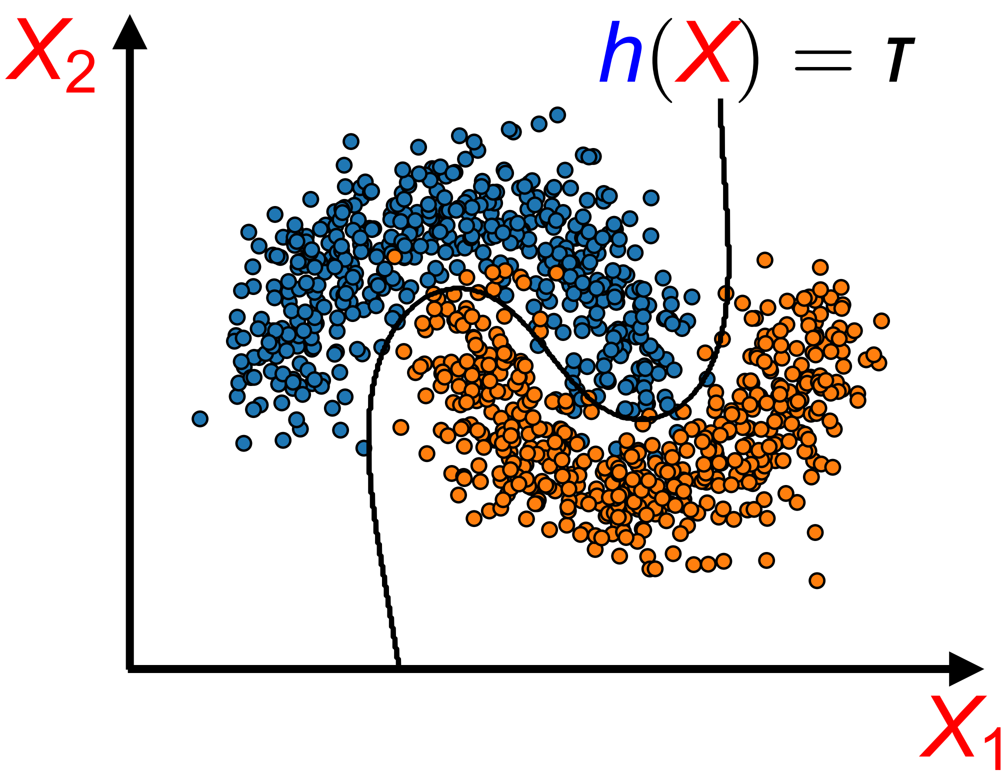
Metrizing Fairness

Yves Rychener, Bahar Taşkesen, Daniel Kuhn

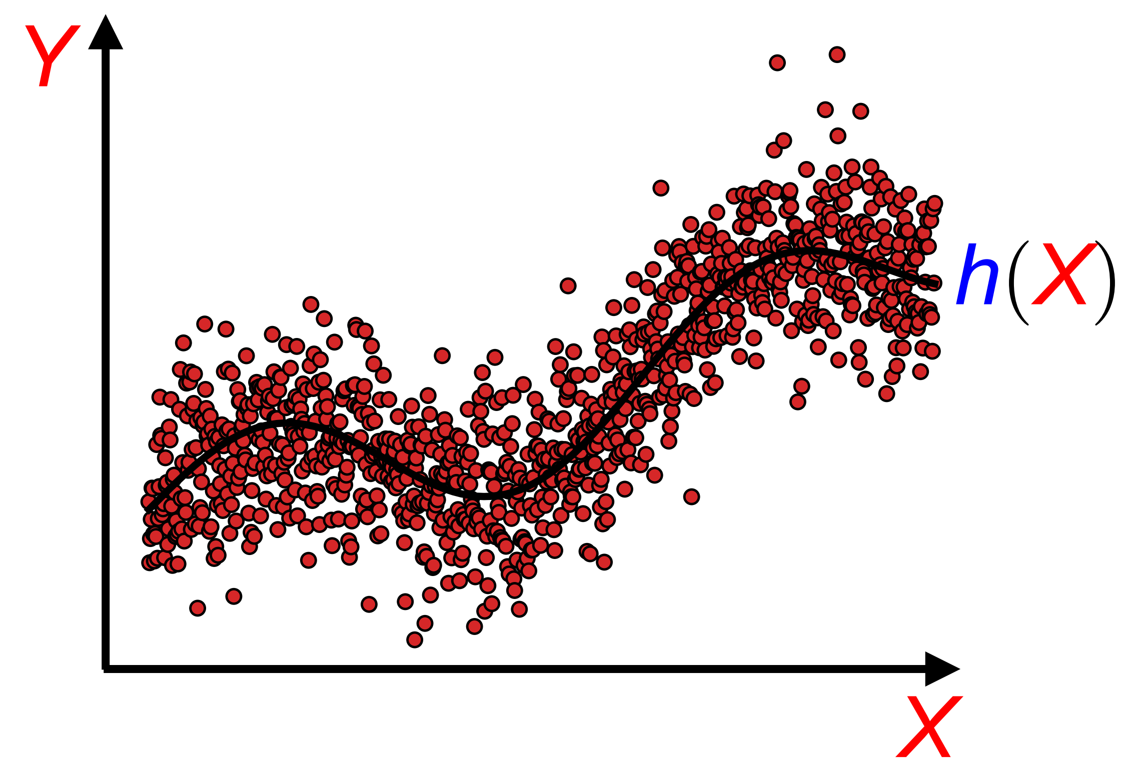
École Polytechnique Fédérale de Lausanne

