Classification and Regression under Statistical Parity

Solenne Gaucher [1]

Evgenii Chzhen [2]

Vincent Divol[3]

Nicolas Schreuder [4]

[1]CREST ENSAE, IPP

[2]CNRS, LMO, Université Paris-Saclay

[3]CEREMADE, Université Paris Dauphine PSL

[4]CNRS, LIGM, Université Gustave Eiffel

Table of contents

~~Introduction to statistical fairness~~

Setting

Observations: $(\underbrace{X}_{\text{feature}}, \underbrace{S}_{\text{sensitive attribute}}, \underbrace{Y}_{\text{outcome}}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

Setting	Regression
Outcome	$\mathcal{Y} = \mathbb{R}$
Predictions	$\mathcal{F} \triangleq \{f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}\}$
Risk	$\mathcal{R}^{sq} = \mathbb{E}\left[(Y - f(X,S))^2\right]$
Fairness criteria	$f(X,S) \perp\!\!\!\perp S$ (Statistical Parity)

Setting

Observations: $(\underbrace{X}_{\text{feature}}, \underbrace{S}_{\text{sensitive attribute}}, \underbrace{Y}_{\text{outcome}}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

Setting	Regression	Classification
Outcome	$\mathcal{Y} = \mathbb{R}$	$\mathcal{Y} = \{0, 1\}$
Predictions	$\mathcal{F} \triangleq \{f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}\}$	$\mathcal{G} \triangleq \{g : \mathcal{X} \times \mathcal{S} \to \{0, 1\}\}$
Risk	$\mathcal{R}^{sq} = \mathbb{E}\left[(Y - f(X, S))^2\right]$	$\mathcal{R}^{0-1} = \mathbb{E}\left[Y \neq g(X, S)\right]$
Fairness criteria	$f(X, S) \perp\!\!\!\perp S$ (Statistical Parity)	$\mathbb{E}[g(X, S)\|S] \perp\!\!\!\perp S$ (Demographic Parity)

Setting

Observations: $(\underbrace{X}_{\text{feature}}, \underbrace{S}_{\text{sensitive attribute}}, \underbrace{Y}_{\text{outcome}}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

Setting	Regression	Classification
Outcome	$\mathcal{Y} = \mathbb{R}$	$\mathcal{Y} = \{0, 1\}$
Predictions	$\mathcal{F} \triangleq \{f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}\}$	$\mathcal{G} \triangleq \{g : \mathcal{X} \times \mathcal{S} \to \{0, 1\}\}$
Risk	$\mathcal{R}^{sq} = \mathbb{E}\left[(Y - f(X, S))^2\right]$	$\mathcal{R}^{0-1} = \mathbb{E}\left[Y \neq g(X, S)\right]$
Fairness criteria	$f(X, S) \perp\!\!\!\perp S$ (Statistical Parity)	$\mathbb{E}[g(X, S)\|S] \perp\!\!\!\perp S$ (Demographic Parity)

This is the awareness framework... in the unawareness framework, f and g cannot depend on S.

Setting

Observations: $(\underbrace{X}_{\text{feature}}, \underbrace{S}_{\text{sensitive attribute}}, \underbrace{Y}_{\text{outcome}}) \sim \mathbb{P}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

Setting	Regression	Classification
Outcome	$\mathcal{Y} = \mathbb{R}$	$\mathcal{Y} = \{0, 1\}$
Predictions	$\mathcal{F} \triangleq \{f : \mathcal{Z} \to \mathbb{R}\}$	$\mathcal{G} \triangleq \{g : \mathcal{Z} \to \{0, 1\}\}$
Risk	$\mathcal{R}^{sq} = \mathbb{E}\left[(Y - f(Z))^2\right]$	$\mathcal{R}^{0-1} = \mathbb{E}\left[Y \neq g(Z)\right]$
Fairness criteria	$f(Z) \perp\!\!\!\perp S$ (Statistical Parity)	$\mathbb{E}[g(Z)\|S] \perp\!\!\!\perp S$ (Demographic Parity)

Awareness framework : $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$, $Z = (X, S)$

Unawareness framework : $\mathcal{Z} = \mathcal{X}$, $Z = X$

Reminders on unconstrained classification

Without fairness constraint, the *Bayes classifier* solution to

$$\underset{g \in \mathcal{G}}{\text{minimize}} \quad \mathbb{P}\left[Y \neq g(Z)\right]$$

is given by

$$g^{Bayes}(z) = \mathbb{1}\left\{\eta(z) \geq \frac{1}{2}\right\}$$

where $\eta(Z) \triangleq \mathbb{E}\left[Y|Z\right]$ is the solution to

$$\underset{f \in \mathcal{F}}{\text{minimize}} \quad \mathbb{E}\left[(Y - f(Z))^2\right].$$

This relationship between classification and regression can be used to design and study classifiers [Yang, 1999], [Massart and Nédélec, 2006], [Audibert and Tsybakov, 2007], [Biau et al., 2008]

Reminders on unconstrained classification

Consider the risks

$$\mathcal{R}_\tau(g) = \tau \mathbb{P}\left[Y = 0, g(Z) = 1\right] + (1-\tau)\mathbb{P}\left[Y = 1, g(Z) = 0\right].$$

We have

$$\mathcal{R}_\tau(g) = (1-\tau)\mathbb{E}[Y] + \mathbb{E}\left[g(Z)(\tau - \eta(Z))\right].$$

$\Longrightarrow$ The Bayes classifier g_τ^{Bayes} is given by

$$g_\tau^{Bayes}(z) = \mathbb{1}\left\{\eta(z) \geq \tau\right\}.$$

There is an equivalence between solving the regression problem and solving the classification problem for all τ.

