Group Fairness with Uncertainty in Sensitive Attributes

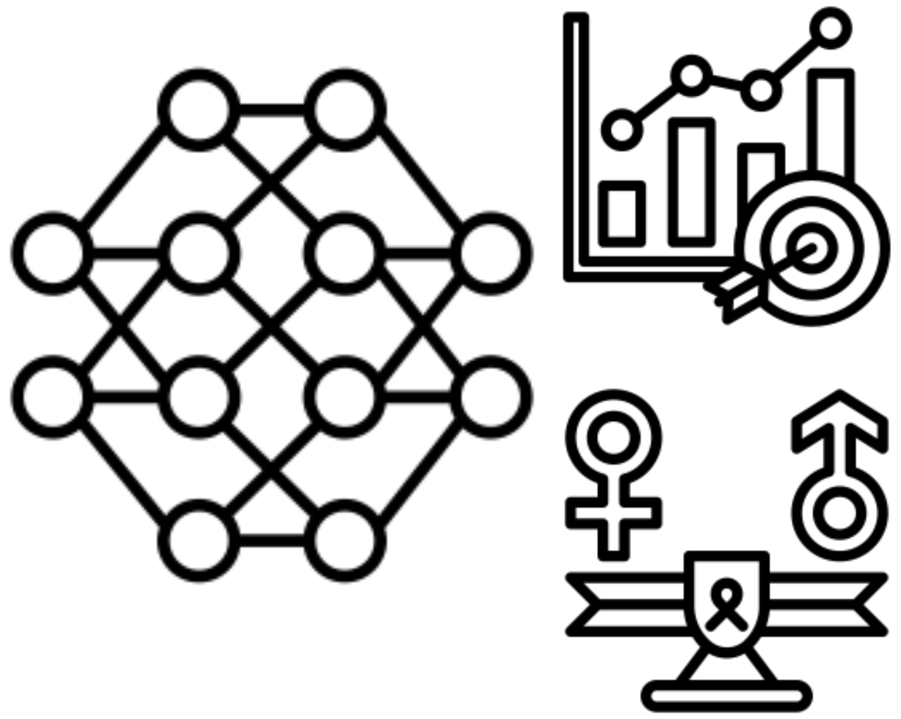
Abhin Shah, Maohao Shen, Jongha Jon Ryu, Subhro Das, Prasanna Sattigeri, Yuheng Bu, Gregory W. Wornell