FNSNF SWISS NATIONAL SCIENCE FOUNDATION   NCCR Automation   EPFL

Learning problem: $\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)]$



Classification

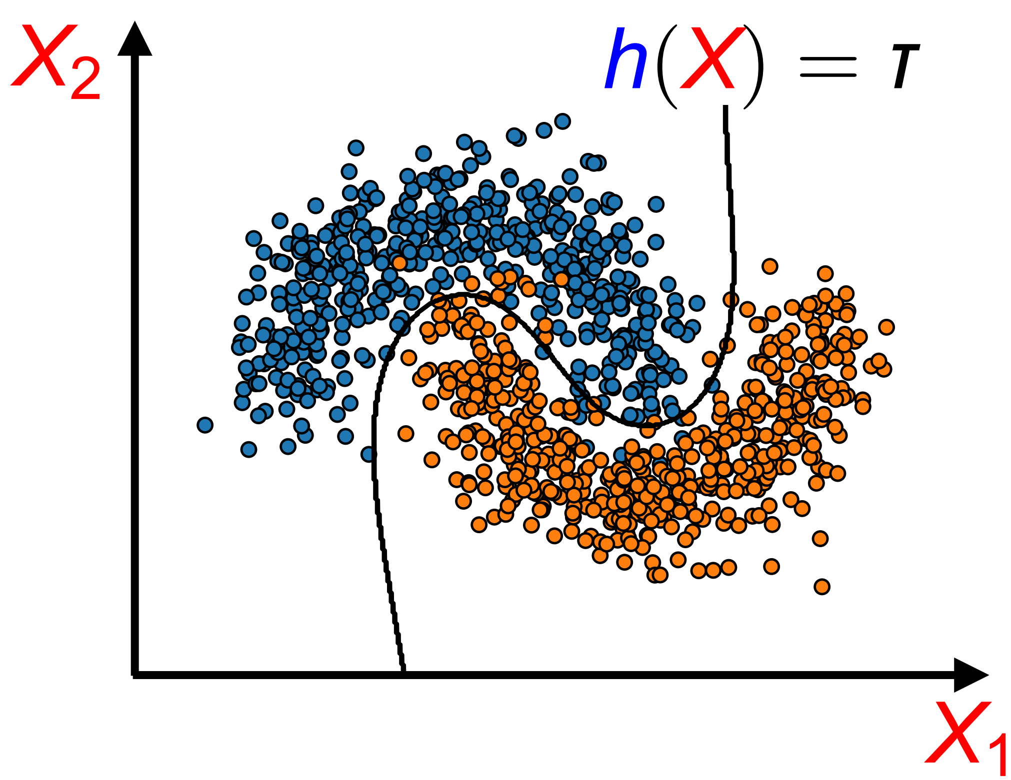


Regression

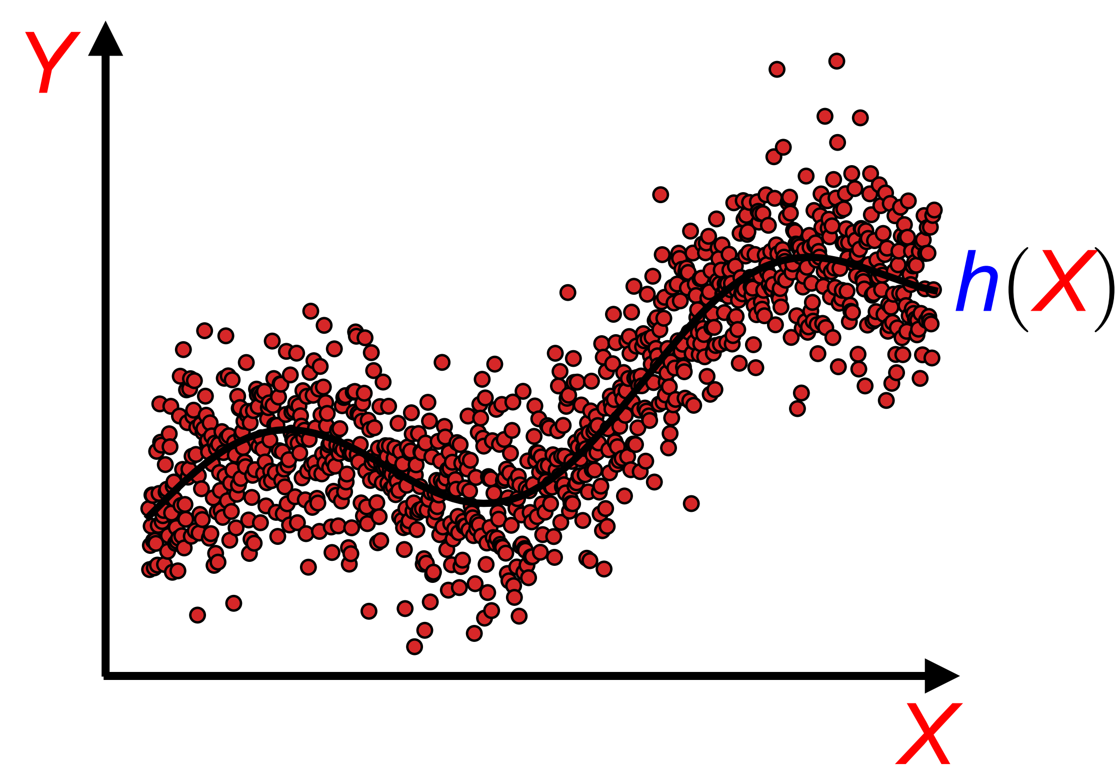
Statistical Learning

Learning problem: $\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)]$

$X \in \mathbb{R}^d$ input



Classification



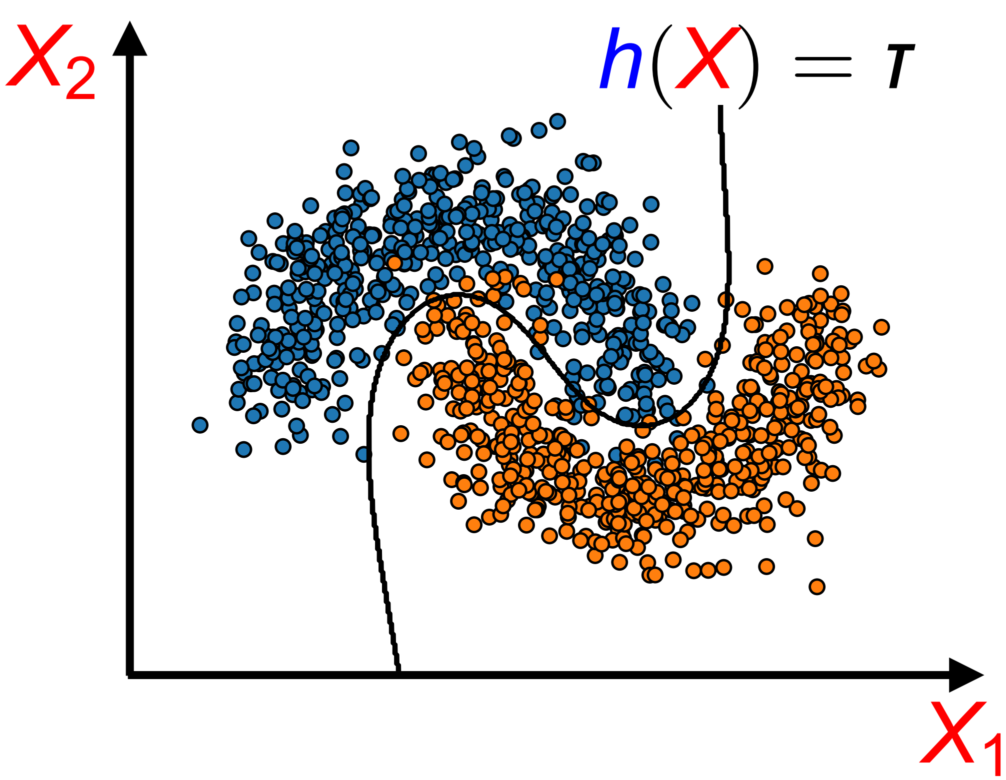
Regression

Statistical Learning

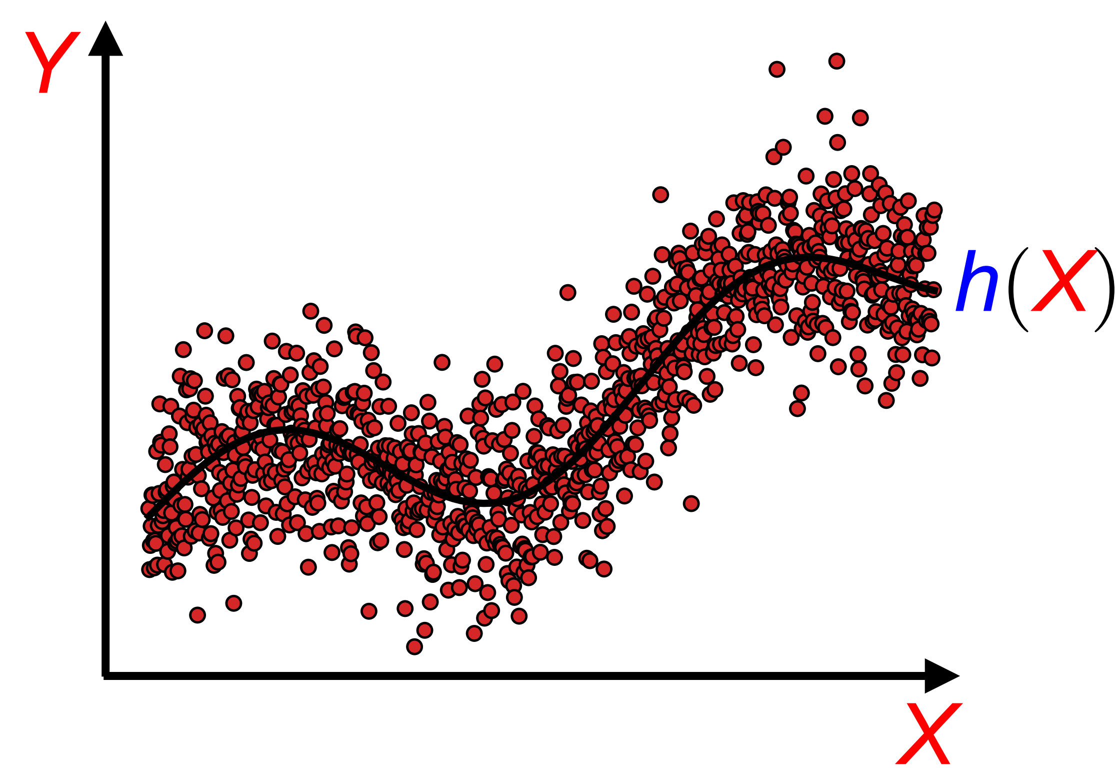
Learning problem: $\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)]$

