Group Fairness with Uncertain Sensitive Attributes


Abhin Shah — MIT
Maohao Shen — MIT
Jongha Jon Ryu — MIT
Subhro Das — MIT-IBM Watson AI Lab
Prasanna Sattigeri — MIT-IBM Watson AI Lab
Yuheng Bu — UF University of Florida
Greg Wornell — MIT

arXiv Link

Group Fairness

Age

Location

BMI

Number of Children

Smoker

Features x

Gender

Sensitive Attribute e

Medical expenses

Outcome y

Machine Learning Model

Accurate

Fair

min Prediction Loss s.t Fairness Loss $\leq \epsilon$

Fairness Loss — measures the degree of violation of the (conditional) independence requirement of group fairness

Uncertainty in Sensitive Attribute

Missing — Additional annotation

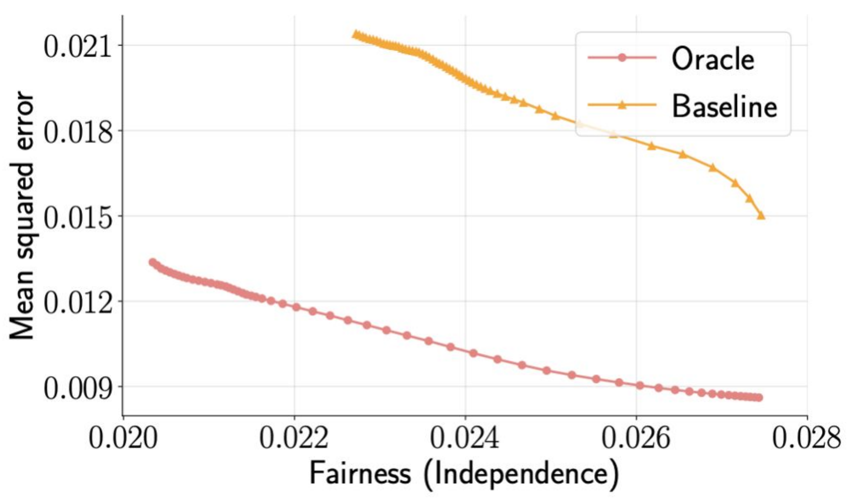
Gender

Unreliable — Noisy; Response bias in a survey; Legal regulations (GDPR General Data Protection Regulation; California Consumer Privacy Act)

Age	Location	BMI	Number of children	Smoker	Medical expenses	Gender	Missing Gender	Unreliable Gender
19	Southwest	27.9	0	Yes	16884	Female	?	Male
28	Southeast	33	3	No	4449	Male	Male	Male
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	Southeast	26.29	0	Yes	27808	Female	?	Female

Features x — Outcome y — Sensitive Attribute e ($\mathcal{D}^{(oracle)}$) — Uncertain Sensitive Attribute $\hat{e}$ ($\mathcal{D}^{(uncertain)}$)

Fair loss estimated with $\mathcal{D}^{(oracle)}$ vs Fair loss estimated with $\mathcal{D}^{(uncertain)}$



Goal — Learn a fair model despite uncertain sensitive attribute data

Limitations of Existing Work

A. Proxy variables — effectiveness depends on the degree of correlation between e and x

B. Perturbed sensitive attribute — focus on specific perturbation models

Problem Formulation

• Predictor f • Loss function ℓ • Fairness measure Φ • Fairness target ϵ

$$f^\star \in \arg\min_{f \in \mathcal{F}} \mathbb{E}[\ell(y, f(x))] \quad \text{s.t.} \quad \Phi(y, f(x), e) \leq \epsilon$$

Fairness measures

- Independence (demographic parity) — $f(x) \perp\!\!\!\perp e$
- Separation (equalized odds) — $f(x) \perp\!\!\!\perp e \mid y$

Choices of Φ

- Independence — $\Phi(y, f(x), e) = \chi^2(p_{e,f(x)} \| p_e p_{f(x)})$
- Separation — $\Phi(y, f(x), e) = \mathbb{E}_{p_y}[\chi^2(p_{e,f(x)|y} \| p_{e|y} p_{f(x)|y})]$