UF UNIVERSITY of FLORIDA

IBM Research

MIT-IBM Watson AI Lab

Group Fairness

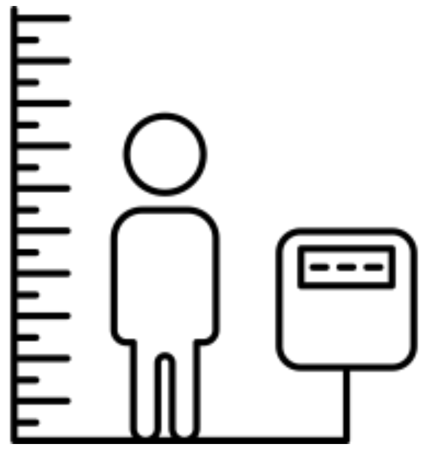


min Prediction Loss

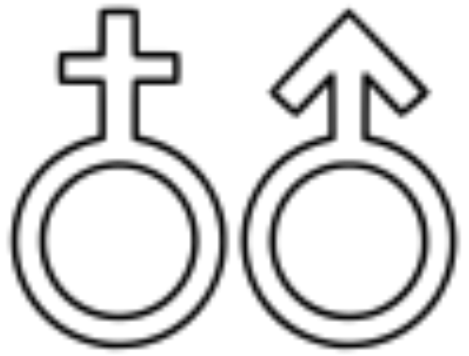
s.t. Fairness Loss $\leq \epsilon$

Age

Location

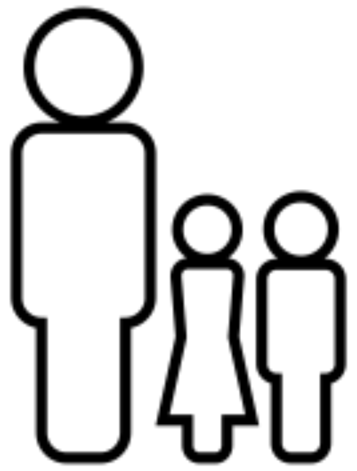


BMI



Gender

Medical expenses



Number of Children

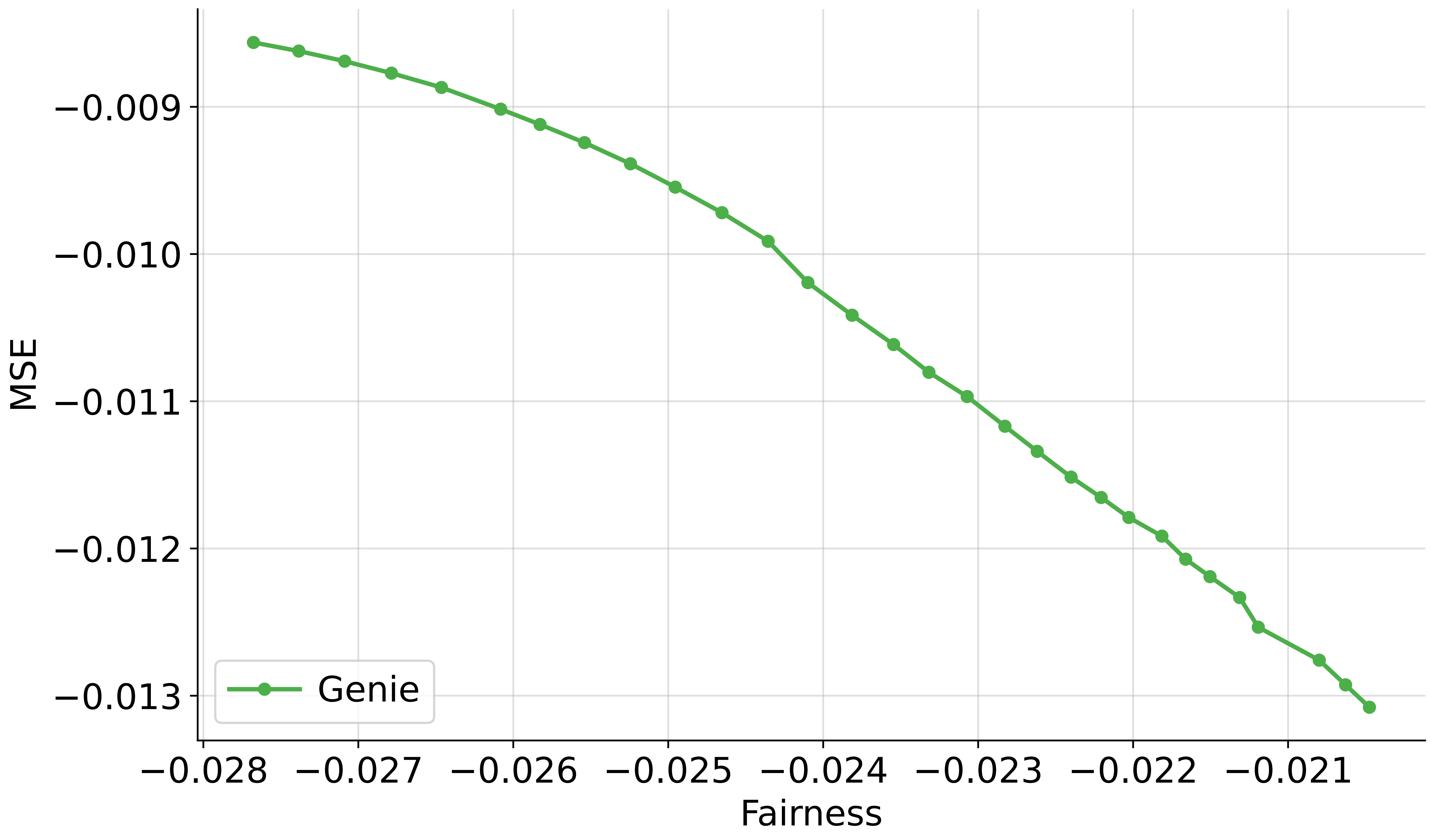
Smoker

Independence :— Prediction $\perp$ Gender

Separation :— Prediction $\perp$ Gender | Medical expenses

Sufficiency :— Medical expenses $\perp$ Gender | Prediction

Insurance Dataset



Lagrangian dual

$$\min \text{Prediction Loss} \quad \text{s.t.} \quad \text{Fairness Loss} \leq \epsilon$$

$$\Downarrow$$

$$\min \max_{\lambda \geq 0} \text{Prediction Loss} + \lambda \left(\underbrace{\text{Fairness Loss} -}_{\text{Sensitive Attribute}} \epsilon \right)$$

Uncertainty in Sensitive Attribute

Gender

Limited
- Additional annotation

Unreliable
- Noisy
- Response bias in a survey
- Legal regulations

European Commission

GDPR

General Data Protection Regulation

CALIFORNIA CONSUMER PRIVACY ACT

Limited sensitive attributes

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

$$\min \max_{\lambda \geq 0} \text{Prediction Loss} + \lambda \, (\underbrace{\text{Fairness Loss}}_{\text{Sensitive Attribute}} - \epsilon)$$

