Demographic Parity Constrained Minimax Optimal Regression under Linear Model

Kazuto Fukuchi[1,3] and Jun Sakuma[2,3]

[1] University of Tsukuba [2] Tokyo Institute of Technology [3] RIKEN AIP

Summary

- We investigate the minimax optimality of regression with the constraint of demographic parity.
- Our model poses the following additional challenges compared to the existing results of Chzhen et al. (2022):
  - (Direct discrimination) Mitigating outcome's variance disparity.
  - (Indirect discrimination) Addressing indirect discrimination.
- We reveal the minimax optimal error rate as $\sigma_\xi^2 B^2 dM/n$.

Setup

Consider X as non-sensitive features in $\mathbb{R}^d$ and S as a sensitive feature within $[M]$. Given noise $\xi \sim N(0, \sigma_\xi^2)$, the outcome Y is:

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

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** requires the learned regressor to approach an (exactly) fair regressor as n tends to infinity.
- We use the 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: (α, δ)-fairness consistency

A learning algorithm is (α, δ)-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

- Goal: to obtain a *fair version of* f^*, defined as $f^*_{\mathrm{DP}} = \arg\min_{f \in \mathcal{F}_{\mathrm{DP}}(\mu.)} \mathbf{E}[(f(X, S) - f^*(X, S))^2]$.
- Inaccuracy of f is measured by the mean squared deviation from f^*_{DP}:

$$\mathcal{E}(f; \beta^*, \mu.) = \mathbf{E}\left[(f(X, S) - f^*_{\mathrm{DP}}(X, S))^2\right]. \tag{4}$$

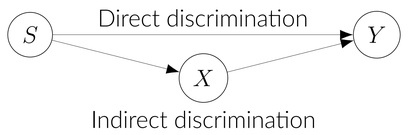
Definition: minimax optimal error

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

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

Sources of Unfairness (Direct v.s. Indirect)

- (Direct discrimination) Sensitive attribute directly affects the outcome, regardless of non-sensitive features.
- (Indirect discrimination) Sensitive attribute indirectly affects the outcome through its correlation with non-sensitive features.



Models

- Chzhen et al. (2022) is the sole study demonstrating minimax optimality in fair regression.
- Contrast with Chzhen et al. (2022): our model accounts for a broader source of discrimination.

	partial coefficients	intercept	non-sensitive features
Chzhen et al. (2022)		✓	
ours	✓	✓	✓

Chzhen et al. (2022)'s model

- Non-sensitive features' model: for a positive semi-definite matrix Σ,

$$X \sim \underbrace{N(0, \Sigma)}. \tag{6}$$

No dependency on S. No indirect discrimination.

- Outcome's model:

$$Y = \underbrace{\langle \beta^*, X \rangle + b_s} + \xi, \tag{7}$$

Intercepts (b_s) may cause direct discrimination.

Our model

- Non-sensitive features' model: for $\sigma_X^2 > 0$,

$$X \sim \underbrace{N(\mu_S, \sigma_X^2 I)}. \tag{8}$$

Means depend on on S, leading to indirect discrimination.

- Outcome's model:

$$Y = \underbrace{\langle \beta^*_s, X \rangle} + \xi, \tag{9}$$

Both partial coefficients and intercept may cause direct discrimination.

Challenges

(Direct discrimination)

- Variability in partial coefficients leads to varied outcome variances against S, while diverse intercepts only change the outcome's mean.
- Our model poses a challenge of addressing both variance and mean disparities.

(Indirect discrimination)

- Our model introduces indirect discrimination via μ_S changes relative to S.
- Counteracting this requires estimating μ_S to fine-tune the regressor, maintaining consistent output across varying μ_S.

Main result

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

$$\|\beta^*_s\| \leq B \text{ and } \frac{(\sum_{s \in [M]} p_s \|\beta^*_s\|)^2}{M} \sum_{s \in [M]} \|\beta^*_s\|^{-2} \leq B^2 \tag{10}$$

Greater variation in $\|\beta^*_s\|$ increases B, characterizing dispersion of outcome variances.

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

Main theorem

Given $\alpha \in (0, 1/2]$ and $\delta \in (0, 1)$, suppose $M(d-1) > 16$ and $n \geq 12(3d \vee 4\ln(M/\delta))/\min_{s \in [M]} p_s$.

$$c\frac{\sigma_\xi^2 B^2 dM}{n} - o\left(\frac{1}{n}\right) \leq \mathcal{E}_n(\alpha, \delta) \leq C\frac{\sigma_\xi^2 B^2 dM \vee \sigma_X^2 B^2 M \vee B^2 U^2}{n} + o\left(\frac{1}{n}\right). \tag{11}$$

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 σ_X^2.
- The term $\sigma_\xi^2 dM/n$ aligns with standard non-fair regression.
- **(Direct discrimination)** The greater the dispersion of outcome variances, the more difficult it becomes to mitigate direct discrimination due to the outcome's variances.
- **(Indirect discrimination)** Indirect discrimination can be mitigated cost-free if X's dependence of S is solely on its mean.

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 arXiv version!

fukuchi@cs.tsukuba.ac.jp