Fair Regression with Wasserstein Barycenters

for Sorbonne Université

Eyal Cohen[1] Eline Pot[1]
[1] LPSM

Supervisor(s):
Christophe Denis[1], Rafael Pinot[1]

Table of Contents

$$Y = f^\star(X, S) + \xi \tag{1}$$

with $\xi \in \mathbb{R}$ a centered r.v.
$(X, S) \sim \mathbb{P}_{X,S}$ on $\mathbb{R}^d \times \mathcal{S}$, where S is the sensible attribute ; $|\mathcal{S}| < \infty$
$f^\star : \mathbb{R}^d \times \mathcal{S} \mapsto \mathbb{R}$ the regression function minimizing the squared risk.
$\forall f : \mathbb{R}^d \times \mathcal{S} \mapsto \mathbb{R}, \nu_{f|s}$ is the distribution of $f(X, S)|S = s$.

$$i.e. \quad F_{\nu_{f|s}}(t) = \mathbb{P}(f(X, S) \leqslant t | S = s) \tag{2}$$

Definition (Demographic Parity (DP))

$g : \mathbb{R}^d \times \mathcal{S} \mapsto \mathbb{R}$ is *fair* if :

$$\forall s, s' \in \mathcal{S}, \quad \sup_{t \in \mathbb{R}} \left|\mathbb{P}(g(X, S) \leqslant t | S = s) - \mathbb{P}(g(X, S) \leqslant t | S = s')\right| = 0$$

In particular, we note that if g is fair, then $\nu_{g|s} = \nu_g$

Table of Contents

Definition (Wasserstein-2 distance)

For all μ, ν univariate probabilities measures, we define the Wasserstein-2 distance :

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Gamma_{\mu,\nu}} \int |x - y|^2 d\gamma(x, y)$$

where $\Gamma_{\mu,\nu}$ is the set of coupling measures on $\mathbb{R} \times \mathbb{R}$

Theorem (Monge map)

Let μ and ν be two univariate measures such that ν has a density. And let $X \sim \nu$. Then there exists a mapping $T : \mathbb{R} \mapsto \mathbb{R}$ such that $\mathcal{W}_2^2(\mu, \nu) = \mathbb{E}(X - T(X))^2$.
i.e. $(X, T(X)) \sim \gamma^\star \in \Gamma_{\mu,\nu}$ with $\gamma^\star$ and optimal coupling measure.
Also, we have : $T = Q_\mu \circ F_\nu$

Theorem (Inverse of the cdf)

Let $\nu_1, \ldots, \nu_{|\mathcal{S}|}$ be $|\mathcal{S}|$ univariate probability measures admitting densities. Let $p_1, \ldots, p_{|\mathcal{S}|} \geqslant 0$ s.t. $\sum_{s=1}^{|\mathcal{S}|} p_s = 1$. Let us denote : $\nu^\star \in \arg\min_\nu \sum_{s=1}^{|\mathcal{S}|} p_s \mathcal{W}_2^2(\nu_s, \nu)$.
Then : $F_\nu^\star(\cdot) = \left(\sum_{s=1}^{|\mathcal{S}} p_s Q_{\nu_s}\right)^{\leftarrow}(\cdot)$.

Theorem (Characterization of fair optimal prediction)

Let us assume $\forall s \in \mathcal{S}$ that the univariate measure $\nu_{f^\star|s}$ has a density and let $p_s = \mathbb{P}(S = s)$. Then :

$$\min_{g \text{ is fair}} \mathbb{E}\left[\left(f^\star(X,S) - g(X,S)\right)^2\right] = \min_{\nu} \sum_{s \in \mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^\star|s}, \nu)$$

Moreover if $g^\star$ solves l.h.s and $v^\star$ the r.h.s, then $\nu^\star = \nu_g^\star$ and :

$$g^\star(x,s) = \left(\sum_{s' \in \mathcal{S}} p_{s'} Q_{f^\star|s'}\right) \bigcirc F_{f^\star|s}(f^\star(x,s))$$

Let us show that

$$\min_{gfair} \mathbb{E}\left[(f^\star(X,S) - g(X,S))^2\right] = \min_{\nu} \sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^\star|s}, \nu)$$

Let us denote $\bar{g}$ a minimizer of the l.h.s., and $\nu_{\bar{g}}$ its ditribution. As $\nu_{f^\star|s}$ admits a density, we can use the precedent theorem : for each $s \in \mathcal{S}$, there exists $T_s = Q_{\nu_{\bar{g}}} \circ F_{f^\star|s}$ such that with $\tilde{g} = T_s = Q_{\nu_{\bar{g}}} \circ F_{f^\star|s} \circ f^\star$:

$$\sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^\star|s}, \nu_{\bar{g}}) = \mathbb{E}\left[(f^\star(X,S) - \tilde{g}(X,S))^2\right] \tag{3}$$