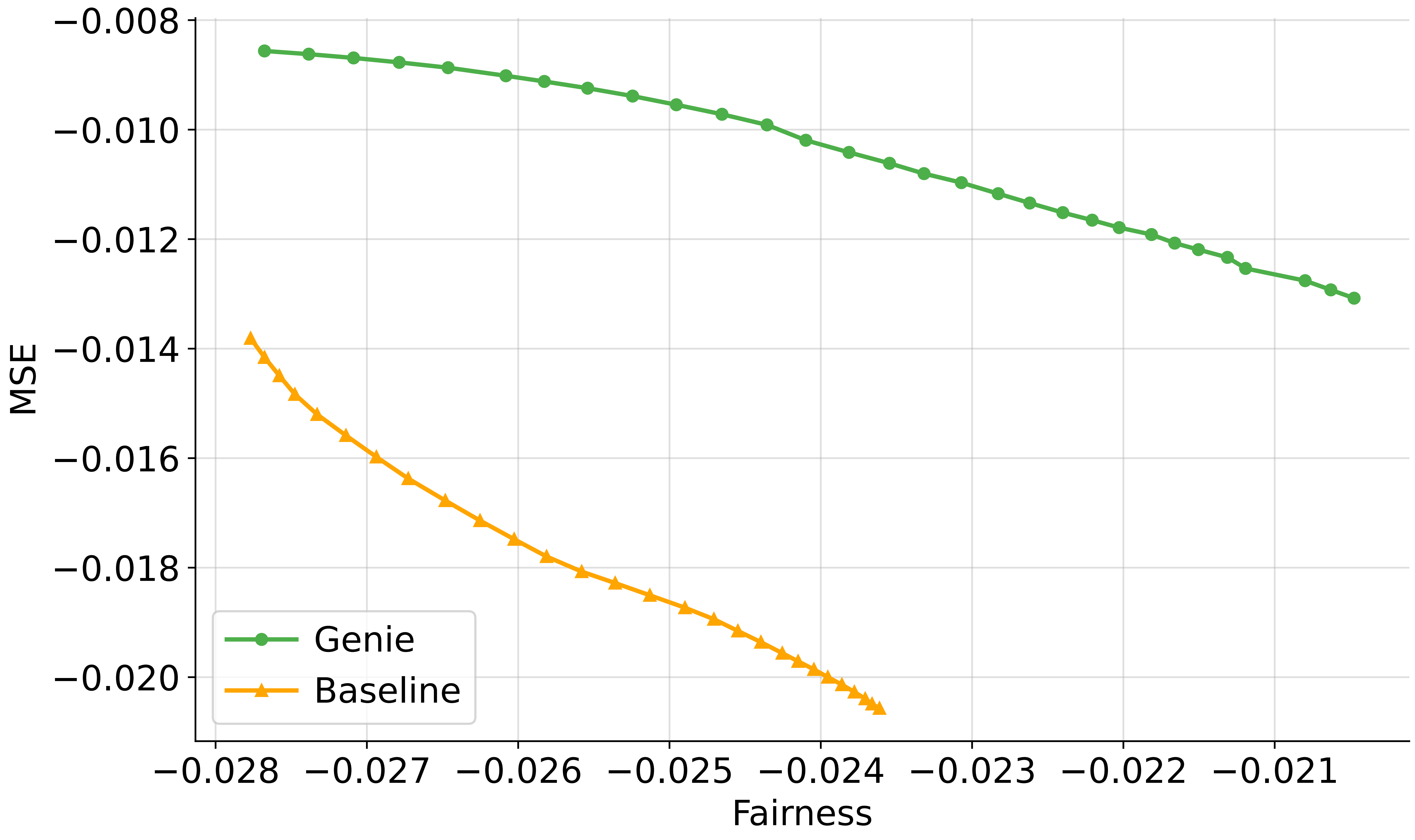
Unreliable sensitive attributes

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

$$\min \max_{\lambda \geq 0} \text{Prediction Loss} + \lambda \, (\underbrace{\text{Fairness Loss}}_{\text{Sensitive Attribute}} - \epsilon)$$