$Y \in \mathbb{R}$ output

$X \in \mathbb{R}^d$ input



Classification



Regression

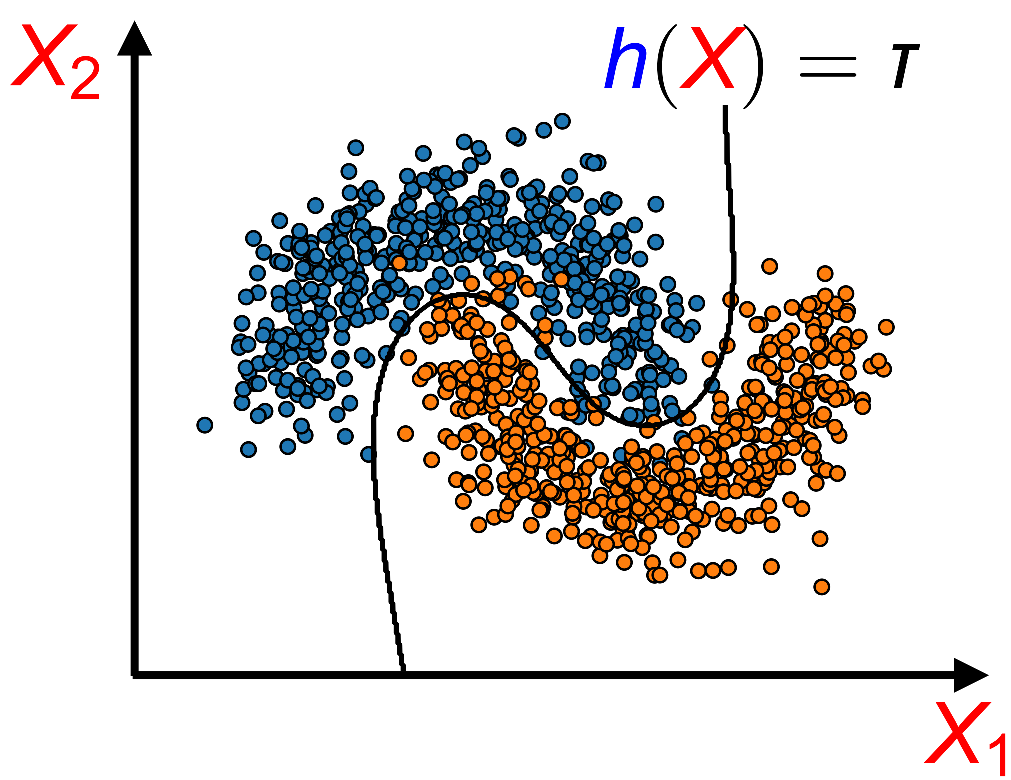
Statistical Learning

Learning problem: $\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)]$

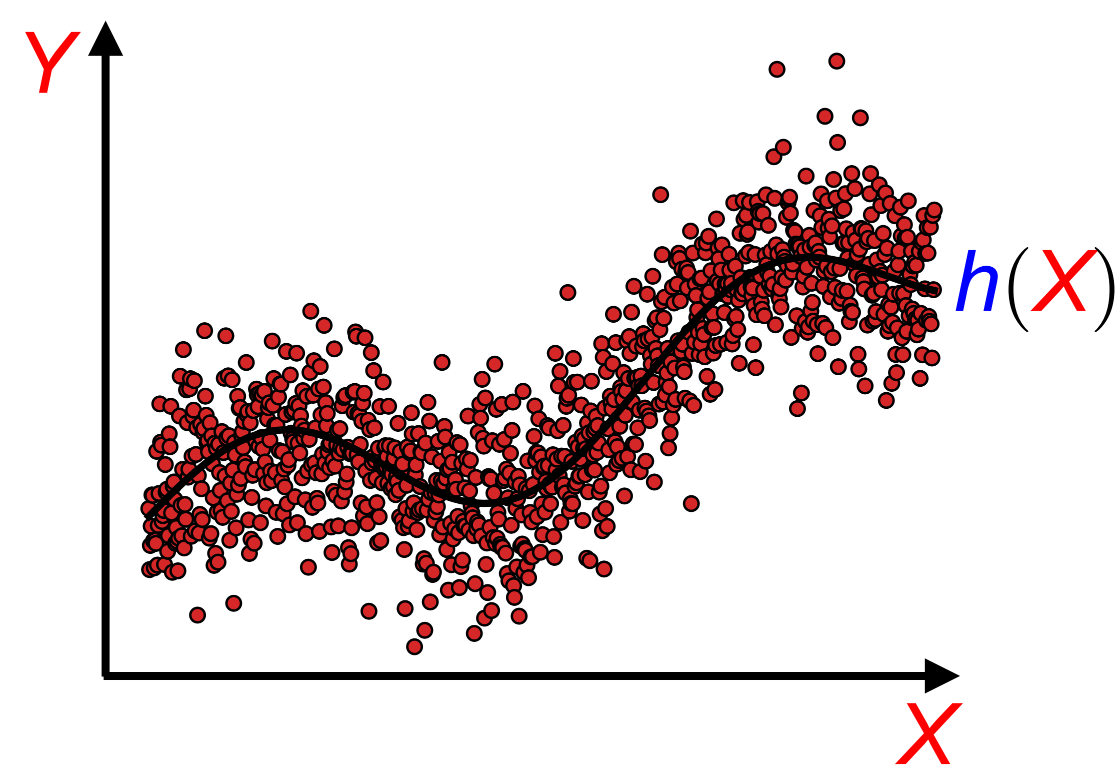
$Y \in \mathbb{R}$ output

hypothesis (predictor)

$X \in \mathbb{R}^d$ input



Classification



Regression

Statistical Learning

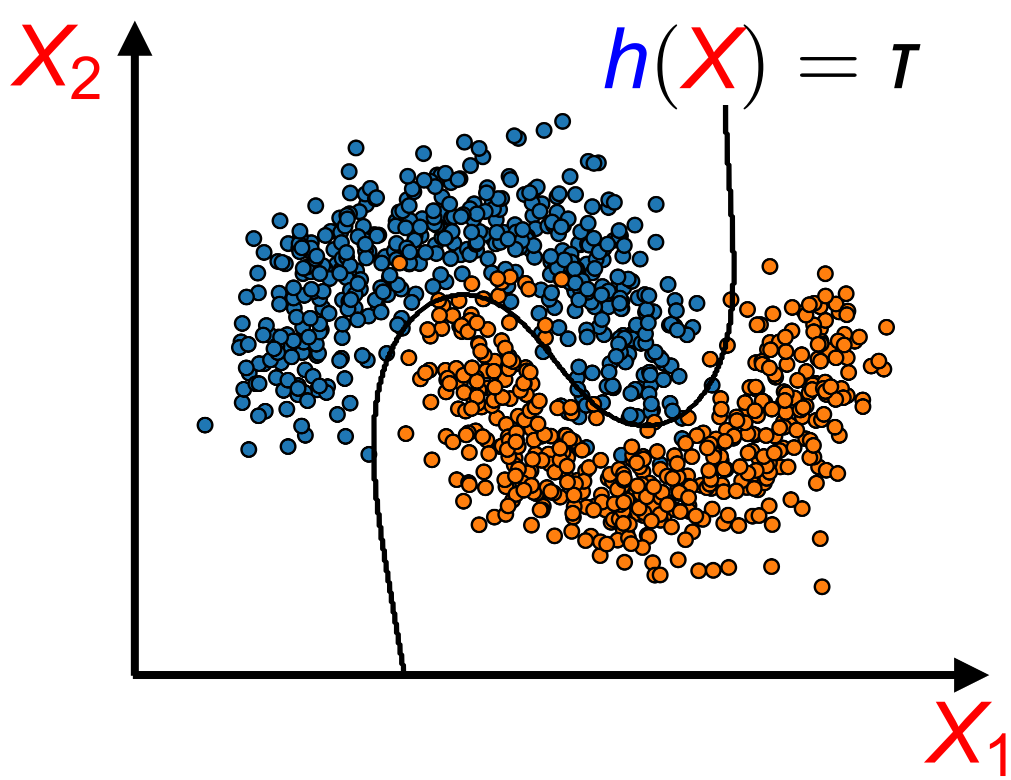
Learning problem: $\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)]$

