Review of ‘Fair Regression with Wasserstein Barycenters’

Chzhen et al., NeurIPS 2020

Presented by Kunwoong Kim

Department of Statistics, Seoul National University

September 30, 2025

- **Problem**: (Group-)fair regression

  We aim to find a function that minimizes the mean squared error under the demographic parity constraint.

- **Idea**: Alignment of predictions using Wasserstein barycenter.

- **Proposed method**: A post-processing algorithm for perfect fairness.

Notation

- Variables
  - $X \in \mathbb{R}^d$: an input random vector
  - $Y \in \mathbb{R}$: a real-valued output
  - $S \in \mathcal{S}$: a sensitive attribute (e.g., $\mathcal{S} = \{0, 1\}$)
- Distributions
  - $\mathbb{P}$: the joint distribution of (X, S, Y).
  - $\mathbb{P}_{X,S}$: the marginal distribution of (X, S).
- Cumulative Distribution Function (CDF)
  For a given probability measure μ, we denote F_μ as the CDF of μ.
- Quantile Function
  For a given probability measure μ, we denote $Q_\mu : [0, 1] \to \mathbb{R}$ as the quantile function of μ. That is, $Q_\mu(t) = \inf\{y \in \mathbb{R} : F_\mu(y) > t\}$ for $t \in (0, 1]$.

- A standard regression model:

$$Y = f(X, S) + \eta,$$

where $\eta \in \mathbb{R}$ is a centered random variable.

- Let f^* be the true regression function such that

$$f^*(x, s) = \mathbb{E}\left(Y | X = x, S = s\right).$$

- Given f, denote $\nu_{f|s}$ as the conditional distribution of $f(X, S)|S = s$.
The CDF of $\nu_{f|s}$ is given by

$$F_{\nu_{f|s}}(t) = \mathbb{P}(f(X, S) \leq t | S = s).$$

Definition 1 ((Strong) demographic parity)

A prediction model $g : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ is fair if, for every $s, s' \in \mathcal{S}$

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}(g(X, S) \leq t | S = s) - \mathbb{P}(g(X, S) \leq t | S = s') \right| = 0. \tag{1}$$

- Strong demographic parity defined in this paper requires the Kolmogorov-Smirnov distance to be zero for all s, s'.

Theorem 2

Let $p_s := \mathbb{P}(S = s)$. *Assume that* $\nu_{f^*|s}$ *has a density for each* $s \in \mathcal{S}$. *Then, we have*

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

Moreover, if g^* *and* ν^* *solve the left-hand-side and the right-hand-side of Equation* (2) *respectively, then* $\nu^* = \nu_{g^*}$ *and*

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

- Implication: We can obtain an optimal fair regression model by: sequentially doing (i) quantile matching and (ii) transforming to barycenter.

In other words, the optimal fair prediction model g^* is a transformation of f^* defined by

$$g^*(x,s) = p_s f^*(x,s) + (1-p_s)t^*(x,s),$$

where t^* is a correction so that the quantile of $f^*(X,s)$ is the same as the quantile of $f^*(X,s')$ for $s \neq s'$.

Fair optimal prediction g^* with $p_1 = 2/5$ and $p_2 = 3/5$

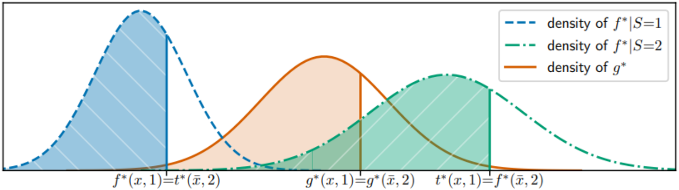

Figure 1: For a new point $(x,1)$, the value $t^*(x,1)$ is chosen such that the shaded Green Area (//) $= \mathbb{P}(f^*(X,S) \leq t^*(x,1)|S=2)$ equals to the shaded Blue Area (\\) $= \mathbb{P}(f^*(X,S) \leq f^*(x,1)|S=1)$. The final prediction $g^*(x,1)$ is a convex combination of $f^*(x,1)$ and $t^*(x,1)$. The same is done for $(\bar{x},2)$.

Main results

- Let $\mathcal{D}_n := \{(x_i, s_i, y_i)\}_{i=1}^n$ be a given dataset. Let $\mathcal{D}_n^s := \{(x_i, s_i, y_i) \in \mathcal{D}_n\}_{i:s_i=s}$ be a subset of $\mathcal{D}_n$ conditional on s and let $n_s = |\mathcal{D}_n^s|$.
- Let $\hat{F}_{f|s}$ and $\hat{Q}_{f|s}$ be the empirical CDF and empirical quantile function for a given f, respectively. Let $\hat{f}$ be a given prediction model (e.g., empirical risk minimizer) and define

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

  where $\epsilon \sim \text{Unif}([-\sigma, \sigma])$.
- Assume that (i) $\nu_{f^*|s}$ admits a bounded density for each $s \in \mathcal{S}$ and (ii) there exists a positive constant c and a sequence b_n such that $\mathbb{E}|f^*(X, S) - \hat{f}(X, S)| \leq c b_n^{-1/2}$.