Gaussian Data

Model the distribution of $(x, y, e, u = f(x))$ as Gaussian

Quadratically Constrained Quadratic Program (QCQP)

$$\max_{a \in \mathcal{B}(0,1)} \langle a, b_{yx} \rangle^2 \quad \text{s.t} \quad \langle a, b_{ex} \rangle^2 \leq \epsilon \quad \text{where} \quad a = b_{ux} \text{ and } b_{vw} \triangleq \Sigma_v^{-1/2} \Sigma_{vw} \Sigma_w^{-1/2}$$

An optimal solution $a^\star$ of the above QCQP lies in the subspace spanned by the vectors b_{yx} and b_{ex}

Baseline

$$\max_{a \in \mathcal{B}(0,1)} \langle a, b_{yx} \rangle^2 \quad \text{s.t} \quad \langle a, \hat{b}_{ex} \rangle^2 \leq \epsilon$$

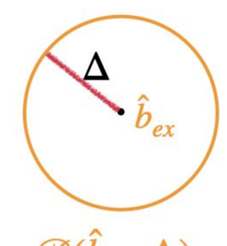
This does not guarantee fairness!

A predictor u satisfying $\Phi_{\mathcal{D}^{(uncertain)}}(y, u, e) \leq \epsilon$ on may not satisfy $\Phi_{\mathcal{D}^{(oracle)}}(y, u, e) \leq \epsilon$

Robust QCQP

Uncertainty in sensitive attributes

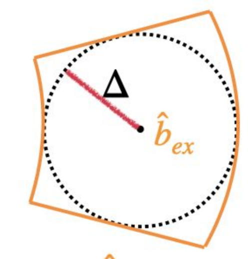
$$\max_{a \in \mathcal{B}(0,1)} \langle a, b_{yx} \rangle^2 \quad \text{s.t} \quad \langle a, b \rangle^2 \leq \epsilon \quad \text{for all} \quad b \in \mathcal{B}(\hat{b}_{ex}, \Delta)$$

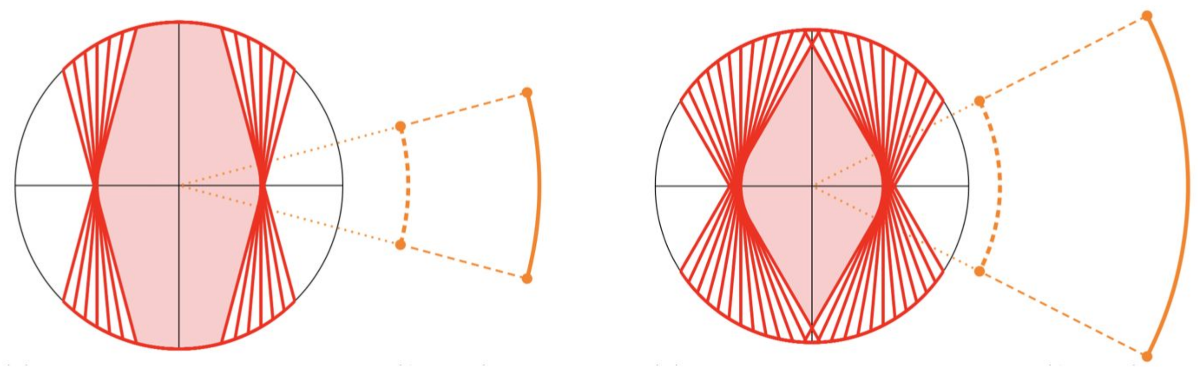


An optimal solution $a^\star$ of the above QCQP lies in the subspace spanned by the vectors b_{yx} and $\hat{b}_{ex}$

Relaxing the uncertainty

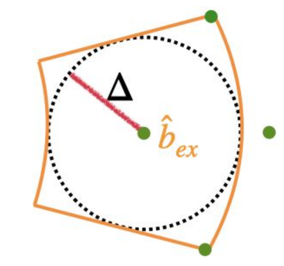
$$\max_{a \in \mathcal{B}(0,1)} \langle a, b_{yx} \rangle^2 \quad \text{s.t} \quad \langle a, b \rangle^2 \leq \epsilon \quad \text{for all} \quad b \in \mathcal{A}(\hat{b}_{ex}, \Delta)$$