penalizes dissimilarity of $h(X)$ and Y

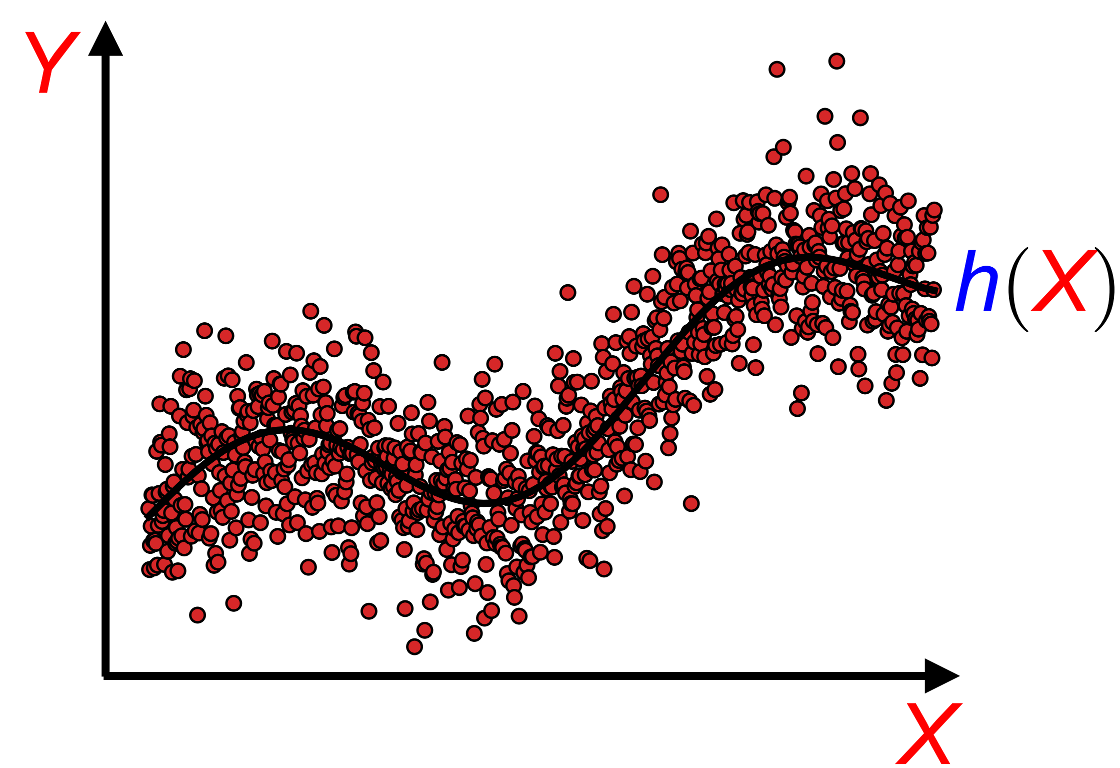
hypothesis (predictor)

$X \in \mathbb{R}^d$ input

$Y \in \mathbb{R}$ output



Classification



Regression

Weapons of Math Destruction

Learning problem: $\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)]$

- X = web browsing history , Y = consumer behavior
- X = credit history, Y = creditworthiness
- X = crime history, Y = recidivism
- X = résumé , Y = skills

Weapons of Math Destruction

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates. Photograph: Brian Snyder/Reuters

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

The data on which the AI hiring algorithm was trained created a preference for male candidates.[1)]

Industry-wide approach affecting millions of patients exhibits significant racial bias.[2)]

---

1) Dastin, *Reuters*, 2018.
2) Obermayer et al., *Science*, 2019.

Sensitive Attributes

X contains a sensitive attribute $A \in \{0, 1\}$ such as:

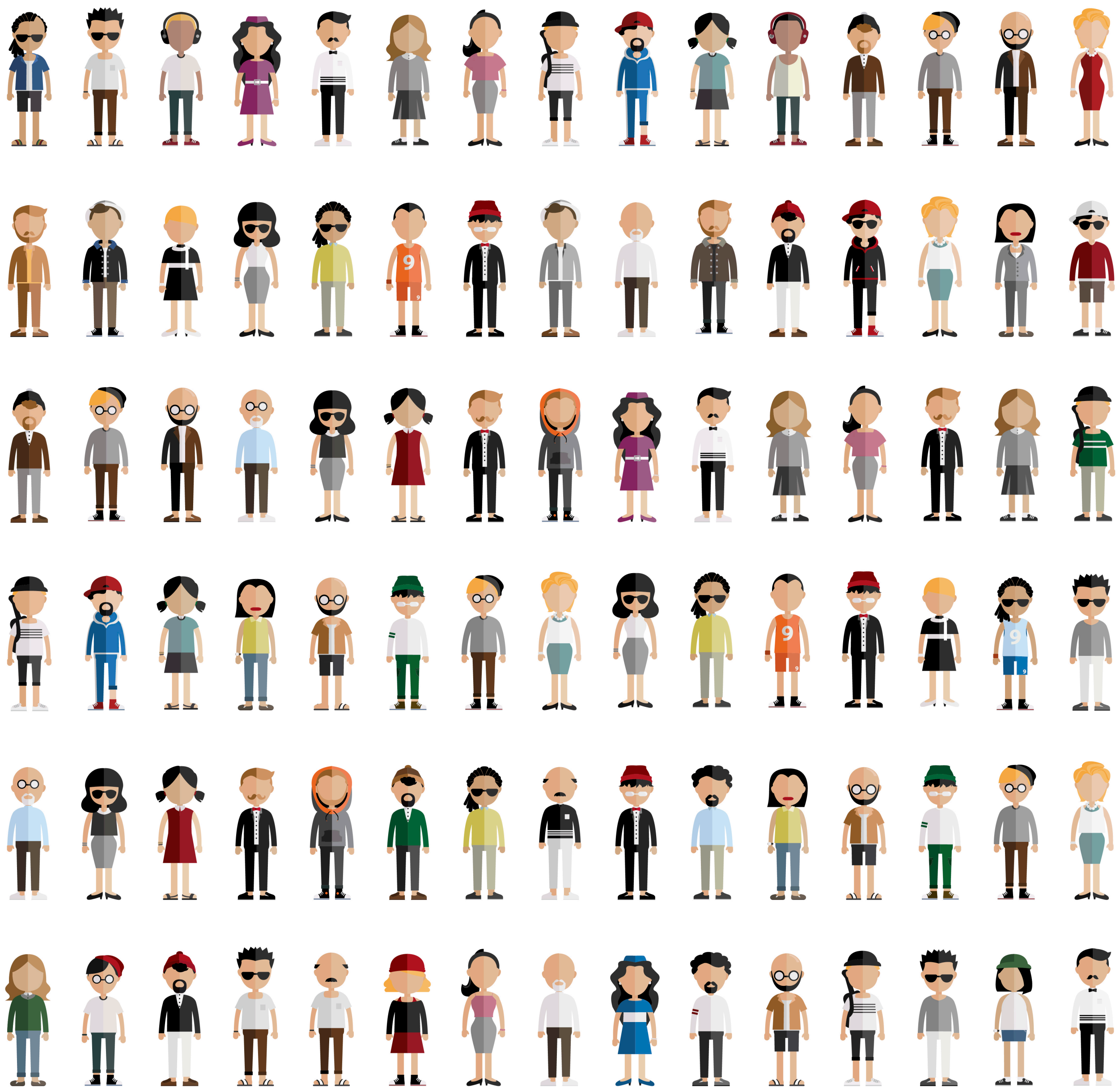


Image by rawpixel.com / Freepik

- Gender
- Ethnicity
- Religion
- Age
- Marital Status
- Citizenship

Fairness Through Unawareness

Idea: Remove A from X

Fairness Through Unawareness

Idea: Remove A from X

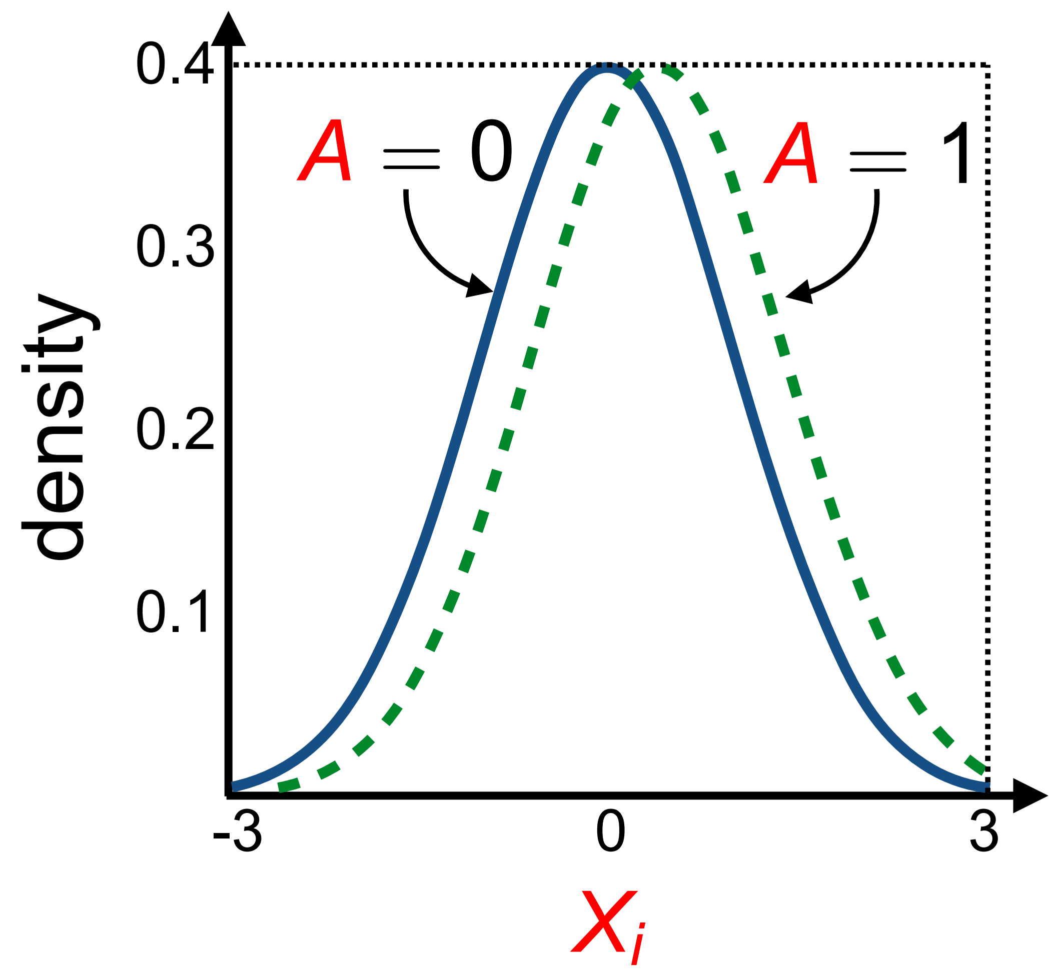
Problem:[1] Can use other features $X_1, X_2, X_3, \ldots$ to predict A

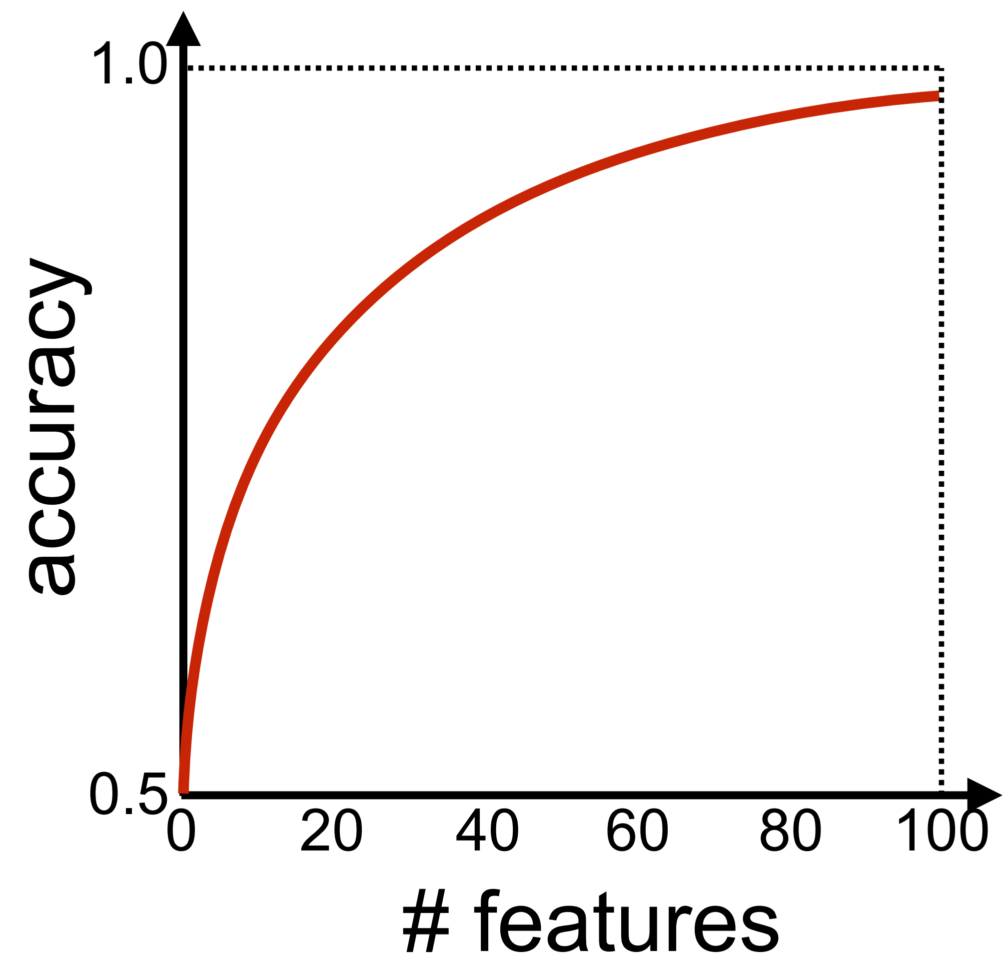
[1] Barocas, Hardt & Narayanan, *fairmlbook.org*, 2019.

Fairness Through Unawareness

Idea: Remove A from X

Problem:[1] Can use other features $X_1, X_2, X_3, \ldots$ to predict A





[1] Barocas, Hardt & Narayanan, *fairmlbook.org*, 2019.

Statistical parity:[1]

$$h(X) \mid A = 0 \;\sim\; h(X) \mid A = 1$$

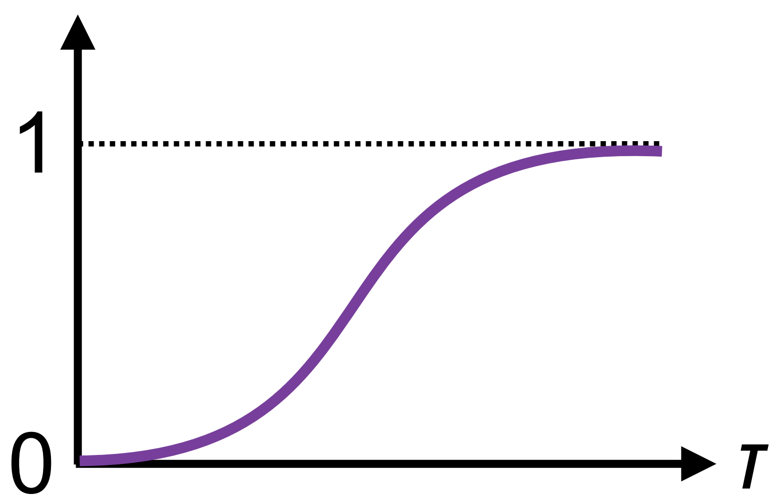
[1] Calders et al., *ICDM*, 2013.