We can write:

$$\mathbb{P}\left(\tilde{g}(X,S) \leqslant t\right) = \sum_{s\in\mathcal{S}} p_s \mathbb{P}\left(Q_{\nu_{\bar{g}}} \circ F_{f^\star|s} \circ f^\star(X,s) \leqslant t\right) \tag{4}$$

$$= \sum_{s\in\mathcal{S}} p_s \mathbb{P}\left(f^\star(X,s) \leqslant Q_{f^\star|s} \circ F_{\nu_{\bar{g}}}(t)\right)$$

$$= \sum_{s\in\mathcal{S}} p_s F_{f^\star(X,s)}\left(Q_{f^\star|s} \circ F_{\nu_{\bar{g}}}(t)\right)$$

$$= \sum_{s\in\mathcal{S}} p_s F_{\nu_{\bar{g}}}(t) \tag{5}$$

$$\mathbb{P}\left(\tilde{g}(X,S) \leqslant t\right) = \sum_{s\in\mathcal{S}} p_s F_{\nu_{\bar{g}}}(t) \tag{5}$$

And since $\bar{g}$ is fair by definition, this implies that $\tilde{g}$ is also fair.
We now define $T^\star = Q_{\nu^\star} \circ F_{f^\star|s}$ as the optimal transport map between $\nu_{f^\star|s}$ and $\nu^\star = argmin_\nu \left(\sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^\star|s}, \nu)\right)$, and we define $g^\star = T^\star \circ f^\star$.
By the result of the Monge map theorem, we get:

$$\min_\nu \sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^\star|s}, \nu) = \mathbb{E}\left[\left(f^\star(X,S) - g^\star(X,S)\right)^2\right]$$

$g^\star$ minimizing the LHS, we note that $\nu^\star$ is independent of s by construction, so $g^\star$ is DP. Thus we can bound the LHS by taking the minimum over the DP-fair set:

$$\min_\nu \sum_{s\in\mathcal{S}} p_s \mathcal{W}_2^2(\nu_{f^\star|s}, \nu) \geqslant \min_{\text{g fair}} \mathbb{E}\left[\left(f^\star(X,S) - g(X,S)\right)^2\right] \tag{6}$$

But, by optimality of $\bar{g}$, we have $\mathbb{E}\left[\left(f^\star(X,S) - \tilde{g}(X,S)\right)^2\right] \geqslant \mathbb{E}\left[\left(f^\star(X,S) - \tilde{g}(X,S)\right)^2\right]$.
But we also have that:

$$\mathcal{W}_2^2\left(\nu_{f^\star|s}, \nu_{\bar{g}}\right) \leqslant \mathbb{E}\left[\left(f^\star(X,S) - g(X,S)\right)^2 |S = s\right]$$

$$\Longrightarrow \min_\nu \sum_{s\in\mathcal{S}} \mathcal{W}_2^2\left(\nu_{f^\star|s}, \nu_{\bar{g}}\right) \leqslant \min_{g \text{ is fair}} \mathbb{E}\left[\left(f^\star(X,S) - g(X,S)\right)^2\right]$$

This gives the equality, we can then set $\bar{g} = g^\star$, as $g^\star$ is fair.

Given a base estimator $\hat{f}$ of $f^\star$, we define the final estimator $\hat{g}$ of $g^\star$ as :

$$\hat{g}(x,s) = \left(\sum_{s'\in\mathcal{S}} \hat{p}_{s'}\hat{Q}_{\hat{f}|s'}\right) \circ \hat{F}_{\hat{f}|s}(\hat{f}(x,s)+\varepsilon)$$

where $\varepsilon \sim \mathcal{U}([-\sigma,\sigma])$, independant from the others r.v.
With $\hat{p}_s$ the empirical frequency of $S=s$ evaluated on $\{S_i\}_{i=1\dots N} \overset{iid}{\sim} \mathbb{P}_S$
With $\hat{F}_{f|s}$ and $\hat{Q}_{f|s}$ the empirical CDF and quantile function of $(f(X,S)+\varepsilon)|S=s$ based on $\{f(X_i^s,S)+\varepsilon_{is}\}_{i\in\mathcal{I}_1^s}$ and $\{f(X_i^s,S)+\varepsilon_{is}\}_{i\in\mathcal{I}_0^s}$ respectively.
They are defined as :

$$\hat{F}_{f|s} = F_{\hat{\nu}^1_{f|s}} \quad \text{and} \quad \hat{Q}_{f|s} = Q_{\hat{\nu}^0_{f|s}}$$