What about fair classification?

Let f^* be the solution to

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\left[(Y - f(Z))^2\right] \\ \text{such that} \quad & f(Z) \perp\!\!\!\perp S. \end{aligned}$$

We can define $g : z \mapsto \mathbb{1}\{f^*(z) \geq \tau\}$.

- Is g optimal for some threshold τ?
- Does g verify Demographic Parity?

What about fair classification?

Let f^* be the solution to

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\left[(Y - f(Z))^2\right] \\ \text{such that} \quad & f(Z) \perp\!\!\!\perp S. \end{aligned}$$

We can define $g : z \mapsto \mathbb{1}\{f^*(z) \geq \tau\}$.

- Is g optimal for some threshold τ?
- Does g verify Demographic Parity? Yes... it verifies **Strong Demographic Parity**

What about fair classification?

Let f^* be the solution to

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\left[(Y - f(Z))^2\right] \\ \text{such that} \quad & f(Z) \perp\!\!\!\perp S. \end{aligned}$$

We can define $g : z \mapsto \mathbb{1}\{f^*(z) \geq \tau\}$.

- Is g optimal for some threshold τ?
- Does g verify Demographic Parity? Yes... it verifies **Strong Demographic Parity**

Strong Demographic parity [JPSJC19]: A classifier g verifies Strong Demographic parity if $g(z) = \mathbb{1}\{f(z) \geq \tau\}$ for some threshold τ, and f verifies Statistical Parity.

What about fair classification?

Let f^* be the solution to

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\left[(Y - f(Z))^2\right] \\ \text{such that} \quad & f(Z) \perp\!\!\!\perp S. \end{aligned}$$

We can define $g : z \mapsto \mathbb{1}\{f^*(z) \geq \tau\}$.

- Is g optimal for some threshold τ? It depends...
- Does g verify Demographic Parity? Yes... it verifies **Strong Demographic Parity**

Strong Demographic parity [JPSJC19]: A classifier g verifies Strong Demographic parity if $g(z) = \mathbb{1}\{f(z) \geq \tau\}$ for some threshold τ, and f verifies Statistical Parity.

The awareness framework

Fair regression

Reminders from Nicolas's talk

Assumption (A1)

μ_s = Law$(\eta(X,S)|S=s)$ are continuous and have finite second moments.

Theorem (Chzhen et al., 2020, Le Gouic et al., 2020)

Let f^* be the solution to

$$\begin{aligned} \textit{minimize} \quad & \mathbb{E}\left[(Y - f(X,S))^2\right] \\ \textit{such that} \quad & f(X,S) \perp\!\!\!\perp S. \end{aligned}$$

Under Assumption (A1),

$$f^*(x,s) = F_{\overline{\mu}}^{-1} \odot F_{\mu_s}(\eta(x,s))$$

where $F_{\mu_s}(t) = \mathbb{P}\left[\eta(X,S) \leq t | S = s\right]$, and $F_{\overline{\mu}}^{-1}(\epsilon) = \sum_s p_s F_{\mu_s}^{-1}(\epsilon)$.

Fair classification

First results

Assumption (A2)

μ_S = Law$(\eta(X,S)|S=s)$ are continuous and supported on an interval.

Theorem (G., Schreuder and Chzhen, 2023)