Fairness

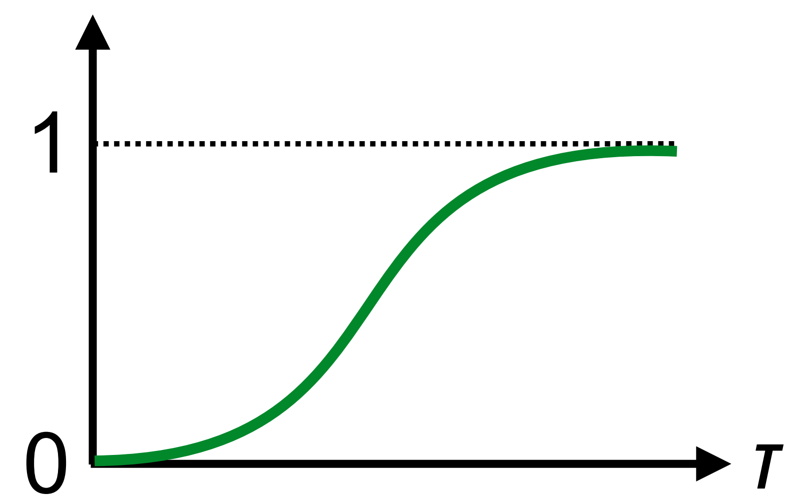
Statistical parity:[1]

$$\mathbb{P}(h(X) \leq \tau | A = 0) \; = \; \mathbb{P}(h(X) \leq \tau | A = 1) \quad \forall \tau \in \mathbb{R}$$

CDF of predictor:



$$\mathbb{P}(h(X) \leq \tau | A = 0)$$



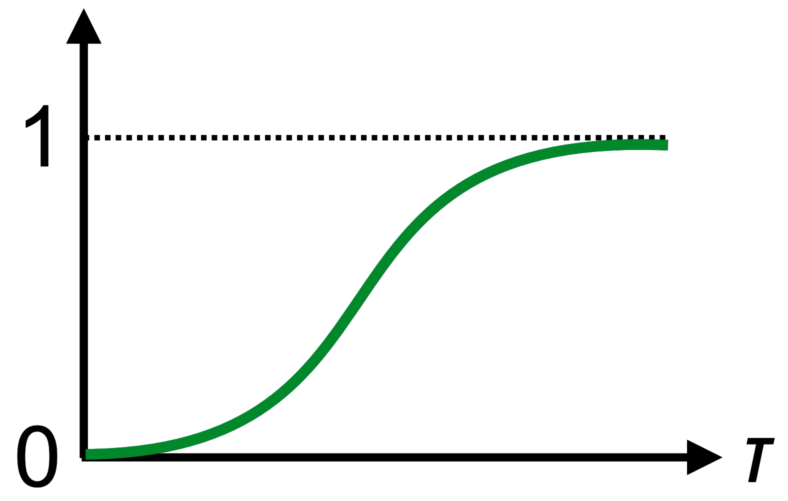
$$\mathbb{P}(h(X) \leq \tau | A = 1)$$

[1] Calders et al., *ICDM*, 2013.

Statistical parity:[1)]

$$\mathbb{P}(h(X) \leq \tau | A = 0) \; = \; \mathbb{P}(h(X) \leq \tau | A = 1) \quad \forall \tau \in \mathbb{R}$$

CDF of predictor:

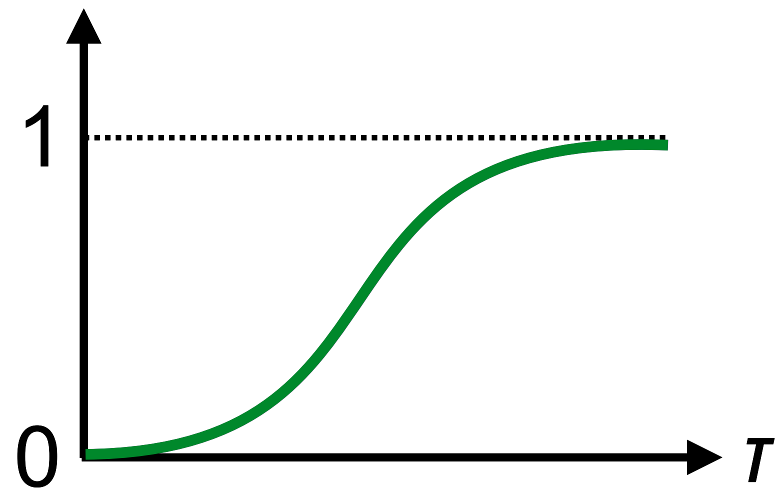


1) Calders et al., *ICDM*, 2013.

Statistical parity:[1)]

$$h(X) \perp A$$

CDF of predictor:



1) Calders et al., *ICDM*, 2013.

Other Group Fairness Definitions

Statistical parity:[1)]

$$h(X) \mid A = 0 \sim h(X) \mid A = 1$$

Equalized odds:[2)]

$$h(X) \mid Y = y, A = 0 \sim h(X) \mid Y = y, A = 1 \quad \forall y \in \mathbb{R}$$

Risk parity:[3)]

$$L(h(X), Y) \mid A = 0 \sim L(h(X), Y) \mid A = 1$$

---

1) Dwork et al., *ITCS*, 2012.
2) Hardt et al., *NeurIPS*, 2016.
3) Donini et al., *NeurIPS*, 2018.

Conceptual Analysis of Statistical Parity

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[L(h(X), Y)]$$
$$\text{s.t.} \quad h(X) \perp A$$

Fair Statistical Learning

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[L(h(X), Y)]$$
$$\text{s.t.} \quad h(X) \perp A$$

SP fairness

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[L(h(X), Y)]$$
$$\text{s.t.} \quad h(X) \perp A$$

SP fairness

$$\iff \mathbb{P}(h(X) \leq \tau | A = 0) = \mathbb{P}(h(X) \leq \tau | A = 1) \ \forall \tau \in \mathbb{R}$$

$\implies$ reminiscent of chance constraint

$\implies$ intractable

Idealized Models

Assumptions:

- $\mathbb{P}$ is known (sample size $= \infty$)
- $\mathcal{H} = \mathcal{L}(\mathbb{R}^d, \mathbb{R})$ (all measurable hypotheses)
- $h^\star$ is essentially unique

Automatic Salary Determination

Goal: Predict the skill levels of job candidates.

X_1	$=$ GPA	(normalized to $[0, 1]$)
X_2	$=$ age group	(0: age $>$ 40, 1: age $\leq$ 40)
Y	$=$ skill level	(normalized to $[0, 1]$)
S	$=$ work experience	(normalized to $[0, 1]$, unobserved)

Automatic Salary Determination

Goal: Predict the skill levels of job candidates.

X_1	$=$ GPA	(normalized to $[0, 1]$)
X_2	$=$ age group	(0: age $>$ 40, 1: age $\leq$ 40)
Y	$=$ skill level	(normalized to $[0, 1]$)
S	$=$ work experience	(normalized to $[0, 1]$, unobserved)

$$X_1, S \sim \mathcal{U}([0,1]), \quad X_2 \sim \mathcal{U}(\{0,1\}), \quad Y = X_1 \cdot X_2 + S \cdot (1 - X_2)$$

Automatic Salary Determination

Goal: Predict the skill levels of job candidates.

X_1 = GPA (normalized to $[0, 1]$)
X_2 = age group (0: age $>$ 40, 1: age $\leq$ 40)
Y = skill level (normalized to $[0, 1]$)
S = work experience (normalized to $[0, 1]$, unobserved)

$$X_1, S \sim \mathcal{U}([0, 1]), \quad X_2 \sim \mathcal{U}(\{0, 1\}), \quad Y = X_1 \cdot X_2 + S \cdot (1 - X_2)$$

$A = X_2$ = sensitive attribute

Automatic Salary Determination

Learning problem with square loss $L(\widehat{y}, y) = (\widehat{y} - y)^2$:

Original learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$
$$\text{s.t.} \quad h(X) \perp A$$

Automatic Salary Determination

Learning problem with square loss $L(\hat{y}, y) = (\hat{y} - y)^2$:

Original learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$
$$\text{s.t.} \quad h(X) \perp A$$

$$\implies h^\star(X) = \begin{cases} \frac{1}{2} & \text{if } X_2 = 0 \\ X_1 & \text{if } X_2 = 1 \end{cases}$$

Automatic Salary Determination

Learning problem with square loss $L(\widehat{y}, y) = (\widehat{y} - y)^2$:

Original learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$
$$\text{s.t.} \quad h(X) \perp A$$

$$\implies h^\star(X) = \begin{cases} \frac{1}{2} & \text{if } X_2 = 0 \\ X_1 & \text{if } X_2 = 1 \end{cases}$$

$$\implies h^\star_{\text{SP}}(X) = \frac{1}{2} + \frac{1}{2}(X_1 - \frac{1}{2})$$

Automatic Salary Determination

Learning problem with square loss $L(\widehat{y}, y) = (\widehat{y} - y)^2$:

Original learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$
$$\text{s.t.} \quad h(X) \perp A$$

$$\implies h^\star(X) = \begin{cases} \frac{1}{2} & \text{if } X_2 = 0 \\ X_1 & \text{if } X_2 = 1 \end{cases}$$

$$\implies h^\star_{\text{SP}}(X) = \frac{1}{2} + \frac{1}{2}(X_1 - \frac{1}{2})$$

$$\implies \mathbb{E}[(h^\star(X) - Y)^2] = \frac{1}{24} < \frac{91}{96} = \mathbb{E}[(h^\star_{\text{SP}}(X) - Y)^2]$$

Automatic Salary Determination

Learning problem with square loss $L(\widehat{y}, y) = (\widehat{y} - y)^2$:

Original learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$
$$\text{s.t.} \quad h(X) \perp A$$

$$\Longrightarrow \; h^\star(X) = \begin{cases} \frac{1}{2} & \text{if } X_2 = 0 \\ X_1 & \text{if } X_2 = 1 \end{cases}$$

$$\Longrightarrow \; h^\star_{\text{SP}}(X) = \frac{1}{2} + \frac{1}{2}\left(X_1 - \frac{1}{2}\right)$$

SP deteriorates predictive power!

