - \Sigma_{SY}\Sigma_{XY}\right) + \beta_0\left(\Sigma_S\Sigma_Y - \Sigma_{SY}^2\right) = 0\}.$$

If moreover Y *and* S *are not linearly dependent, it can be exactly computed as*

$$\hat{\beta}_{0,fair} = \hat{\beta}_{fair}^T C_{S,X,Y}$$
$$\hat{\beta}_{fair} = \Sigma_Z^{-1}\Sigma_{ZY},$$

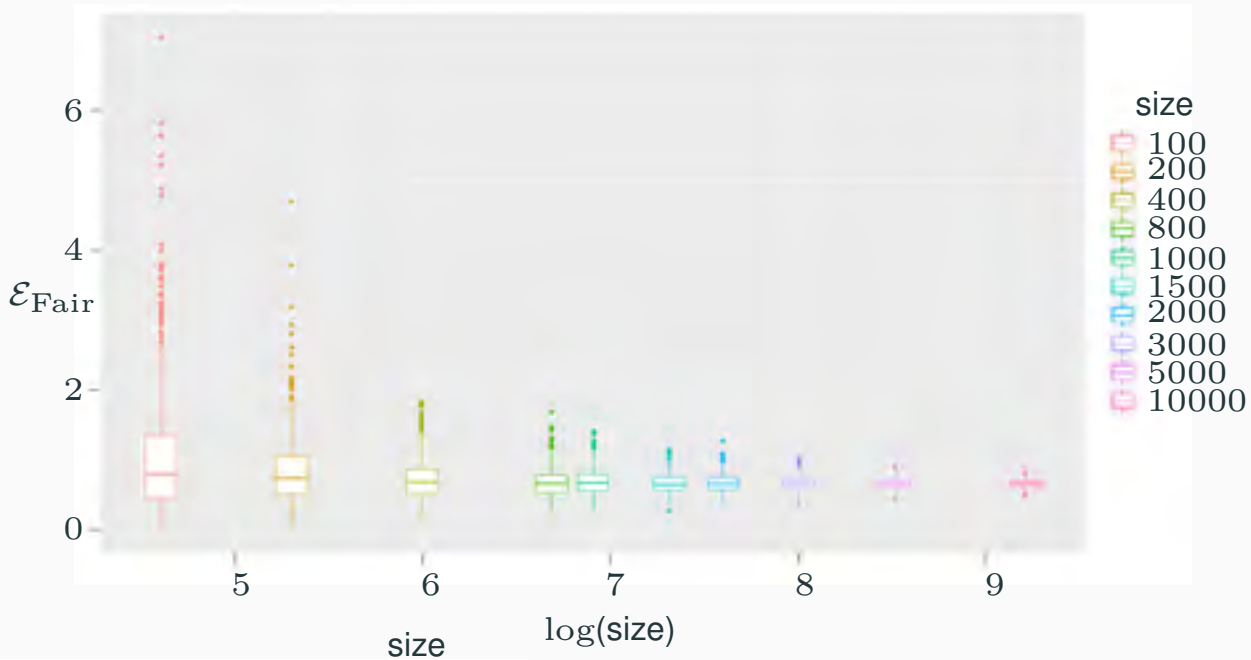
where

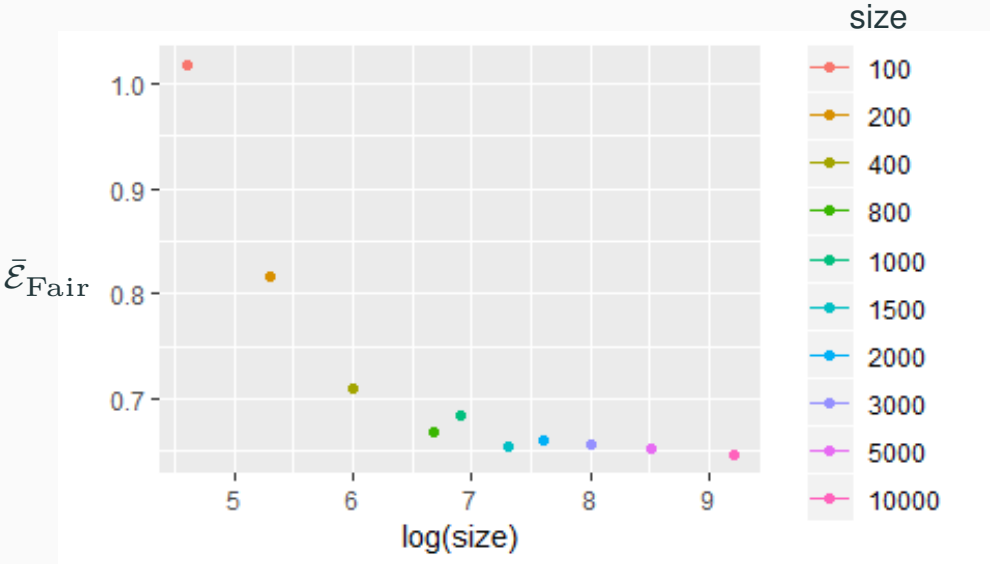
$$\Sigma_Z = \Sigma_X + \Sigma_S C_{S,X,Y} C_{S,X,Y}^T + C_{S,X,Y}\Sigma_{XS}^T + \Sigma_{XS} C_{S,X,Y}^T$$
$$\Sigma_{ZY} = \Sigma_{XY} + \Sigma_{SY} C_{S,X,Y}.$$

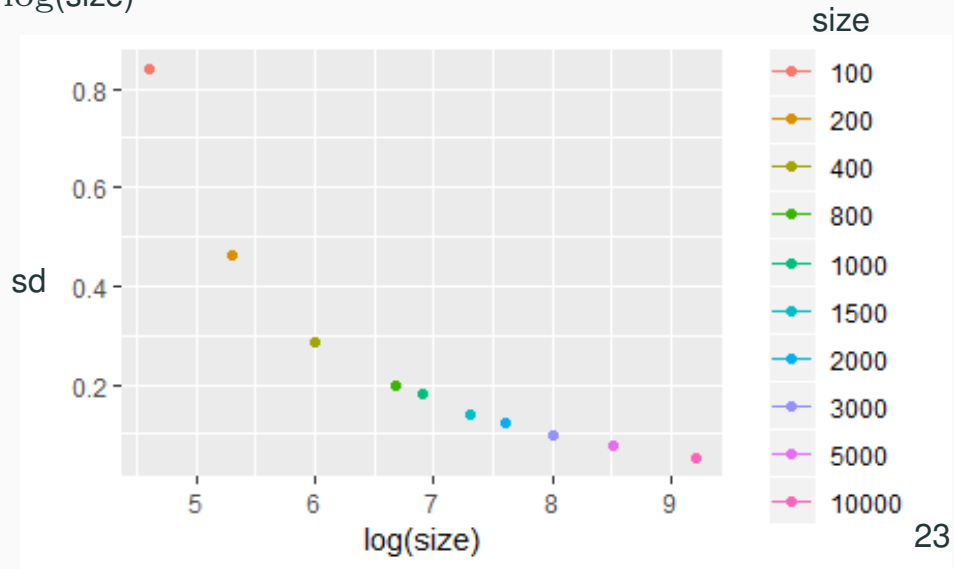
and

$$C_{S,X,Y} := \left(\frac{\Sigma_{XS}\Sigma_Y - \Sigma_{SY}\Sigma_{XY}}{\Sigma_S\Sigma_Y - \Sigma_{SY}^2}\right) \in \mathbb{R}^{p\times 1} \textit{ vector of correction for fairness}$$

Simulations: $S \sim \mathcal{N}(0, 10)$ and $X \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}\right)$







Price for equality of odds in classification

Protected attribute	Visible attributes	Target	Outcome
$S \in \mathcal{S} = \{0, 1\}$	$X \in \mathcal{X} \subset \mathbb{R}^d$	$Y \in \{0, 1\}$	$\hat{Y} = g(X, S),\ g \in \mathcal{G}$ $g : \mathbb{R}^d \to \{0, 1\}$