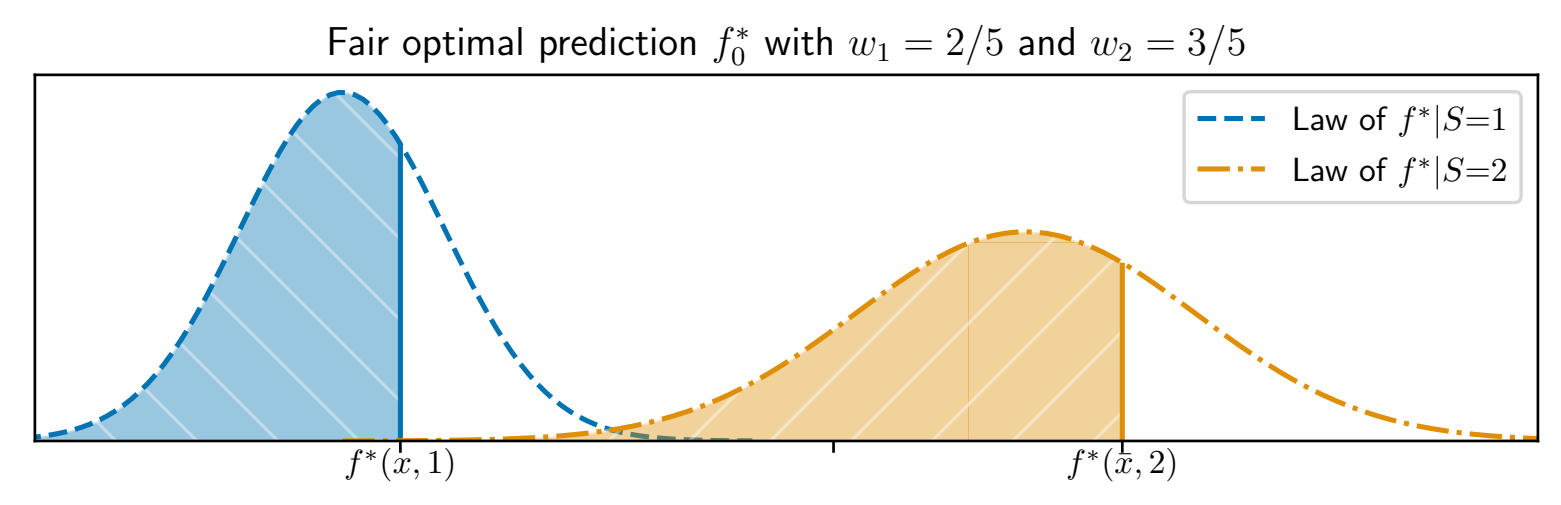
Let g^* be the solution to

$$\begin{aligned} &\textit{minimize} && \mathbb{P}\left[Y \neq g(X,S)\right] \\ &\textit{such that} && g(X,S) \perp\!\!\!\perp S. \end{aligned}$$

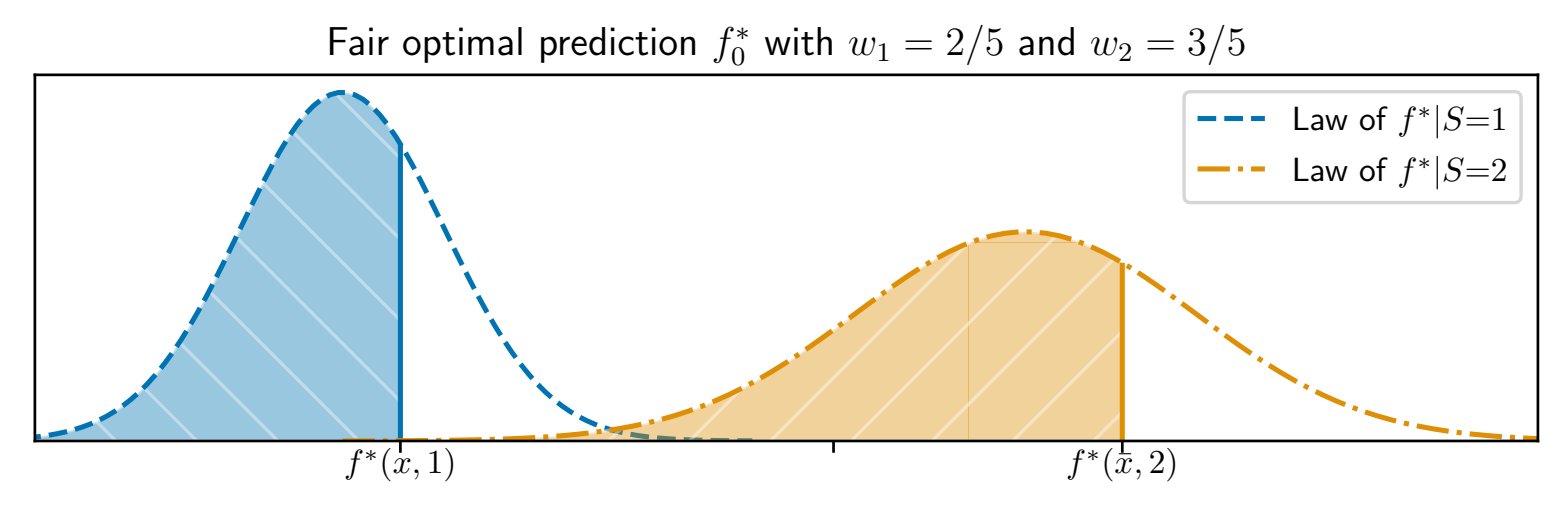
Under Assumption (A2), g^* can be expressed as