Automatic Salary Determination

Original learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$
$$\text{s.t.} \quad h(X) \perp A$$

$$\Longrightarrow \quad h^\star(X) = \begin{cases} \frac{1}{2} & \text{if } X_2 = 0 \\ X_1 & \text{if } X_2 = 1 \end{cases}$$

$$\Longrightarrow \quad h^\star_{\text{SP}}(X) = \frac{1}{2} + \frac{1}{2}(X_1 - \frac{1}{2})$$

$A = X_2 = 1$ (junior candidate) :

$$\Longrightarrow \quad h^\star(X) = Y$$

$$\Longrightarrow \quad h^\star_{\text{SP}}(X) = \frac{1}{2} + \frac{1}{2}(Y - \frac{1}{2})$$

Automatic Salary Determination

Original learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$
$$\text{s.t.} \quad h(X) \perp A$$

$$\implies h^\star(X) = \begin{cases} \frac{1}{2} & \text{if } X_2 = 0 \\ X_1 & \text{if } X_2 = 1 \end{cases}$$

$$\implies h^\star_{\text{SP}}(X) = \tfrac{1}{2} + \tfrac{1}{2}(X_1 - \tfrac{1}{2})$$

$A = X_2 = 1$ (junior candidate) :

$$\implies h^\star(X) = Y$$

$$\implies h^\star_{\text{SP}}(X) = \tfrac{1}{2} + \tfrac{1}{2}(Y - \tfrac{1}{2})$$

salary grows with skill level

Automatic Salary Determination

Original learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$
$$\text{s.t.} \quad h(X) \perp A$$

$$\Longrightarrow \quad h^\star(X) = \begin{cases} \frac{1}{2} & \text{if } X_2 = 0 \\ X_1 & \text{if } X_2 = 1 \end{cases}$$

$$\Longrightarrow \quad h^\star_{\text{SP}}(X) = \frac{1}{2} + \frac{1}{2}(X_1 - \frac{1}{2})$$

$A = X_2 = 0$ (senior candidate) :

$$\Longrightarrow \quad h^\star(X) = \frac{1}{2}$$

$$\Longrightarrow \quad h^\star_{\text{SP}}(X) = \frac{1}{2} + \frac{1}{2}(X_1 - \frac{1}{2})$$

Automatic Salary Determination

Original learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$
$$\text{s.t.} \quad h(X) \perp A$$

$$\Longrightarrow \quad h^\star(X) = \begin{cases} \frac{1}{2} & \text{if } X_2 = 0 \\ X_1 & \text{if } X_2 = 1 \end{cases}$$

$$\Longrightarrow \quad h^\star_{\text{SP}}(X) = \frac{1}{2} + \frac{1}{2}(X_1 - \frac{1}{2})$$

$A = X_2 = 0$ (senior candidate) :

$$\Longrightarrow \quad h^\star(X) = \frac{1}{2}$$

uniform salary

$$\Longrightarrow \quad h^\star_{\text{SP}}(X) = \frac{1}{2} + \frac{1}{2}(X_1 - \frac{1}{2})$$

random salary

Automatic Salary Determination

Original learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \quad \mathbb{E}[(h(X) - Y)^2]$$
$$\text{s.t.} \quad h(X) \perp A$$

$$\implies h^\star(X) = \begin{cases} \frac{1}{2} & \text{if } X_2 = 0 \\ X_1 & \text{if } X_2 = 1 \end{cases}$$

$$\implies h^\star_{SP}(X) = \frac{1}{2} + \frac{1}{2}(X_1 - \frac{1}{2})$$

$A = X_2 = 0$ (senior candidate) :

Treatment of senior candidates seems less "fair" when SP is enforced!

random salary

Optimality Implies Statistical Parity

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)]$$

Theorem. $\mathbb{P}_{Y|X} \perp A \implies h^\star(X) \perp A$

Optimality Implies Statistical Parity

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)]$$

Theorem. $\mathbb{P}_{Y|X} \perp A \implies h^\star(X) \perp A$

$\implies$ SP is a necessary optimality condition!

Training with Biased Data

True learning problem:

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y_0)]$$

Biased learning problem:

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y_\delta)]$$

Training with Biased Data

True learning problem:

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y_0)]$$

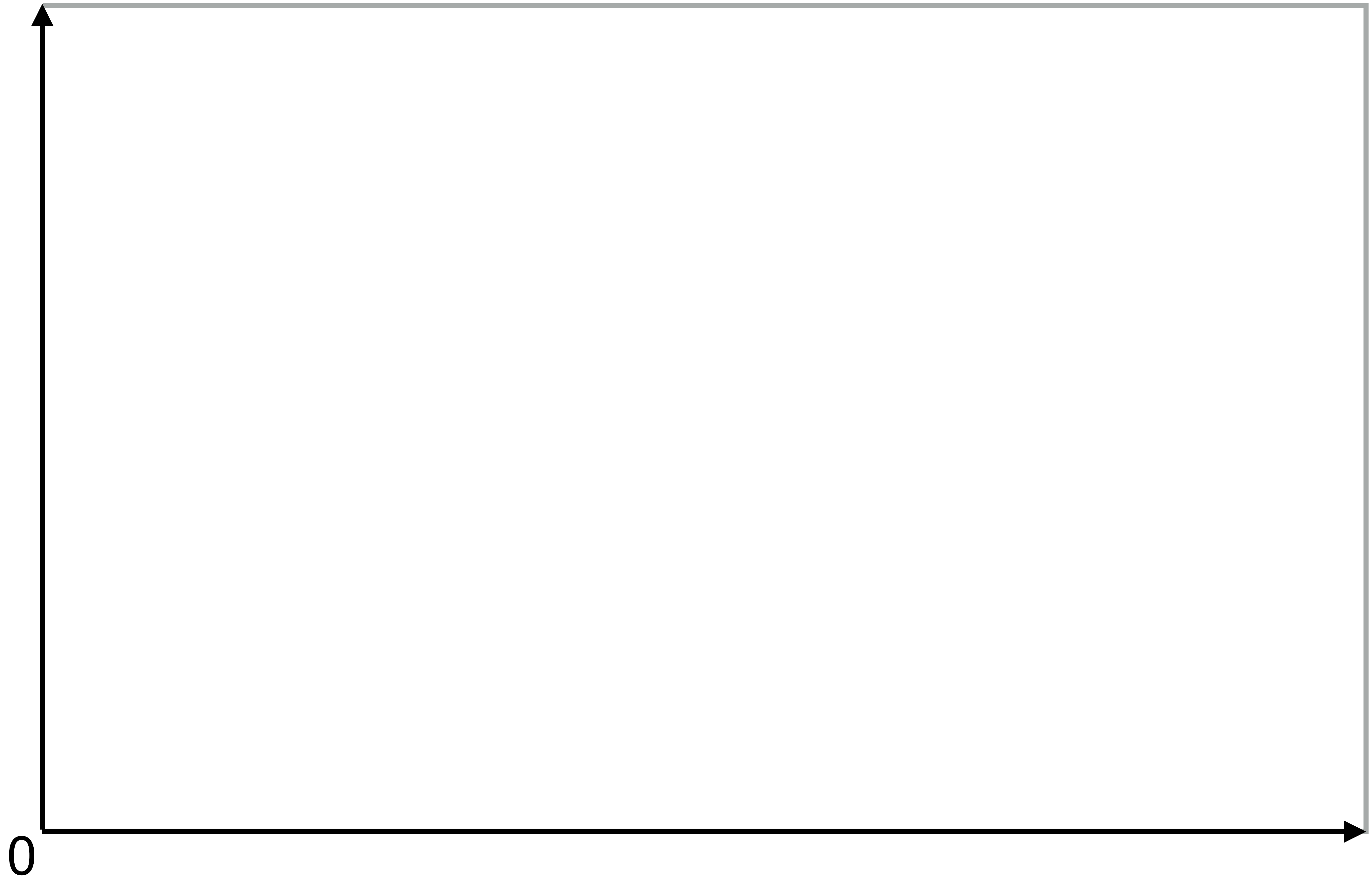
true target
(no data available)