Assume $\mathbb{P}(Y = 1) \in (0, 1),\ \mathbb{P}(S = 1) \in (0, 1)$ and $\mathbb{P}(Y = 1, S = 1) \in (0, 1)$

Fair classifier:

$$\mathcal{F}_{EO} := \{g \in \mathcal{G} : \mathbb{P}(g(X, S) = i \mid Y = i, S = 0) = \mathbb{P}(g(X, S) = i \mid Y = i, S = 1), i = 0, 1\}.$$

Extention of **optimal fair (equality of opportunity) classifier** (Chzhen et al., 2019)

equality of opportunity = equality of TP rates across both groups ($i = 1$)

Proposition (del Barrio et al., 2020)

For each $s \in \{0,1\}$*, assume that the mapping* $t \in \mathbb{P}(\eta(X,S) \leq t \mid S = s)$ *is continuous on* $(0,1)$*. An* ***optimal fair classifier*** g^* *can be obtained for all* $(x,s) \in \mathbb{R}^d \times \{0,1\}$ *as*

$$g^*(x,1) = \mathbb{1}_{\{1 \leq 2\eta(X,1) - \theta_1^* \frac{\eta(X,1)}{\mathbb{P}(Y=1,S=1)} + \theta_0^* \frac{1-\eta(X,1)}{\mathbb{P}(Y=0,S=1)}\}}$$
$$g^*(x,0) = \mathbb{1}_{\{1 \leq 2\eta(X,0) + \theta_1^* \frac{\eta(X,0)}{\mathbb{P}(Y=1,S=0)} - \theta_0^* \frac{1-\eta(X,0)}{\mathbb{P}(Y=0,S=0)}\}},$$

where $(\theta_0^*, \theta_1^*) \in \mathbb{R}^2$ *is determined from equations*

$$\frac{\mathbb{E}_{X|S=1}\Big[\eta(X,1)g^*(X,1)\Big]}{\mathbb{P}(Y=1 \mid S=1)} = \frac{\mathbb{E}_{X|S=0}\Big[\eta(X,0)g^*(X,0)\Big]}{\mathbb{P}(Y=1 \mid S=0)}$$
$$\frac{\mathbb{E}_{X|S=1}\Big[(1-\eta(X,1))g^*(X,1)\Big]}{\mathbb{P}(Y=0 \mid S=1)} = \frac{\mathbb{E}_{X|S=0}\Big[(1-\eta(X,0))g^*(X,0)\Big]}{\mathbb{P}(Y=0 \mid S=0)}.$$

- $\theta_0^* = 0 \rightarrow$ optimal fair *equality of opportunity* classifier in Chzhen et al. (2019)
- If moreover $\theta_1^* = 0 \rightarrow$ classical Bayes rule $g_B(X,S) := \mathbb{1}_{\{1 \leq 2\eta(X,S)\}}$.

Methods for imposing a level of fairness

From a procedural point of view (Oneto and Chiappa, 2020)

Fairness through Optimal Transport

(A) Pre-processing the training data

Kamiran and Calders (2009,2010,2012)
Zemel et al. (2013)
Feldman et al. (2015)
Johndrow and Lum (2017)
Gordaliza et al. (2019)

(C) Post-processing the model outputs

Pedreschi et al. (2009)
Hardt et al. (2016)
Kusner et al. (2017)
Chzhen et al. (2019)

(B) In-processing to control the training phase of the algorithm

(i) Lagrange multipliers
Berk et al. (2017a)
Zafar et al. (2017a, 2019)
Agarwal et al. (2018)

(ii) Add penalties to the objective
Bechavod and Ligett (2017)
Dwork et al. (2018)
Donini et al. (2018)

Fairness through empirical risk minimization

Recently very important : **causal approach** S. Chiappa et al. (2019) and **counterfactual approach** de Lara et al. (2021)

Target variable	Level of repair	Transformation	$\mu = \mu_S{}_\sharp T_S$
$Z \sim \mu_B$	$\lambda \in [0,1]$	o.t.m. T_S	$T_S^{-1}(Z) \sim \mu_S$