$$g^*(x,s) = \mathbb{1}\left\{f^*(x,s) \geq \frac{1}{2}\right\}.$$

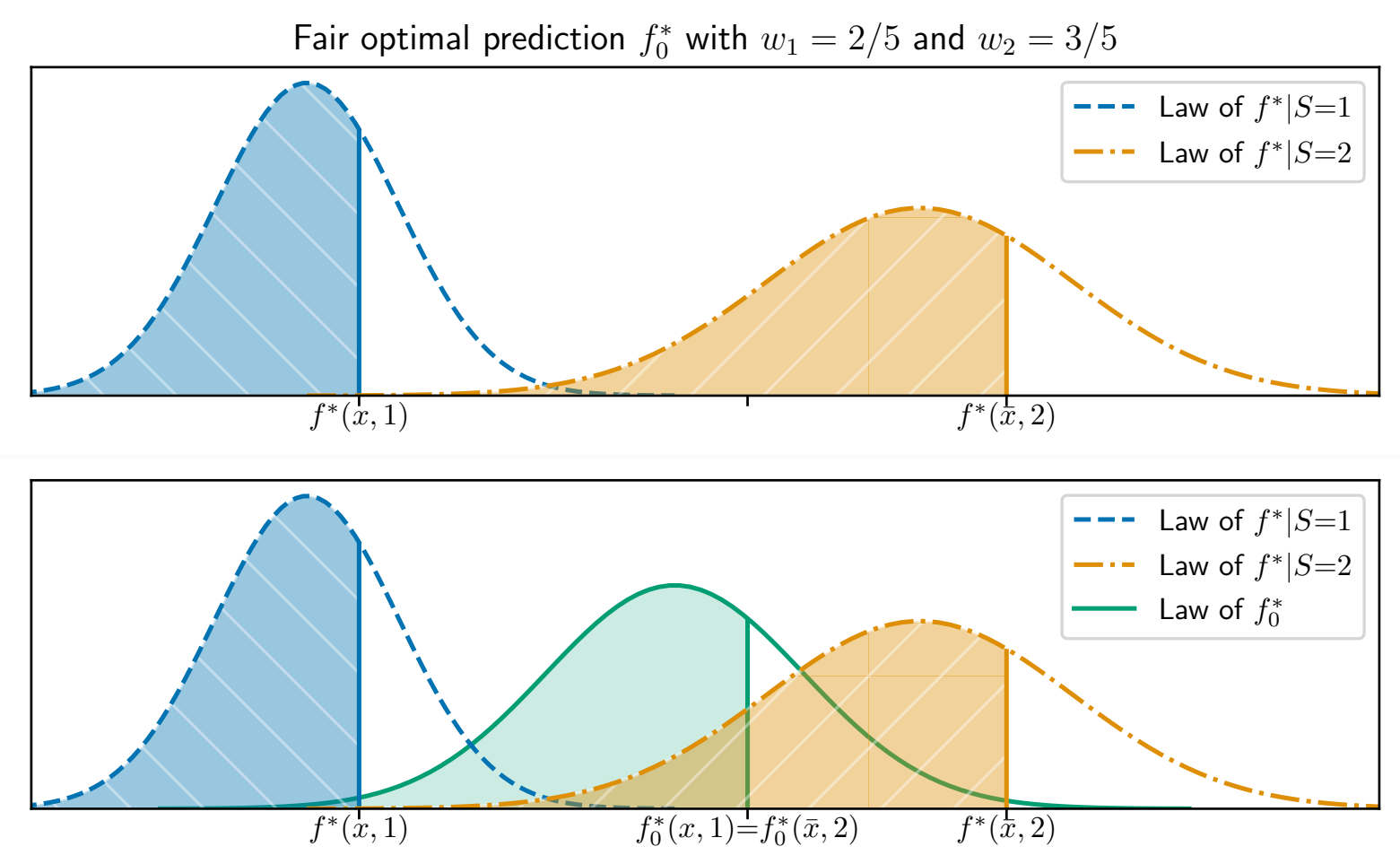
We look for classifiers $g(x, s) = \mathbb{1}\{\eta(x, s) \geq \kappa_s\}$.



We look for classifiers $g(x, s) = \mathbb{1}\{\eta(x, s) \geq \kappa_s\}$.



We look for classifiers $g(x, s) = \mathbb{1}\{\eta(x, s) \geq \kappa_s\}$.



General performance measures

We consider general performance measures

$$\mathcal{U}_{n,d}(g) \triangleq \frac{n_0 + n_1 \mathbb{P}[g(X,S)=1, Y=1] + n_2 \mathbb{P}[g(X,S)=1]}{d_0 + d_1 \mathbb{P}[g(X,S)=1, Y=1] + d_2 \mathbb{P}[g(X,S)=1]}$$

where n, d can depend on $\mathbb{P}[Y=1]$ (but not on g).

- Accuracy $\mathcal{U}_{n,d}(g) = -\mathbb{P}[g(X,S) \neq Y]$.
- F_b-score $\mathcal{U}_{n,d}(g) = \frac{(1+b^2)\mathbb{P}[g(X,S)=1, Y=1]}{b^2 \mathbb{P}[Y=1] + \mathbb{P}[g(X,S)=1]}$.
- Jaccard Index $\mathcal{U}_{n,d}(g) = \frac{\mathbb{P}[g(X,S)=1, Y=1]}{\mathbb{P}[g(X,S)=1, Y=0] + \mathbb{P}[g(X,S)=1]}$.
- AM Measure $\mathcal{U}_{n,d}(g) = \frac{1}{2}\left(\mathbb{P}[g(X,S)=1|Y=1] + \mathbb{P}[g(X,S)=0|Y=0]\right)$.