Insurance

Uncertainty — limited sensitive attribute

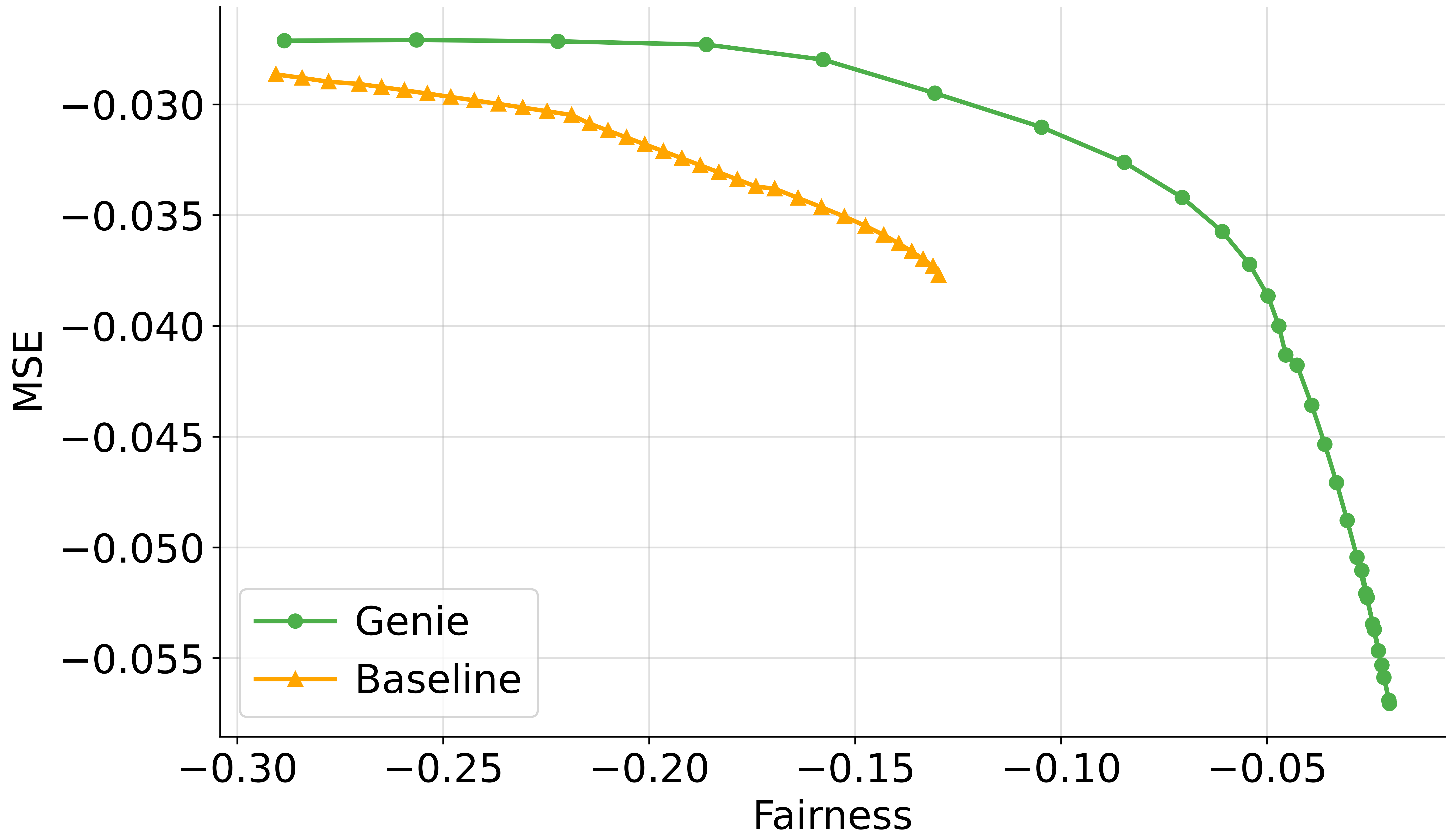


Communities and Crime

Task— predict density of violent crimes (Regression)

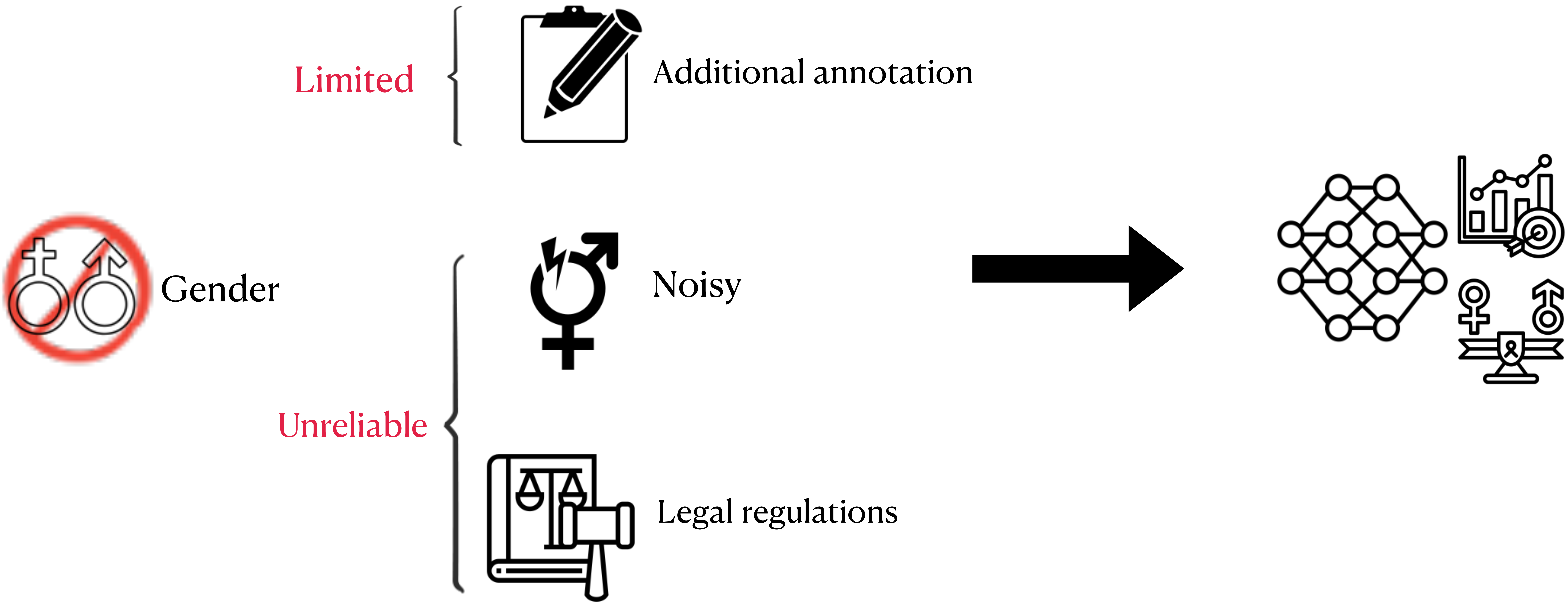
Sensitive attribute — Race (Continuous)