Geometric repair (Feldman et al., 2015)

$$\tilde{\mu}_{S,\lambda} = \mathcal{L}(\lambda T_S(X) + (1-\lambda)X)$$

Random repair (Gordaliza et al., 2019)

$B \sim \mathcal{B}(\lambda)$, independent of (X, S, Y)

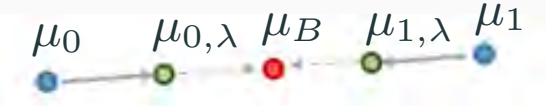
$$\tilde{\mu}_{S,\lambda} = \mathcal{L}(BT_S(X) + (1-B)X)$$

Unmodified variable	$0 \longleftarrow \lambda \longrightarrow 1$	Totally repaired variable
$\tilde{\mu}_{s,0} = \mathcal{L}(X \mid S = s)$	$\Updownarrow$	$\tilde{\mu}_{s,1} = \mathcal{L}(Z) = \mu_B$
Accuracy of $g(\tilde{X})$	$\longleftarrow$ **Trade-off** $\longrightarrow$	**Non-predictability** of S

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

Target variable	Level of repair	Transformation	$\mu = \mu_{S\sharp}T_S$
$Z \sim \mu_B$	$\lambda \in [0,1]$	o.t.m. T_S	$R_S := T_S^{-1}(Z) \sim \mu_S$

Geometric repair (Feldman et al., 2015)



$$\tilde{\mu}_{S,\lambda} = \mathcal{L}(\lambda T_S(X) + (1-\lambda)X)$$

✗ The level of repair does not affect d_{TV}:

$= $ in some examples

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

Random repair (Gordaliza et al., 2019)

$B \sim \mathcal{B}(\lambda)$, independent of (X, S, Y)

$$\tilde{\mu}_{S,\lambda} = \mathcal{L}(BT_S(X) + (1-B)X)$$

✓ The level of repair controls d_{TV}:

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

✓ The new risk is a mixture of the two errors:

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

New criteria for statistical parity assessment

$$\varepsilon^* := \min_{g\in\mathcal{G}} BER(g, X, S) = \tfrac{1}{2}\left(1 - d_{TV}\left(\mu_0, \mu_1\right)\right),\ \mu_s = \mathcal{L}(X \mid S = s)$$

(Gordaliza et al., 2019)

Rejection of $H_0 :\ \rho(\mu_0, \mu_1) \geq \Delta_0 \Longrightarrow$ **Statistical certification that** $\mu_0 \approx \mu_1$

$H_0 : \mathcal{W}_p(\mu_0, \mu_1) \geq \Delta_0$ vs $H_a : \mathcal{W}_p(\mu_0, \mu_1) < \Delta_0$, for $\Delta_0 > 0$ and $p \geq 1$ ✓

Goal: CLT $\begin{cases} r_n\left(\mathcal{W}_p^p(\mu_{0,n}, \mu_1) - a_n\right) \\ r_{n,m}\left(\mathcal{W}_p^p(\mu_{0,n}, \mu_{1,m}) - a_{n,m}\right) \end{cases}$ in the case $\mu_0 \neq \mu_1$

Proposition (Central Limit Theorem for $\mathcal{W}_p$ on the real line with $p > 1$)

del Barrio et al., 2019b *Assume that $F, G \in \mathcal{F}_{2p}$ and G^{-1} is continuous on $(0, 1)$ and $p > 1$.*

+ Technical assumptions

(i) If $X_1, \ldots, X_n$ are i.i.d. F and F_n is the empirical d.f. based on the X_i's

$$\sqrt{n}(\mathcal{W}_p^p(F_n, G) - \mathcal{W}_p^p(F, G)) \to_w N(0, \sigma_p^2(F, G)).$$

(ii) If, furthermore, F^{-1} is continuous, $Y_1, \ldots, Y_m$ are i.i.d. G, independent of the X_i's, G_m is the empirical d.f. based on the Y_j's and $\frac{n}{n+m} \to \lambda \in (0, 1)$ then

$$\sqrt{\tfrac{nm}{n+m}}(\mathcal{W}_p^p(F_n, G_m) - \mathcal{W}_p^p(F, G)) \to_w N(0, (1-\lambda)\sigma_p^2(F, G) + \lambda\sigma_p^2(G, F)).$$