Results

Theorem (G., Schreuder and Chzhen, 2023 - Informal)

Let $g^*_{\mathsf{n},\mathsf{d}}$ *be the solution to*

$$\begin{aligned} &\textit{maximize} && \mathcal{U}_{\mathsf{n},\mathsf{d}}(g) \\ &\textit{such that} && g(X,S) \perp\!\!\!\perp S. \end{aligned}$$

Under Assumption (A2) and assumptions on n *and* d, $g^*_{\mathsf{n},\mathsf{d}}$ *can be expressed as*

$$g^*_{\mathsf{n},\mathsf{d}}(x,s) = \mathbb{1}\left\{f^*(x,s) \geq \theta^*_{\mathsf{n},\mathsf{d}}\right\}.$$

Assumption (A3)

1. $d_0 + (d_1 + (d_2)_+)_+ \geq 0$
2. $\begin{cases} d_2 n_1 > n_2 d_1 \textit{ and } d_0 n_1 - n_0 d_1 \geq (n_0 d_2 - d_0 n_2)_+ \\ \frac{n_0 d_2 - d_0 n_2}{n_2 d_1 - d_2 n_1} \leq \mathbb{P}[Y = 1] \end{cases}$
3. $\begin{cases} d_2 n_1 = n_2 d_1 \textit{ and } n_1 d_0 > d_1 n_0 \\ \frac{d_0 n_2 - n_0 d_2}{n_0 d_1 - d_0 n_1} \in [0, 1] \end{cases}$

Theorem (G., Schreuder and Chzhen, 2023)

Under (A2) + (A3)(1.) + (A3)(2. or 3.), $g^*_{n,d}(x, s) = \mathbb{1}\left\{f^*(x, s) \geq \theta^*_{n,d}\right\}$, *where*

- *if (A3)(2.) holds,* $\theta^*_{n,d}$ *solves*

$$\mathbb{E}\left[(f^*(X, S) - \theta)_+\right] = \theta \cdot \left\{\frac{n_0 d_1 - d_0 n_1}{n_2 d_1 - d_2 n_1}\right\} + \left\{\frac{n_0 d_2 - d_0 n_2}{n_2 d_1 - d_2 n_1}\right\}$$

- *if (A3)(3.) holds,* $\theta^*_{(n,d)} = \frac{d_0 n_2 - n_0 d_2}{n_0 d_1 - d_0 n_1}$.

Fair learning for classification

Algorithm:

1. Estimate η
2. Estimate f^*
3. Estimate n and d
4. Estimate threshold $\theta^*_{n,d}$ (explicit formula or fixed-point equation)
5. Use double-plug-in estimator

$$\widehat{g}_{n,d}(x,s) = \mathbb{1}\left\{\widehat{f}^*(x,s) \geq \widehat{\theta}_{\widehat{n},\widehat{d}}\right\}$$

Summary

In the awareness framework, DP-fair optimal classification with performance measure $\mathcal{U}_{n,d}$

- is given by $g^*_{n,d}(x,s) = \mathbb{1}\left\{f^*(x,s) \geq \theta^*_{n,d}\right\}$
- has desirable properties :
  - **does no harm to the protected group**
  - preserves **rational ordering**
  - preserves **monotonicity**

Summary

In the awareness framework, DP-fair optimal classification with performance measure $\mathcal{U}_{n,d}$

- is given by $g^*_{n,d}(x,s) = \mathbb{1}\left\{f^*(x,s) \geq \theta^*_{n,d}\right\}$
- has desirable properties :
  - **does no harm to the protected group**
  - preserves **rational ordering**
  - preserves **monotonicity**

(Informal) A family of classifiers g^*_θ preserves **monotonicity** if μ-almost-surely,
$\{\mathbb{P}[g^*_\theta(X,S)=1] > \mathbb{P}[g^*_{\theta'}(X,S)=1]\} \implies \{g^*_\theta(x,s) \geq g^*_{\theta'}(x,s)\}$.

The unawareness framework

Fair regression

Fair regression under unawareness

Problem: We want to solve

$$\begin{aligned} &\text{minimize} && \mathbb{E}\left[(Y - f(X))^2\right] \\ &\text{such that} && f(X) \perp\!\!\!\perp S. \end{aligned}$$

In the following, we assume $\mathcal{S} = \{1, 2\}$.

Notations: $X \sim \mu$, $X|S = 1 \sim \mu_1$, $X|S = 2 \sim \mu_2$.

An equivalent fairness constraint

Jordan decomposition We write $\mu_1 - \mu_2 = \quad (\mu_+ - \mu_-)$, where

- μ_+ and μ_- are positive measures
- $\mathcal{X}_+ \triangleq \text{supp}(\mu_+)$ and $\mathcal{X}_- \triangleq \text{supp}(\mu_-)$ are disjoint.

An equivalent fairness constraint

Jordan decomposition We write $\mu_1 - \mu_2 = M(\mu_+ - \mu_-)$, where

- μ_+ and μ_- are probability measures
- $\mathcal{X}_+ \triangleq \text{supp}(\mu_+)$ and $\mathcal{X}_- \triangleq \text{supp}(\mu_-)$ are disjoint.

An equivalent fairness constraint

Jordan decomposition We write $\mu_1 - \mu_2 = M(\mu_+ - \mu_-)$, where

- μ_+ and μ_- are probability measures
- $\mathcal{X}_+ \triangleq \text{supp}(\mu_+)$ and $\mathcal{X}_- \triangleq \text{supp}(\mu_-)$ are disjoint.