Uncertainty — unreliable sensitive attribute

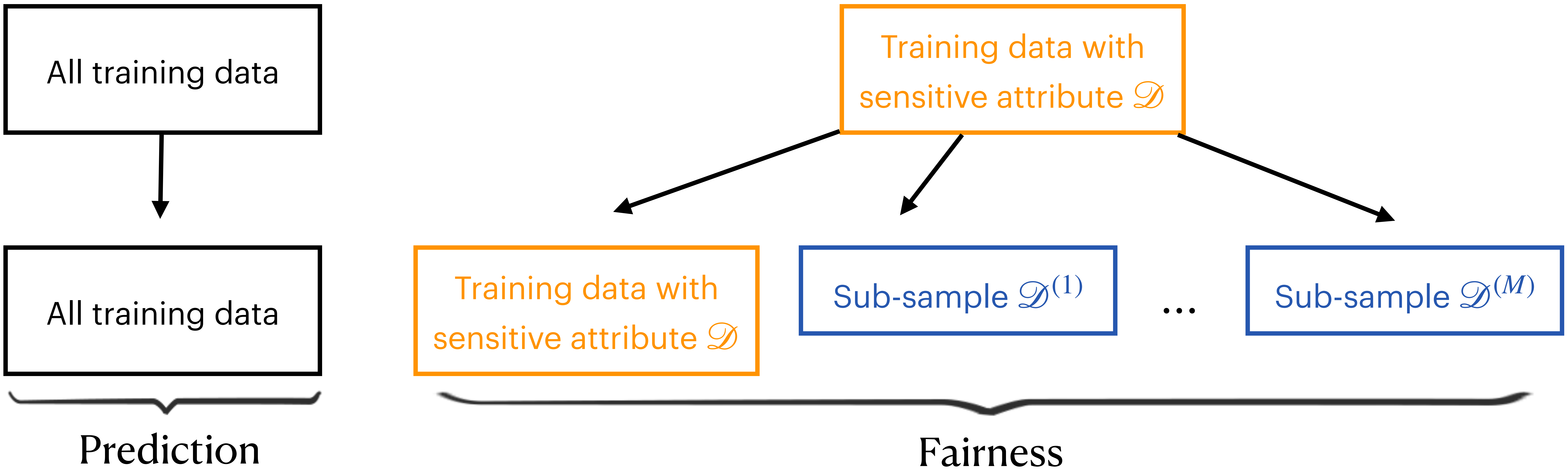


Goal

Learn a fair model despite uncertain sensitive attribute data.



A General Purpose Algorithm



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

$$\min \max_{\lambda \geq 0} \text{Prediction Loss} + \lambda\ (\text{Fairness Loss}(\mathscr{D}) - \epsilon) + \boxed{\sum_{i \in [M]} \lambda_i\ (\text{Fairness Loss}(\mathscr{D}^{(i)}) - \epsilon)}$$