Here, $\hat{\nu}^0_{f|s}$ and $\hat{\nu}^1_{f|s}$ are estimators of $\nu_{f|s}$. Fixing $\mathcal{I}_0^s$ and $\mathcal{I}_1^s$ as an equal partition of $[N_s]$, these estimators read as :

$$\hat{\nu}^0_{f|s} = \frac{1}{|\mathcal{I}_0^s|}\sum_{i\in\mathcal{I}_0^s}\delta(f(X_i^s,s)+\varepsilon_{is}-\cdot) \quad \text{and} \quad \hat{\nu}^1_{f|s} = \frac{1}{|\mathcal{I}_1^s|}\sum_{i\in\mathcal{I}_1^s}\delta(f(X_i^s,s)+\varepsilon_{is}-\cdot) \quad \text{with} \quad \varepsilon_{is} \overset{iid}{\sim} \mathcal{U}([-\sigma,\sigma])$$

Table of Contents

Theorem (Fairness guarantee)

For any $\mathbb{P}$ on (X, S, Y), for any $\hat{f}$ constructed on labeled data, and for any $s, s' \in \mathcal{S}$, we have fairness guarantee in expectation over the data :

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}(\hat{g}(X,S) \leqslant t | S = s) - \mathbb{P}(\hat{g}(X,S) \leqslant t | S = s')\right| \leqslant \frac{2}{\min(N_s, N_{s'}) + 2} \mathbb{1}_{N_s \neq N_{s'}}$$

We also have the bound over the expected (over data) violation of the fairness definition (with $\mathcal{D}$ the union of all the datasets):

$$\mathbb{E}\left[\sup_{t\in\mathbb{R}} \left|\mathbb{P}(\hat{g}(X,S) \leqslant t | S = s, \mathcal{D}) - \mathbb{P}(\hat{g}(X,S) \leqslant t | S = s', \mathcal{D})\right|\right] \leqslant \frac{6}{\sqrt{\min(N_s, N_{s'}) + 1}}$$

PROOF IDEAS: $\longrightarrow$ *For the first result, we use that*

$\forall s, s' \in \mathcal{S}, \sup_{t\in\mathbb{R}} |\mathbb{P}(\hat{g}(X^s, s) \leqslant t) - \mathbb{P}(\hat{g}(X^{s'}, s) \leqslant t)| \leqslant \sup_{t\in(0,1)} |\mathbb{P}(\hat{F}_{\hat{f}|s}(\hat{f}(X^s, s) + \varepsilon) \leqslant t) - \mathbb{P}(\hat{F}_{\hat{f}|s'}(\hat{f}(X^{s'}, s') + \varepsilon) \leqslant t)|$

Then, fixing $t \in (0,1)$ and $k_s(t) \in \{1, \dots, |\mathcal{I}_1^s|\}$ such that $t \in \left[\frac{k_s(t)-1}{|\mathcal{I}_1^s|}, \frac{k_s(t)}{|\mathcal{I}_1^s|}\right)$, and using the fact that conditionally on labeled data, the random variables

$\sum_{i\in\mathcal{I}_1^s} \mathbb{1}_{\hat{f}(X_i^s)+\varepsilon_{is} \leqslant \hat{f}(x,s)+\varepsilon} \sim \mathcal{U}(\{0, \dots, |\mathcal{I}_1^s|\})$, *we have:* $\mathbb{P}(\hat{F}_{\hat{f}|s}(\hat{f}(X^s, s) + \varepsilon) \leqslant t) = \frac{k_s(t)}{|\mathcal{I}_1^s|+1}$

We use the same argument for s' to get: $\sup_{t\in\mathbb{R}} \left|\mathbb{P}(\hat{g}(X^s, s) \leqslant t) - \mathbb{P}(\hat{g}(X^{s'}, s') \leqslant t)\right| \leqslant \sup_{t\in(0,1)} \left|\frac{k_s(t)}{(N_s/2+1)} - \frac{k_{s'}(t)}{(N_{s'}/2+1)}\right|$