Lemma (Chzhen and Schreuder, 2020)

A function $f : \mathcal{X} \to \mathbb{R}$ verifies Statistical Parity if and only if

$$f\sharp\mu_+ = f\sharp\mu_-.$$

An equivalent fairness constraint

Jordan decomposition We write $\mu_1 - \mu_2 = M(\mu_+ - \mu_-)$, where

- μ_+ and μ_- are probability measures
- $\mathcal{X}_+ \triangleq \text{supp}(\mu_+)$ and $\mathcal{X}_- \triangleq \text{supp}(\mu_-)$ are disjoint.

Lemma (Chzhen and Schreuder, 2020)

A function $f : \mathcal{X} \to \mathbb{R}$ verifies Statistical Parity if and only if

$$f\sharp\mu_+ = f\sharp\mu_-.$$

Consequence We can look for functions

$$f(x) = \begin{cases} f_+(x) & \text{if} \quad x \in \mathcal{X}_+ \\ f_-(x) & \text{if} \quad x \in \mathcal{X}_- \\ \eta(x) & \text{if} \quad x \in \mathcal{X}_= \triangleq \{x : \mu_1(x) = \mu_2(x)\}. \end{cases}$$

such that $f_+\sharp\mu_+ = f_-\sharp\mu_-$.

$$\mathcal{R}^{sq}(f) = \int_{\mathcal{X}_+} c\Big(\big(\eta(x), \Delta(x)\big), f_+(x)\Big) d\mu_+(x) + \int_{\mathcal{X}_-} c\Big(\big(\eta(x), \Delta(x)\big), f_-(x)\Big) d\mu_-(x) + cste$$

where $c : \left(\mathbb{R} \times \mathbb{R}_+^*\right) \times \mathbb{R} \to \mathbb{R}_+$ is a cost, and $\Delta(x) = \begin{cases} \frac{d\mu_+}{d\mu}(x) \text{ if } x \in \mathcal{X}_+ \\ \frac{d\mu_-}{d\mu}(x) \text{ if } x \in \mathcal{X}_- \end{cases}$.

.

$$\mathcal{R}^{sq}(f) = \int_{\mathcal{X}_+} c\Big(\big(\eta(x), \Delta(x)\big), f_+(x)\Big) d\mu_+(x) + \int_{\mathcal{X}_-} c\Big(\big(\eta(x), \Delta(x)\big), f_-(x)\Big) d\mu_-(x) + cste$$

where $c : \left(\mathbb{R} \times \mathbb{R}_+^*\right) \times \mathbb{R} \to \mathbb{R}_+$ is a cost, and $\Delta(x) = \begin{cases} \frac{d\mu_+}{d\mu}(x) \text{ if } x \in \mathcal{X}_+ \\ \frac{d\mu_-}{d\mu}(x) \text{ if } x \in \mathcal{X}_- \end{cases}$.

- $f_\pm$ should depend only on $\big(\eta(x), \Delta(x)\big)$: $f_\pm(x) = \tilde{f}_\pm\big(\eta(x), \Delta(x)\big)$.

$$\mathcal{R}^{sq}(f) = \int_{\mathcal{X}_+} c\Big(\big(\eta(x), \Delta(x)\big), f_+(x)\Big) d\mu_+(x) + \int_{\mathcal{X}_-} c\Big(\big(\eta(x), \Delta(x)\big), f_-(x)\Big) d\mu_-(x) + cste$$

where $c : \left(\mathbb{R} \times \mathbb{R}_+^*\right) \times \mathbb{R} \to \mathbb{R}_+$ is a cost, and $\Delta(x) = \begin{cases} \frac{d\mu_+}{d\mu}(x) \text{ if } x \in \mathcal{X}_+ \\ \frac{d\mu_-}{d\mu}(x) \text{ if } x \in \mathcal{X}_- \end{cases}$.

- $f_\pm$ should depend only on $\Big(\eta(x), \Delta(x)\Big)$: $f_\pm(x) = \tilde{f}_\pm\Big(\eta(x), \Delta(x)\Big)$.
- Let $\tilde{\mu}_\pm = \text{Law}\Big((\eta(X), \Delta(X)) \,\Big|\, X \sim \mu_\pm\Big)$. $\tilde{f}_\pm$ should solve:

$$\text{minimize} \int c\Big((\eta, \Delta), \tilde{f}_+(\eta, \Delta)\Big) d\tilde{\mu}_+(\eta, \Delta) + \int c\Big((\eta, \Delta), \tilde{f}_-(\eta, \Delta)\Big) d\tilde{\mu}_-(\eta, \Delta)$$

$$\text{such that} \quad \tilde{f}_+ \sharp \tilde{\mu}_+ = \tilde{f}_- \sharp \tilde{\mu}_-$$