Role of the centering constants:

Kantorovich duality (Villani, 2003) $\Rightarrow \mathbb{E}(\mathcal{W}_p^p(F_n, G)) \geq \mathcal{W}_p^p(F, G)$

We can replace the centering constants in CLT provided:

$$0 \leq \sqrt{n}\Big(\mathbb{E}(\mathcal{W}_p^p(F_n, G)) - \mathcal{W}_p^p(F, G)\Big) \to 0$$

$$\Uparrow$$

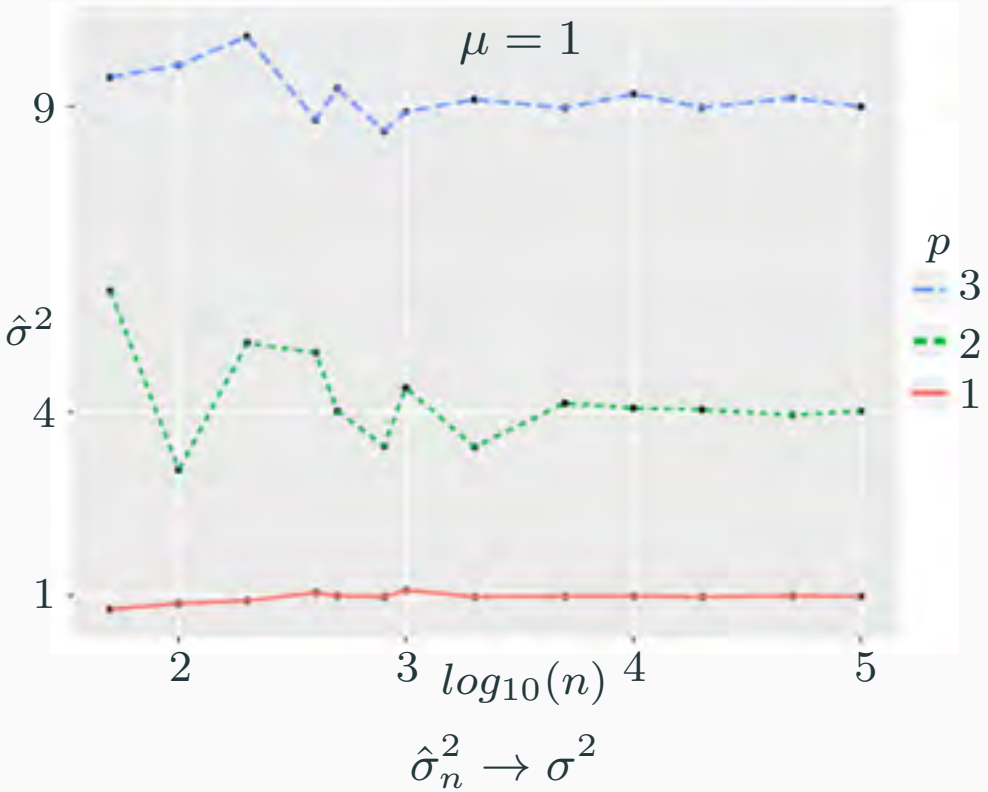
Sufficient conditions

Proposition (Consistency of variance estimates. del Barrio et al., 2019)

If $F, G \in \mathcal{F}_{2p}$, F^{-1}, G^{-1} *are continuous on* $(0,1)$ *and* $\frac{n}{n+m} \to \lambda \in (0,1)$, *then*

$$\hat{\sigma}^2_{n,m} = \frac{m}{n+m}\hat{\sigma}^2_{1,n,m} + \frac{n}{n+m}\hat{\sigma}^2_{2,n,m} \to (1-\lambda)\sigma^2_p(F,G) + \lambda\sigma^2_p(G,F) \text{ a.s.}$$