Theorem 3

Set $\sigma \lesssim \min_{s \in \mathcal{S}} n_s^{-1/2} \wedge b_n^{-1/2}$. *Then, we have*

$$\mathbb{E}|g^*(X, S) - \hat{g}(X, S)| \lesssim b_n^{-1/2} \bigvee \left(\sum_{s \in \mathcal{S}} p_s n_s^{-1/2} \right) \bigvee \sqrt{\frac{|\mathcal{S}|}{n}}. \quad (3)$$

Experiments

- Performance measures
  - Prediction

$$\mathsf{MSE}(g) = \frac{1}{n} \sum_{(x_i, s_i, y_i) \in \mathcal{D}_n} (y_i - g(x_i, s_i))^2$$

  - Fairness

$$\mathsf{KS}(g) = \max_{s, s' \in \mathcal{S}} \sup_{t \in \mathbb{R}} \left| \frac{1}{n_s} \sum_{(x_i, s_i, y_i) \in \mathcal{D}_n^s} \mathbb{I}(g(x_i, s_i) \leq t) - \frac{1}{n_{s'}} \sum_{(x_i, s_i, y_i) \in \mathcal{D}_n^{s'}} \mathbb{I}(g(x_i, s_i) \leq t) \right| \quad (4)$$

Method	CRIME		LAW		NLSY		STUD		UNIV	
	MSE	KS	MSE	KS	MSE	KS	MSE	KS	MSE	KS
RLS	$.033 \pm .003$	$.55 \pm .06$	$.107 \pm .010$	$.15 \pm .02$	$.153 \pm .016$	$.73 \pm .07$	$4.77 \pm .49$	$.50 \pm .05$	$2.24 \pm .22$	$.14 \pm .01$
RLS+Berk	$.037 \pm .004$	$.16 \pm .02$	$.121 \pm .013$	$.10 \pm .01$	$.189 \pm .019$	$.49 \pm .05$	$5.28 \pm .57$	$.32 \pm .03$	$2.43 \pm .23$	$.05 \pm .01$
RLS+Oneto	$.037 \pm .004$	$.14 \pm .01$	$.112 \pm .012$	$.07 \pm .01$	$.156 \pm .016$	$.50 \pm .05$	$5.02 \pm .54$	$.23 \pm .02$	$2.44 \pm .26$	$.05 \pm .01$
RLS+Ours	$.041 \pm .004$	$.12 \pm .01$	$.141 \pm .014$	$.02 \pm .01$	$.203 \pm .019$	$.09 \pm .01$	$5.62 \pm .52$	$.04 \pm .01$	$2.98 \pm .32$	$.02 \pm .01$
KRLS	$.024 \pm .003$	$.52 \pm .05$	$.040 \pm .004$	$.09 \pm .01$	$.061 \pm .006$	$.58 \pm .06$	$3.85 \pm .36$	$.47 \pm .05$	$1.43 \pm .15$	$.10 \pm .01$
KRLS+Oneto	$.028 \pm .003$	$.19 \pm .02$	$.046 \pm .004$	$.05 \pm .01$	$.066 \pm .007$	$.06 \pm .01$	$4.07 \pm .39$	$.18 \pm .02$	$1.46 \pm .13$	$.04 \pm .01$
KRLS+Perez	$.033 \pm .003$	$.25 \pm .02$	$.048 \pm .005$	$.04 \pm .01$	$.065 \pm .007$	$.08 \pm .01$	$3.97 \pm .38$	$.14 \pm .02$	$1.50 \pm .15$	$.06 \pm .01$
KRLS+Ours	$.034 \pm .004$	$.09 \pm .01$	$.056 \pm .005$	$.01 \pm .01$	$.081 \pm .008$	$.03 \pm .01$	$4.46 \pm .43$	$.03 \pm .01$	$1.71 \pm .16$	$.02 \pm .01$
RF	$.020 \pm .002$	$.45 \pm .04$	$.046 \pm .005$	$.11 \pm .01$	$.055 \pm .006$	$.55 \pm .06$	$3.59 \pm .39$	$.45 \pm .05$	$1.31 \pm .13$	$.10 \pm .01$
RF+Raff	$.030 \pm .003$	$.21 \pm .02$	$.058 \pm .006$	$.06 \pm .01$	$.066 \pm .006$	$.08 \pm .01$	$4.28 \pm .40$	$.09 \pm .01$	$1.38 \pm .12$	$.02 \pm .01$
RF+Agar	$.029 \pm .003$	$.13 \pm .01$	$.050 \pm .005$	$.04 \pm .01$	$.065 \pm .006$	$.07 \pm .01$	$3.87 \pm .41$	$.07 \pm .01$	$1.40 \pm .13$	$.02 \pm .01$
RF+Ours	$.033 \pm .003$	$.08 \pm .01$	$.064 \pm .006$	$.02 \pm .01$	$.070 \pm .007$	$.03 \pm .01$	$4.18 \pm .38$	$.02 \pm .01$	$1.49 \pm .14$	$.01 \pm .01$

Table 1: Results for all the datasets and all the methods concerning MSE and KS.

- Performs well for various datasets and models.
- MSE is slightly larger than the baselines, while KS is slightly lower than the baselines.

Thank you