# Minimax Optimal Fair Regression under Linear Model

Kazuto Fukuchi and Jun Sakuma

University of Tsukuba / RIKEN

## Summary

- We investigate the minimax optimality of fair regression in terms of demographic parity under a certain linear model.
- Our model poses the following additional challenges compared to the existing results of Chzhen et al. (2022):
  1. Presence of *redlining effect*
  2. Mitigation biases in the second moment of the outcome
- We reveal the minimax optimal error of $\sigma_\xi^2 B^2 dM/n$.

## Model

Let $X$ be non-sensitive features on $\mathbb{R}^d$, let $S$ be a sensitive feature on $[M]$ where $M \geq 2$, and let $Y$ be an outcome on $\mathbb{R}$.

### Our model

- Conditioned on $S = s$, $X \sim N(\mu_s, \sigma_X^2 I)$ for $\mu \in \mathbb{R}^d$ and $\sigma_X^2 > 0$.
- Outcome's model:

$$Y = f^*(X, S) + \xi = \langle \beta_S^*, X \rangle + \xi, \tag{1}$$

where $\xi \sim N(0, \sigma_\xi^2)$ for $\sigma_\xi^2 > 0$.

## Fairness

### Definition: demographic parity (Pedreshi et al. 2008)

A regressor $f$ satisfies (strong) demographic parity if for all $s, s' \in [M]$, and for all $E \in \sigma(f(X, S))$,

$$\mathbb{P}\{f(X, S) \in E | S = s\} = \mathbb{P}\{f(X, S) \in E | S = s'\}. \tag{2}$$

- **Fairness consistency** is an approximation of the exact fairness guarantee and requires the learned regressor to approach an (exactly) fair regressor as $n$ tends to infinity.
- We use the following Wasserstain distance-based unfairness score to define "approaching".

$$U(f) = \max_{s, s' \in [M]} W_2(\nu_{f|s}, \nu_{f|s'}) \tag{3}$$

### Definition: $(\alpha, \delta)$-fairness consistency

A learning algorithm is $(\alpha, \delta)$-consistently fair for an unfairness score $U$ if there exists constants $n_0 \geq 0$ and $C > 0$ independent of $n$ such that $\mathbb{P}\{U(\hat{f}_n) > Cn^{-\alpha}\} \leq \delta$ for all $n \geq n_0$.

## Accuracy

- The goal of the leaner is to obtain a *fair version of* $f^*$.
- By employing the $L^2$ distance for assessing the closeness, we define it as $f_{\mathrm{DP}}^* = \arg\min_{f \in \mathcal{F}_{\mathrm{DP}}(\mu_\cdot)} \mathbf{E}[(f(X, S) - f^*(X, S))^2]$.
- We evaluate the inaccuracy of a regressor $f$ by the mean squared deviation from $f_{\mathrm{DP}}^*$, defined as

$$\mathcal{E}(f; \beta_\cdot^*, \mu_\cdot) = \mathbf{E}\left[\left(f(X, S) - f_{\mathrm{DP}}^*(X, S)\right)^2\right], \tag{4}$$

### Definition: minimax optimal error

Given $\alpha > 0$ and $\delta \in (0, 1)$, the minimax optimal error is defined as

$$\mathcal{E}_n(\alpha, \delta) = \inf_{\hat{f}_n : (\alpha, \delta)\text{-consistently fair}} \sup_{\beta^* \in \mathcal{B}, \mu_\cdot \in \mathcal{M}} \mathbf{E}\left[\mathcal{E}(\hat{f}_n; \beta^*, \mu_\cdot)\right], \tag{5}$$

## Challenges

Chzhen et al. (2022) is the only work that reveals the minimax optimal error for fair regression.

### Chzhen et al. (2022)'s model

- $X \sim N(0, \Sigma)$ for a positive semi-definite matrix $\Sigma$.
- Outcome's model:

$$Y = \langle \beta^*, X \rangle + b_s + \xi, \tag{6}$$

- **(Redlining)** $X$ is independent of $S$ in Chzhen et al. (2022)'s model, which cannot simulate a crucial phenomenon, *redlining effect*. In contrast, our model varies the mean of $X$ by $S$, by which we can (partly) simulate the presence of redlining effect.
- **(Second-order bias)** In Chzhen et al. (2022)'s model, $\mathbf{E}[Y|S = s]$ is changed by $s$, but the higher conditional (central) moments do not. In contrast, $\mathbf{Var}[Y|S = s]$ varies by $s$ in addition to the mean in our model.

## Main result

1. There is a finite universal constant $B > 0$ such that

$$\|\beta_s\| \leq B \text{ and } \frac{(\sum_{s \in [M]} p_s \|\beta_s\|)^2}{M} \sum_{s \in [M]} \|\beta_s\|^{-2} \leq B^2 \tag{7}$$

2. There exists a finite universal constant $U > 0$ such that $\|\mu_s\| \leq U$.

### Main theorem

If $\alpha \in (0, 1/2]$, $M(d-1) > 16$, and $n \geq 12(3d \vee 4\ln(M/\delta))/\min_{s \in [M]} p_s$, there exist universal constants $C > 0$ and $c > 0$ such that for any $\delta \in (0, 1)$,

$$c\frac{\sigma_\xi^2 B^2 dM}{n} - o\left(\frac{1}{n}\right) \leq \mathcal{E}_n(\alpha, \delta)$$
$$\leq C\frac{\sigma_X^2 \sigma_\xi^2 B^2 dM \vee U^2 \sigma_\xi^2 \vee U^2 B^2}{n} + o\left(\frac{1}{n}\right). \tag{8}$$

The upper bound is achieved by a carefully designed plugin estimator (See our paper for detail).

## Implications

- The constructed estimator is minimax optimal up to constant depending on $U$ and $\sigma_X^2$.
- The term $\sigma_\xi^2 dM/n$ is natural and the same as the non-fair regression.
- **(Redlining)** There is no term regarding redlining, meaning that the dependency of $S$ in the mean of $X$ is too simple not to have an effect on statistical efficiency.
- **(Second-order bias)** $B$ assesses the diversity of the outcomes' variances, meaning that $B$ represents the difficulty in mitigating the second-order bias.

### References

Chzhen, Evgenii and Nicolas Schreuder (Aug. 2022). "A minimax framework for quantifying risk-fairness trade-off in regression". In: *The Annals of Statistics* 50.4, pp. 2416–2442. ISSN: 0090-5364. DOI: 10.1214/22-AOS2198.

Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini (2008). "Discrimination-aware data mining". In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, pp. 560–568. ISBN: 9781605581934. DOI: 10.1145/1401890.1401959.

Check out the arXiv version!