4 constraints

$$\max_{a \in \mathcal{B}(0,1)} \langle a, b_{yx} \rangle^2 \quad \text{s.t} \quad \langle a, \hat{b}_{ex} \rangle^2 \leq \epsilon \quad \text{and} \quad \langle a, b^{(i)} \rangle^2 \leq \epsilon \quad \text{for all} \quad i \in [3]$$



Algorithm

Baseline :— $\min_u \mathbb{E}[\ell(y, u)]$ s.t. $\Phi_{\mathcal{D}^{(uncertain)}}(y, u, e) \leq \epsilon$

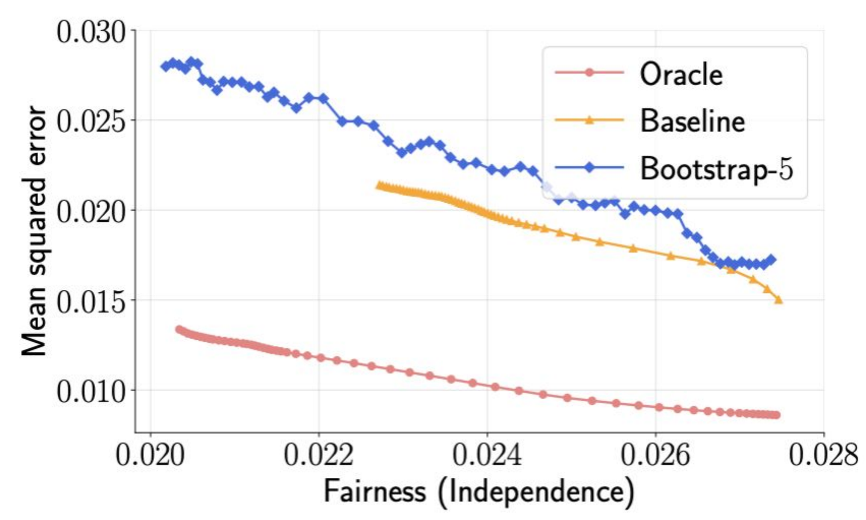
$\mathcal{D}^{(uncertain)}$ → $\mathcal{D}^{(uncertain)}$, Sub-sample $\mathcal{D}^{(1)}$, …, Sub-sample $\mathcal{D}^{(S)}$

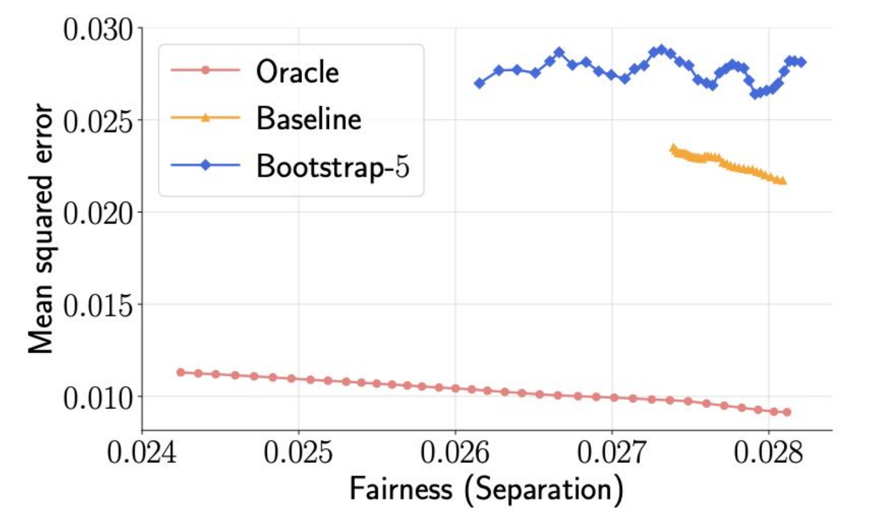
For some $k \in [n]$ and $S \geq 1$, uniformly draw $\mathcal{D}^{(1)}, \cdots, \mathcal{D}^{(S)}$ each of size k from $\mathcal{D}^{(uncertain)}$ with replacement.