Example$(n = m)$: $F \sim N(0,1), G \sim N(\mu,1) \Rightarrow \sigma^2_p(F,G) = \sigma^2_p(G,F) = p^2\mu^{2p-2}$



$$\hat{\sigma}^2_n \to \sigma^2$$

n	$p=1$	$p=2$	$p=3$
50	0.03076	2.28517	79.70453
100	0.01434	1.25248	36.57057
200	0.00634	0.74908	15.10497
400	0.00290	0.32747	6.15403
500	0.00237	0.21351	5.50914
800	0.00148	0.18638	3.20970
1,000	0.00112	0.13431	2.59728
2,000	0.00054	0.0711	1.41032
5,000	0.00021	0.0304	0.52269
10,000	0.00011	0.0145	0.24127
σ^2	1	4	9

$$MSE = \frac{1}{N}\sum_{j=1}^{N}\left|\hat{\sigma}^2_j - \sigma^2\right|^2, N = 1000$$

Finite performance of the test. Simulations.

Example: Normal location model ($n = m$)

$$F \sim N(0,1), G \sim N(\mu, 1)$$

$$H_0 : \mathcal{W}_p(F,G) \geq \Delta_0,$$

vs

$$H_a : \mathcal{W}_p(F,G) < \Delta_0$$

Asymptotic level $\alpha = 0.05$

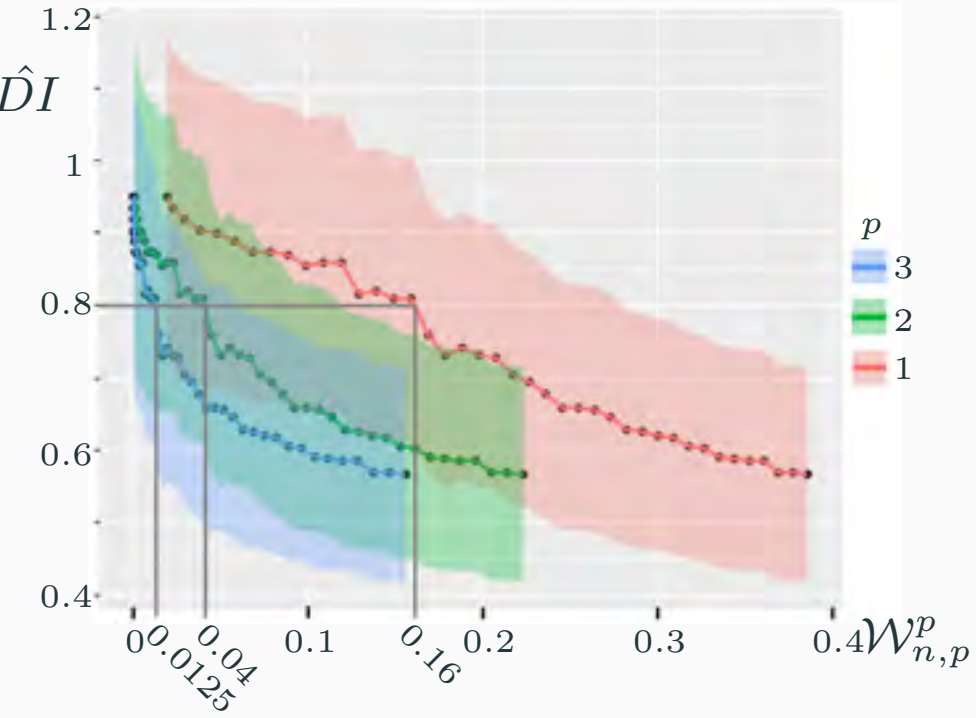
Simulations:

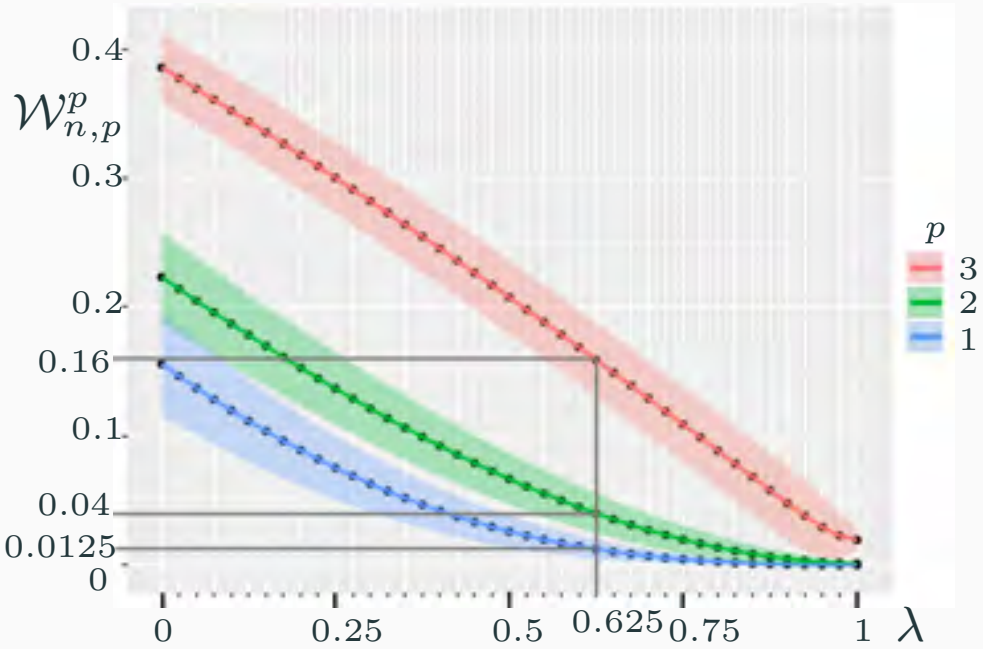
$$\Delta_0 = \mathcal{W}_p(N(0,1), N(1,1)) = 1$$

$\mu = 1 \rightarrow$ Level of the test

$\mu = 0.9, 0.7, 0.5 \rightarrow$ Power of the test

p	n	μ=1	μ=0.9	μ=0.7	μ=0.5
1	50	0.062	0.146	0.481	0.825
	100	0.055	0.193	0.698	0.974
	200	0.053	0.275	0.918	1
	400	0.051	0.413	0.995	1
	500	0.051	0.481	0.999	1
	800	0.052	0.64	1	1
	1,000	0.054	0.728	1	1
	2,000	0.047	0.937	1	1
2	50	0.074	0.167	0.513	0.839
	100	0.063	0.198	0.717	0.979
	200	0.059	0.272	0.927	1
	400	0.055	0.422	0.995	1
	500	0.05	0.484	0.999	1
	800	0.053	0.651	1	1
	1,000	0.053	0.736	1	1
	2,000	0.051	0.935	1	1
3	50	0.071	0.154	0.515	0.822
	100	0.066	0.206	0.715	0.973
	200	0.057	0.266	0.925	1
	400	0.052	0.422	0.992	1
	500	0.057	0.497	0.997	1
	800	0.053	0.652	1	1
	1,000	0.053	0.733	1	1
	2,000	0.051	0.937	1	1





DI and BER depend on a given classifier...

while $\mathcal{W}_p$ is a global condition on the fairness of the dataset

References

- E. DEL BARRIO, P. GORDALIZA, H. LESCORNEL AND J.-M. LOUBES. Central Limit Theorem and bootstrap procedure for Wasserstein's variations with application to structural relationships between distributions *JMVA*, 2019.
- P. GORDALIZA, E. DEL BARRIO, F. GAMBOA AND J.-M. LOUBES. Obtaining Fairness using Optimal Transport Theory. *Proceedings of the 36th International Conference on Machine Learning,* 2019.
- E. DEL BARRIO, P. GORDALIZA AND J.-M. LOUBES. A central limit theorem on the real line with application to fairness assessment in machine learning. *Information and Inference: a journal of the IMA*, 2019.
- P. BESSE, E. DEL BARRIO, P. GORDALIZA, J.-M. LOUBES. AND L. RISSER. A survey of bias in Machine Learning through the prism of Statistical Parity. *The American Statistician*, 2021.
- E. DEL BARRIO, P. GORDALIZA AND J.-M. LOUBES. Review of Mathematical Frameworks for fairness in Machine Learning. *Working paper*.

INSTITUT de MATHEMATIQUES de TOULOUSE

imUVa

Thanks for your attention!