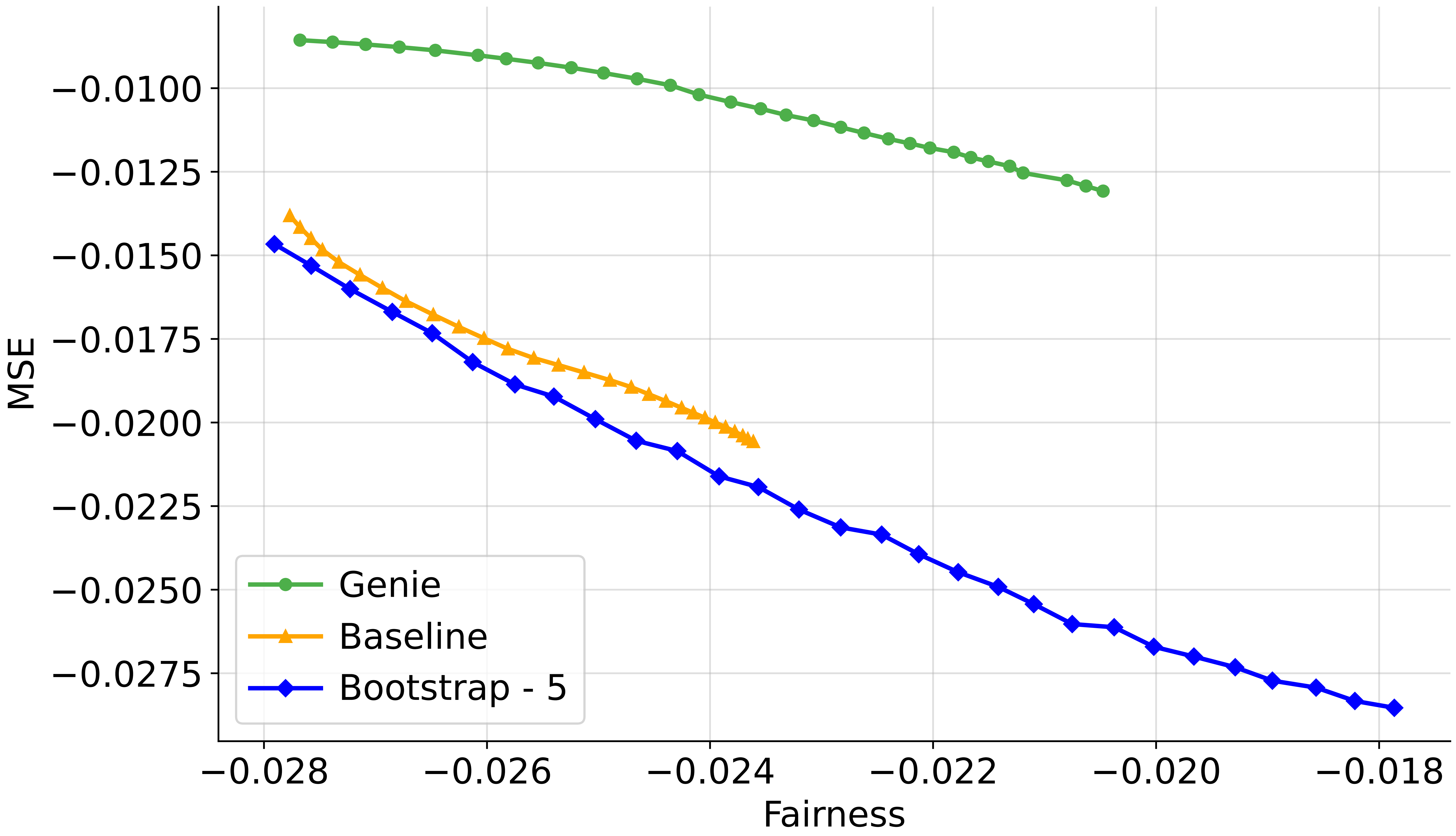
A General Purpose Algorithm

Bootstrap-M

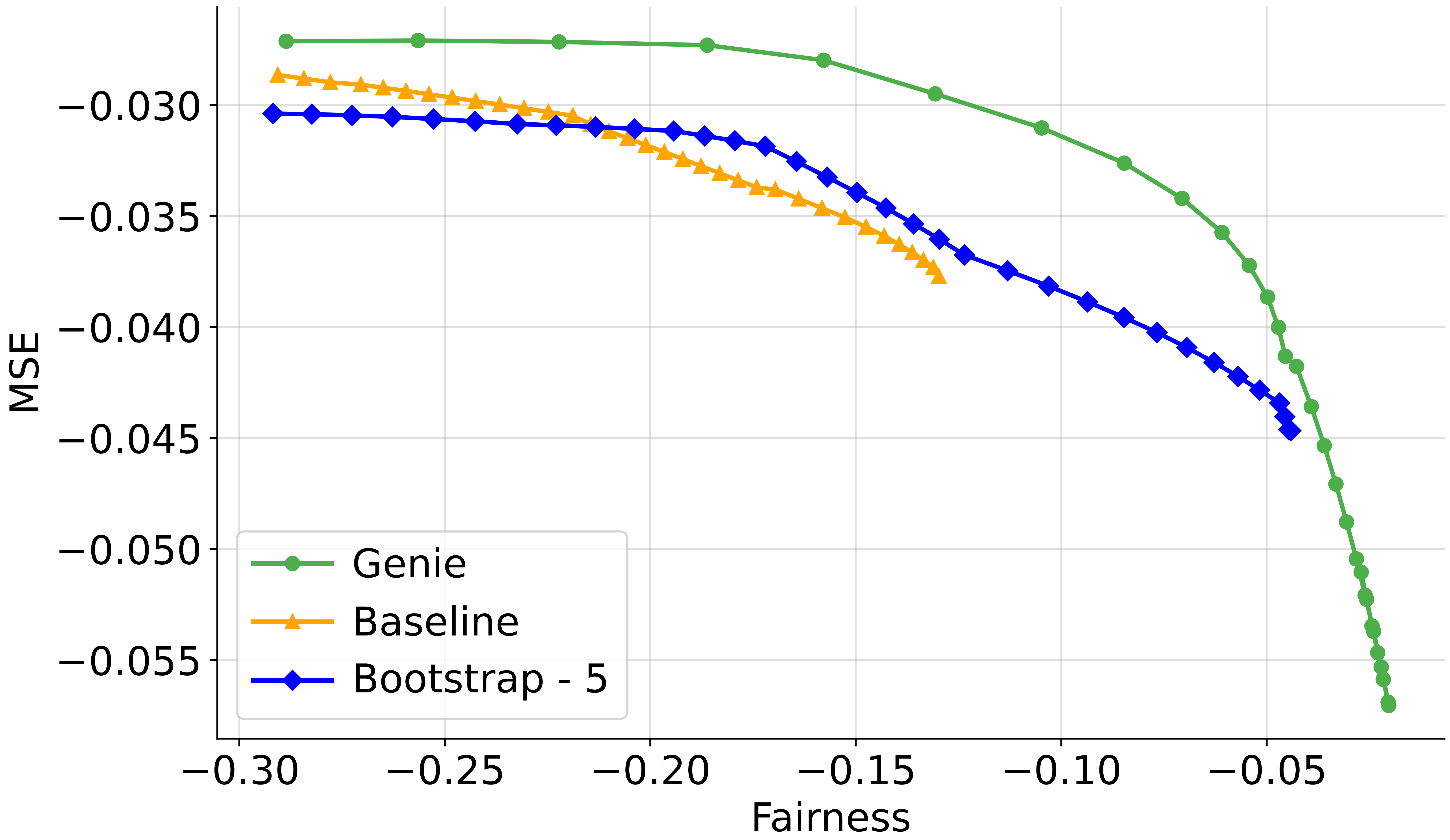
$$\min \max_{\lambda \geq 0} \text{Prediction Loss} + \lambda \, (\text{Fairness Loss}(\mathscr{D}) - \epsilon) + \boxed{\sum_{i \in [M]} \lambda_i \, (\text{Fairness Loss}(\mathscr{D}^{(i)}) - \epsilon)}$$