- Let $\tilde{\mu}_\pm = \text{Law}\Big((\eta(X), \Delta(X)) \;\Big|\; X \sim \mu_\pm\Big)$. $\tilde{f}_\pm$ should solve:

$$\text{minimize} \int c\Big((\eta, \Delta), \tilde{f}_+(\eta, \Delta)\Big) d\tilde{\mu}_+(\eta, \Delta) + \int c\Big((\eta, \Delta), \tilde{f}_-(\eta, \Delta)\Big) d\tilde{\mu}_-(\eta, \Delta)$$

$$\text{such that} \quad \tilde{f}_+ \sharp \tilde{\mu}_+ = \tilde{f}_- \sharp \tilde{\mu}_-$$

- Let $\tilde{\mu}_\pm = \text{Law}\Big((\eta(X), \Delta(X)) \,\Big|\, X \sim \mu_\pm\Big)$. $\tilde{f}_\pm$ should solve:

$$\text{minimize} \int c\Big((\eta, \Delta), \tilde{f}_+(\eta, \Delta)\Big) d\tilde{\mu}_+(\eta, \Delta) + \int c\Big((\eta, \Delta), \tilde{f}_-(\eta, \Delta)\Big) d\tilde{\mu}_-(\eta, \Delta)$$

$$\text{such that} \quad \tilde{f}_+ \sharp \tilde{\mu}_+ = \tilde{f}_- \sharp \tilde{\mu}_-$$

- $\tilde{f}_+ \sharp \tilde{\mu}_+ = \tilde{f}_- \sharp \tilde{\mu}_-$ should solve the barycenter problem:

$$\min_{\nu} OT_c(\tilde{\mu}_+, \nu) + OT_c(\tilde{\mu}_-, \nu)$$

where

$$OT_c(\tilde{\mu}, \nu) = \inf_{\gamma \in \Pi(\tilde{\mu}, \nu)} \int c((\eta, \Delta), y) d\gamma((\eta, \Delta), y).$$

Result

Assumption (A4)

The measures $\tilde{\mu}_+$ and $\tilde{\mu}_-$ are continuous, and the interior of their support has measure 1.

Theorem (Divol and G., 2024)

Under Assumptions (A1) and (A4), the solution ν^{bar} of the barycenter problem

$$\min_{\nu} OT_c(\tilde{\mu}_+, \nu) + OT_c(\tilde{\mu}_-, \nu)$$

is such that $OT_c(\tilde{\mu}_+, \nu^{bar})$ and $OT_c(\tilde{\mu}_-, \nu^{bar})$ are solved by transport maps $\tilde{f}_+^*$ and $\tilde{f}_-^*$. Moreover, the optimal fair prediction is given by

$$f(x) = \begin{cases} \tilde{f}_+^*(\eta(x), \Delta(x)) & \text{if} \quad x \in \mathcal{X}_+ \\ \tilde{f}_-^*(\eta(x), \Delta(x)) & \text{if} \quad x \in \mathcal{X}_- \\ \eta(x) & \text{else.} \end{cases}$$

Remarks:

- Under Assumptions (A1) and (A4), the optimal prediction is **deterministic**.
- The optimal prediction tries to guess the sensitive attribute.
- Unless η verifies Statistical Parity, the optimal fair prediction does not verify **rational ordering** within group.

Fair classification

Fair classification under unawareness

We want to solve

$$\begin{aligned} &\text{minimize} && \mathcal{R}_\tau(g) \triangleq \tau\mathbb{P}[Y=0, g(X)=1] + (1-\tau)\mathbb{P}[Y=1, g(X)=0] \\ &\text{such that} && g(X) \perp\!\!\!\perp S. \end{aligned}$$

Remark:

$$\mathcal{R}_\tau(g) = (1-\tau)\mathbb{E}[Y] + \mathbb{E}\left[g(X)(\tau - \eta(X))\right].$$

$\Longrightarrow$ The Bayes classifier g_τ^{Bayes} is given by

$$g_\tau^{Bayes}(x) = \mathbb{1}\left\{\eta(x) \geq \tau\right\}.$$

- Family of risks $\mathcal{R}_\tau$ with corresponding thresholds τ.

Lemma (Chzhen and Schreuder, 2020)

A function $f : \mathcal{X} \to \mathbb{R}$ verifies Statistical Parity if and only if

$$f\sharp\mu_+ = f\sharp\mu_-.$$

- g verifies Demographic Parity if and only if

$$\mathbb{E}_{\mu_+}\left[g(X)\right] = \mathbb{E}_{\mu_-}\left[g(X)\right].$$

Lemma (Chzhen and Schreuder, 2020)

A function $f : \mathcal{X} \to \mathbb{R}$ verifies Statistical Parity if and only if

$$f\sharp\mu_+ = f\sharp\mu_-.$$

- g verifies Demographic Parity if and only if

$$\mathbb{E}_{\mu_+}\left[g(X)\right] = \mathbb{E}_{\mu_-}\left[g(X)\right].$$

- We can look for $g(x) = \begin{cases} g_+(x) & \text{if} \quad x \in \mathcal{X}_+ \\ g_-(x) & \text{if} \quad x \in \mathcal{X}_- \\ \mathbb{1}\left\{\eta(x) \geq \tau\right\} & \text{else.} \end{cases}$

Lemma (Chzhen and Schreuder, 2020)