Bootstrap-S :— $\min_u \mathbb{E}[\ell(y, u)]$ s.t. $\Phi_{\mathcal{D}^{(uncertain)}}(y, u, e) \leq \epsilon$ and $\Phi_{\mathcal{D}^{(i)}}(y, u, e) \leq \epsilon$ for all $i \in [S]$

Insurance Dataset (regression)

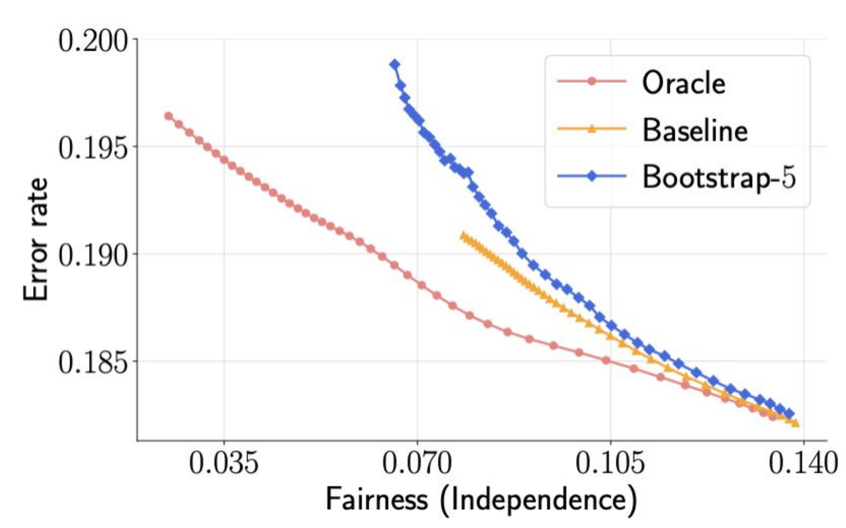
Sensitive attribute — Gender (Discrete) // Uncertainty — limited sensitive attribute

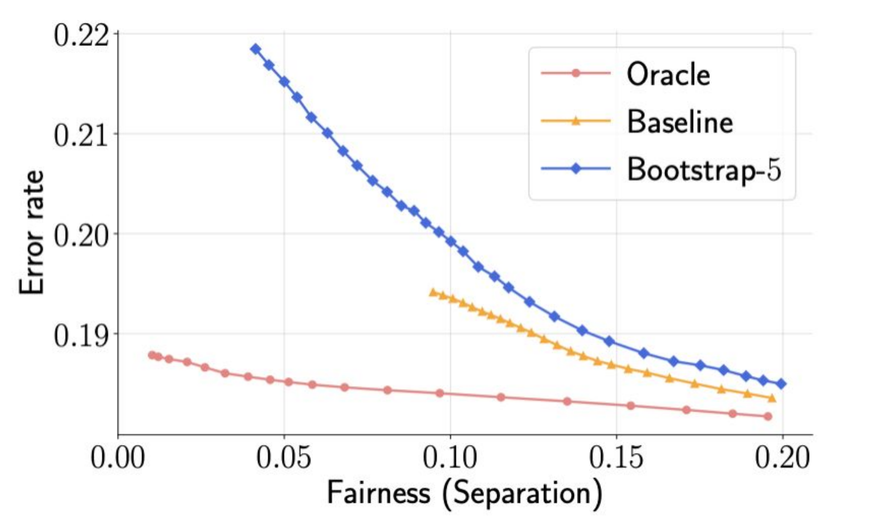




Adult Dataset (classification)

Sensitive attribute — Gender (Discrete) // Uncertainty — limited sensitive attribute





Crime Dataset (regression)

Sensitive attribute — Race (Continuous) // Uncertainty — unreliable sensitive attribute