$$\begin{aligned} &\min \text{Prediction Loss} \quad \text{s.t.} \quad \text{Fairness Loss}(\mathscr{D}) \leq \epsilon \\ &\qquad\qquad\qquad\qquad\quad\; \text{s.t.} \quad \text{Fairness Loss}(\mathscr{D}^{(i)}) \leq \epsilon \ \text{ for all } i \in [M] \end{aligned}$$

Insurance Dataset



Crime Dataset



Setup

- x :— d-dimensional input
- y :— 1-dimensional target
- e :— 1-dimensional sensitive attribute
- $p_{x,y}$ and p_e :— known

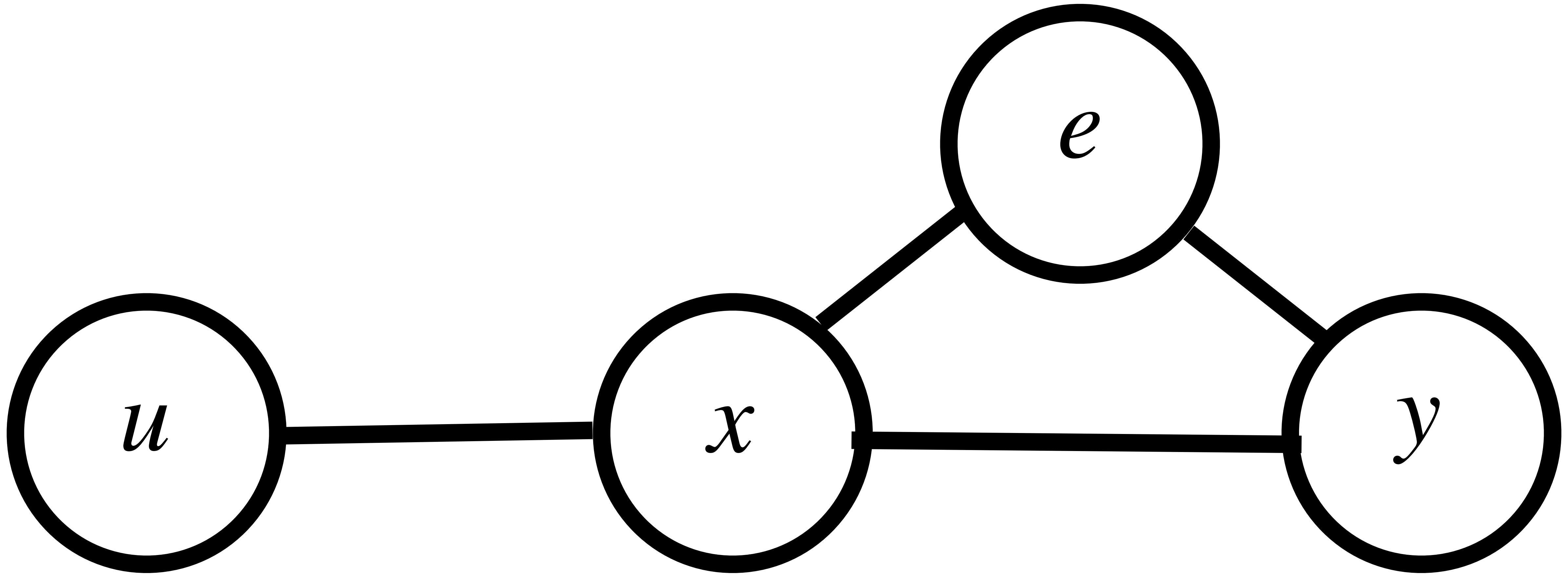
Gaussian Data

Quadratically Constrained Quadratic Program (QCQP)

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

$$\min \text{ Prediction Loss} \quad \text{s.t.} \quad \text{Fairness Loss } \leq \epsilon$$

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



Gaussian Data

Quadratically Constrained Quadratic Program (QCQP)