A function $f : \mathcal{X} \to \mathbb{R}$ verifies Statistical Parity if and only if

$$f\sharp\mu_+ = f\sharp\mu_-.$$

- g verifies Demographic Parity if and only if

$$\mathbb{E}_{\mu_+}\left[g(X)\right] = \mathbb{E}_{\mu_-}\left[g(X)\right].$$

- We can look for $g(x) = \begin{cases} g_+(x) & \text{if} \quad x \in \mathcal{X}_+ \\ g_-(x) & \text{if} \quad x \in \mathcal{X}_- \\ \mathbb{1}\left\{\eta(x) \geq \tau\right\} & \text{else.} \end{cases}$

Decomposition + change of measure:

$$\mathcal{R}_\tau(g) = \mathbb{E}_{\mu_+}\left[\left(\frac{\tau - \eta(X)}{\Delta(X)}\right) g_+(X)\right] + \mathbb{E}_{\mu_-}\left[\left(\frac{\tau - \eta(X)}{\Delta(X)}\right) g_-(X)\right] + cste.$$

We should choose $g_\pm(x) = \mathbb{1}\left\{\frac{\eta(X)-\tau}{\Delta(X)} \geq \kappa_\pm\right\}$.

Optimal fair classification under unawareness

Theorem (G., Schreuder and Chzhen, 2023)

Let g_τ^* be the classifier minimizing $\mathcal{R}_\tau$ under the Demographic parity constraint. Under Assumption (A4), g_τ^* can be expressed as

$$g_\tau^*(x) = \begin{cases} \mathbb{1}\left\{\frac{\eta(x)-\tau}{\Delta(x)} \geq \kappa(\tau)\right\} & \text{if} \quad x \in \mathcal{X}_+ \\ \mathbb{1}\left\{\frac{\eta(x)-\tau}{\Delta(x)} \geq -\kappa(\tau)\right\} & \text{if} \quad x \in \mathcal{X}_- \\ \mathbb{1}\left\{\eta(x) \geq \tau\right\} & \text{else} \end{cases}$$

where $\kappa(\tau)$ is such that

$$\mathbb{P}_{\mu_+}\left[\frac{\eta(X)-\tau}{\Delta(X)} \geq \kappa(\tau)\right] = \mathbb{E}_{\mu_-}\left[\frac{\eta(X)-\tau}{\Delta(X)} \geq -\kappa(\tau)\right].$$

Remarks: In the unawareness framework, DP-fair optimal classification with risk measure $\mathcal{R}_\tau$

- tries to guess the sensitive attribute;
- **harms** some individuals from **protected group**;
- does not preserves **rational ordering**

Remarks: In the unawareness framework, DP-fair optimal classification with risk measure $\mathcal{R}_\tau$

- tries to guess the sensitive attribute;
- **harms** some individuals from **protected group**;
- does not preserves **rational ordering**
- may not preserves **monotonicity**

Remarks: In the unawareness framework, DP-fair optimal classification with risk measure $\mathcal{R}_\tau$

- tries to guess the sensitive attribute;
- **harms** some individuals from **protected group**;
- does not preserves **rational ordering**
- may not preserves **monotonicity**

Monotonicity would imply that μ-almost surely,

$$\tau \geq \tau' \implies \begin{cases} \left\{ x \in \mathcal{X}_+ : \underbrace{\frac{\eta(x) - \tau}{\Delta(x)} \geq \kappa(\tau)}_{g^*_\tau(x) = 1} \right\} \subset \left\{ x \in \mathcal{X}_+ : \underbrace{\frac{\eta(x) - \tau'}{\Delta(x)} \geq \kappa(\tau')}_{g^*_{\tau'}(x) = 1} \right\} \\ \left\{ x \in \mathcal{X}_- : \underbrace{\frac{\eta(x) - \tau}{\Delta(x)} \geq -\kappa(\tau)}_{g^*_\tau(x) = 1} \right\} \subset \left\{ x \in \mathcal{X}_- : \underbrace{\frac{\eta(x) - \tau'}{\Delta(x)} \geq -\kappa(\tau')}_{g^*_{\tau'}(x) = 1} \right\}. \end{cases}$$

Relationship between classification and regression

Original question: Is $\mathbb{1}\{f^*(x) \geq \tau\}$ optimal?

- Without monotonicity, no! (The optimal classifier cannot be of the form $\mathbb{1}\{f(x) \geq \tau\}$.)
- **(Divol and G., 2024 - Informal)** With monotonicity, yes!
- Both behaviours can be observed.

Conclusion

In the awareness framework:

- Optimal fair classifier verifies desirable properties : does no harm to the protected group, preserves rational ordering, preserves monotonicity.
- For general performance measures, $g^*(x, s) = \mathbb{1}\{f^*(x, s) \geq \theta\}$

In the unawareness framework:

- Both the optimal fair classifier and the optimal fair regression function rely on guessing the sensitive attribute;
- Neither of them preserve rational ordering, fair classification harms indivual from the protected group.
- If monotonicity is verified, $g^*_\tau(x) = \mathbb{1}\{f^*(x) \geq \tau\}$
- If monotonicity is not verified, $g^*_\tau(x) \neq \mathbb{1}\{f^*(x) \geq \tau\}$.