$\longrightarrow$ *For the second result, the beginning of the demonstration is the same, then we use the triangle inequality to finally get the bound* $2\mathbb{E}\|F_{\bar{\nu}_{\hat{f}|s}} - \hat{F}_{\hat{f}|s}\|_\infty + 2\mathbb{E}\|F_{\bar{\nu}_{\hat{f}|s'}} - \hat{F}_{\hat{f}|s'}\|_\infty$, *where* $F_{\bar{\nu}_{\hat{f}|s}} = \mathbb{P}(\hat{f}(X^s, s) + \varepsilon \leqslant t | \mathcal{D})$. *Then, we conclude applying DKL inequality, conditionally on $\mathcal{L}$.*

Let us make the following assumptions :

- For any $s \in \mathcal{S}, \nu_{f^\star|s}$ admits a density q_s such that $0 < \underline{\lambda_s} \leqslant q_s \leqslant \bar{\lambda}_s$.
- There exist c and C positive and independent of $n, N, N_1, \ldots, N_{|\mathcal{S}|}$ and b_n a positive sequence such that for all $\delta > 0$, we have :

$$\mathbb{P}(|f^\star(x,s) - \hat{f}(x,s)| \geqslant \delta) \leqslant \exp(-Cb_n\delta^2)$$

Theorem (Estimation guarantee)

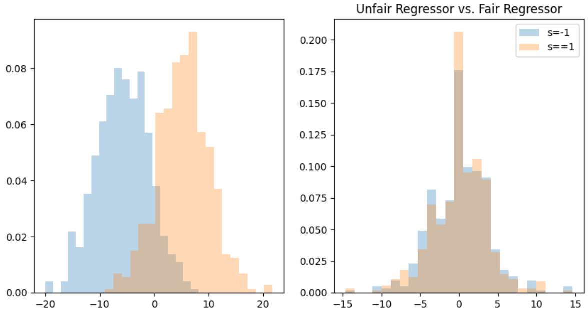
Under these assumptions, let us set $\sigma \lesssim \min_{s\in\mathcal{S}}\{\max(\frac{1}{\sqrt{N_s}}, \frac{1}{\sqrt{bn}})\}$, *then* $\hat{g}$ *verifies :*

$$\mathbb{E}[|g^\star(X,S) - \hat{g}(X,S)|] \lesssim \min\left\{\frac{1}{\sqrt{b_n}}; \sum_{s\in\mathcal{S}} p_s \frac{1}{\sqrt{N_s}}; \sqrt{\frac{|\mathcal{S}|}{N}}\right\}$$

Table of Contents

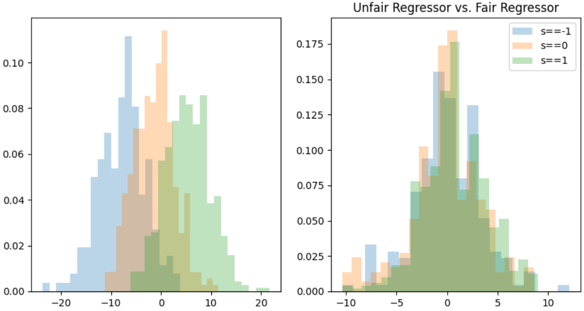
Gaussian distributions:

- Binary:



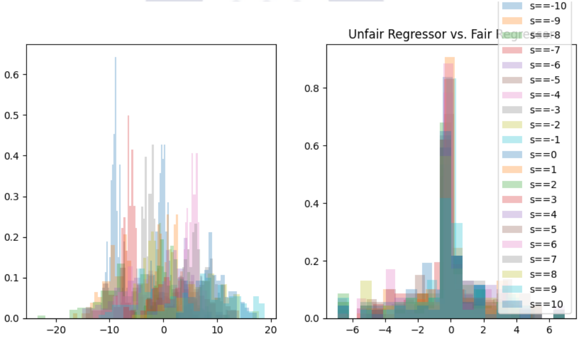
Gaussian distributions:

- Three values:



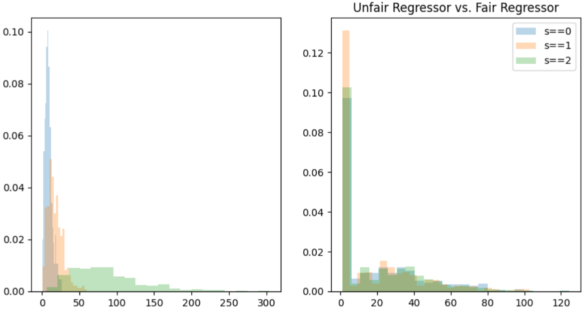
Gaussian distributions:

- More values (21):



Exponential distributions:

- Three values:



Fair Regression with Wasserstein Barycenters

Thank you for listening !
Any Questions ?