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

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

Baseline

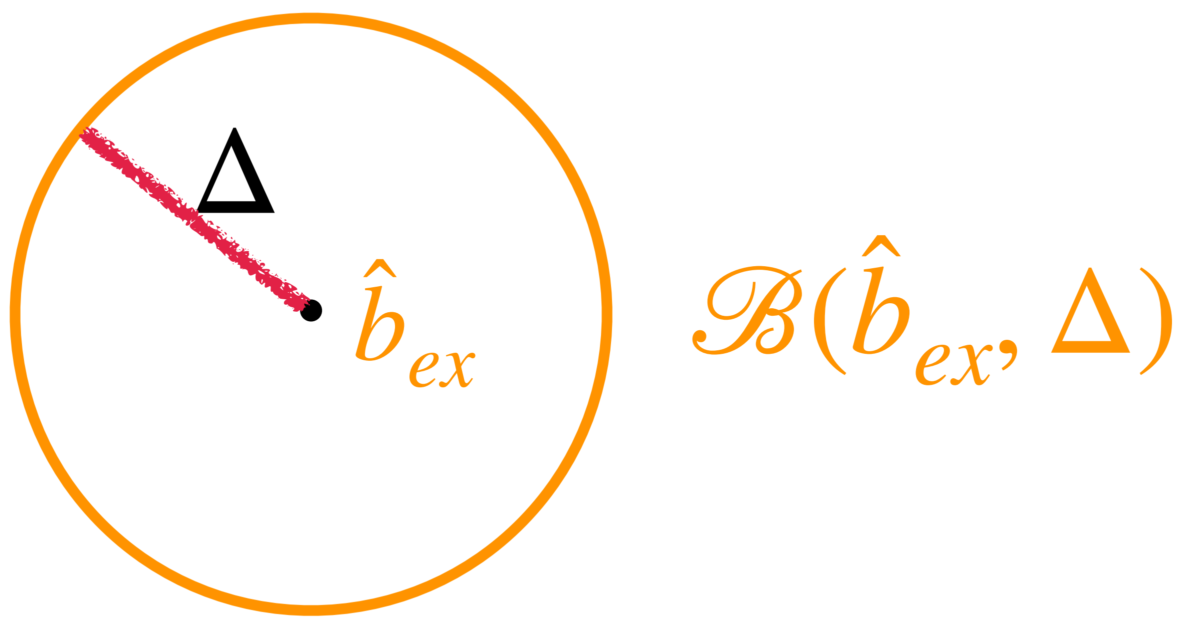
$$\max_{a \in \mathcal{B}(0,1)} \langle a, b_{yx} \rangle^2 \quad \text{s.t} \quad \langle a, \hat{b}_{ex} \rangle^2 \leq \epsilon$$

This does not guarantee fairness

$$b_{vw} \triangleq \Sigma_v^{-1/2} \Sigma_{vw} \Sigma_w^{-1/2}$$

Robust QCQP

Uncertainty in sensitive attributes

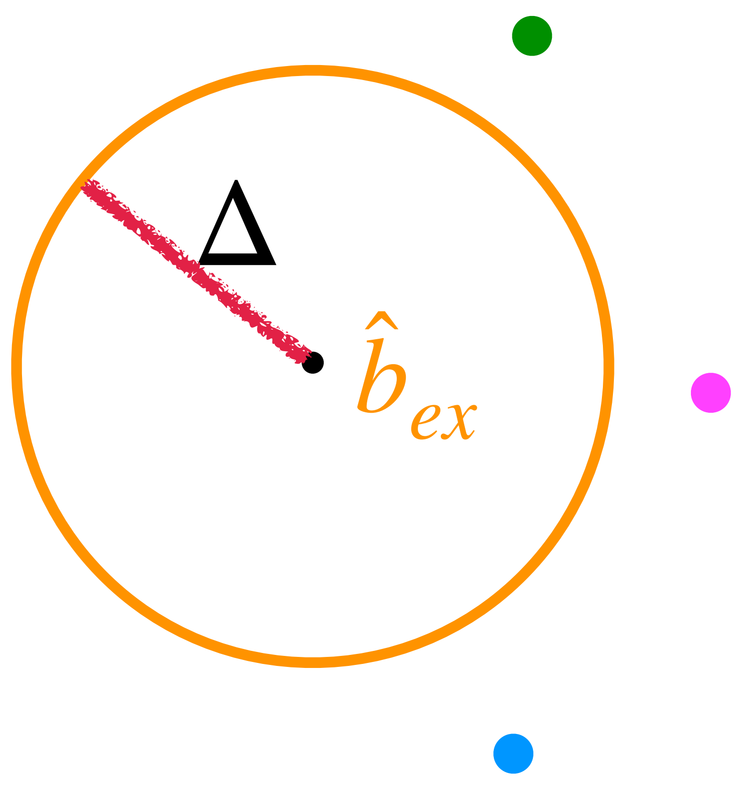


$$\max_{a \in \mathcal{B}(0,1)} \langle a, b_{yx} \rangle^2 \quad \text{s.t} \quad \langle a, b \rangle^2 \leq \epsilon \quad \text{for all} \quad b \in \mathcal{B}(\hat{b}_{ex}, \Delta)$$

$b_{vw} \triangleq \Sigma_v^{-1/2}\Sigma_{vw}\Sigma_w^{-1/2}$

Robust QCQP

Relaxing the uncertainty



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