Biased learning problem:

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y_\delta)]$$

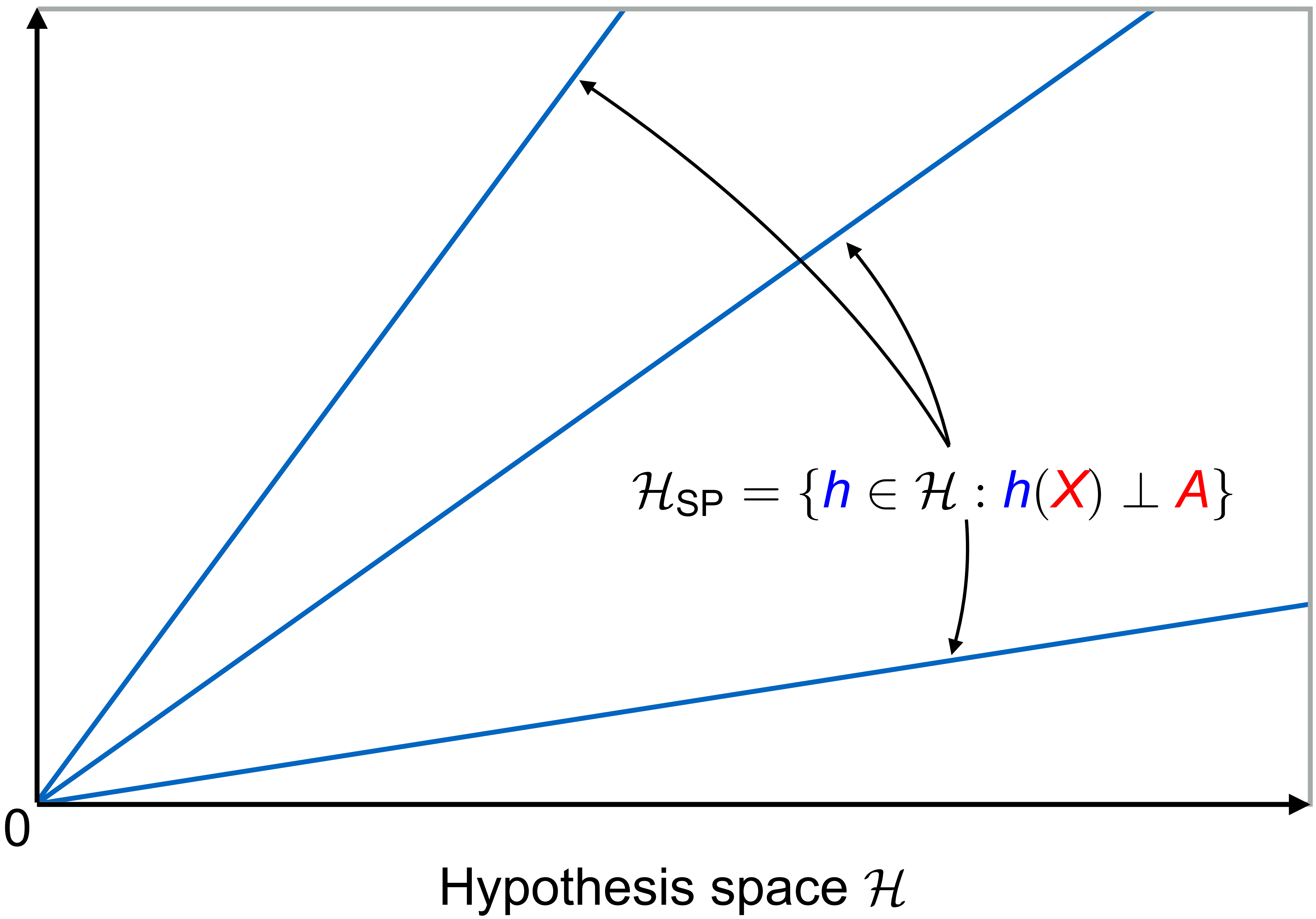
biased target
(data available)

The Geometry of Statistical Parity



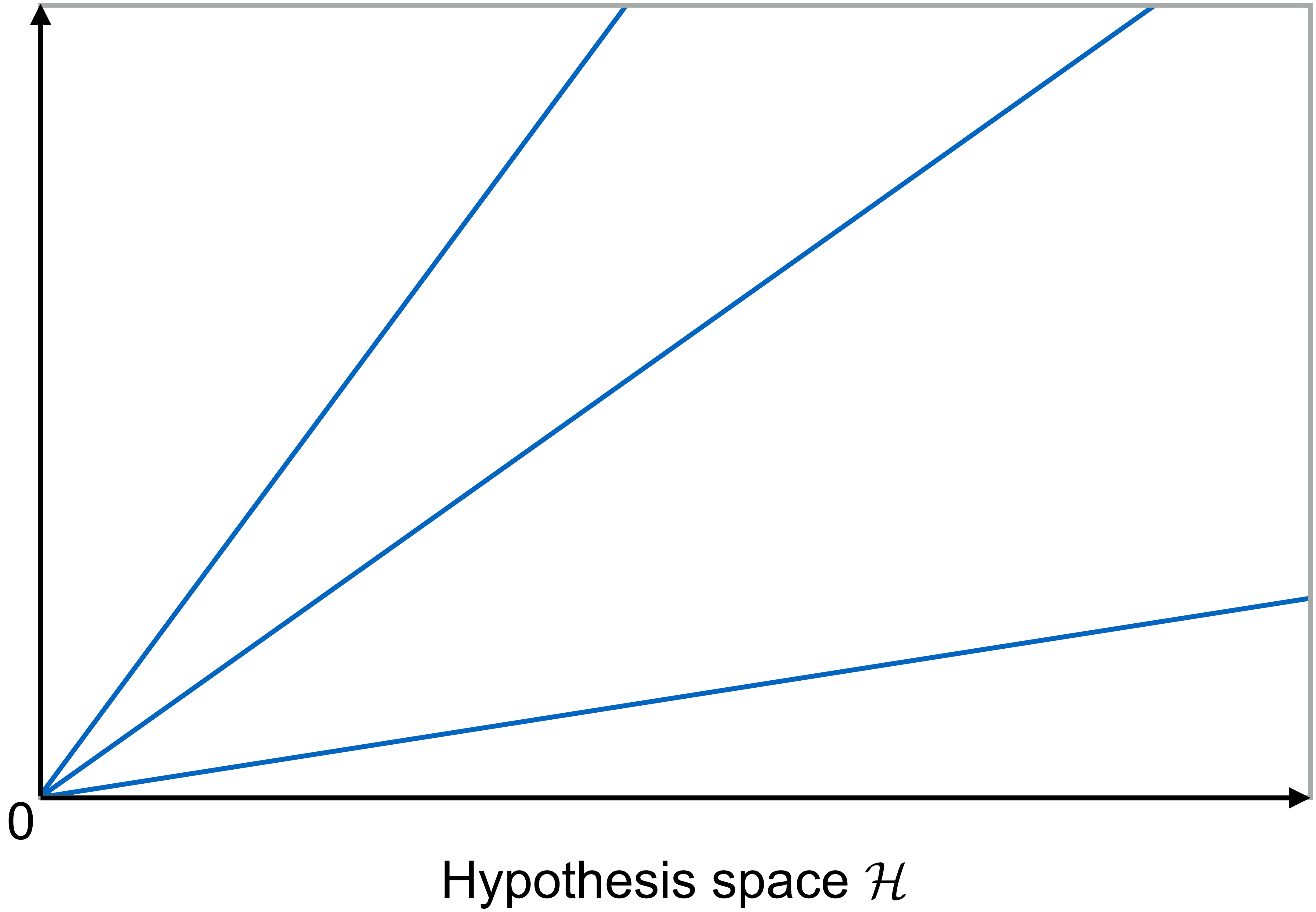
Hypothesis space $\mathcal{H} = \mathcal{L}(\mathbb{R}^d, \mathbb{R})$

The Geometry of Statistical Parity



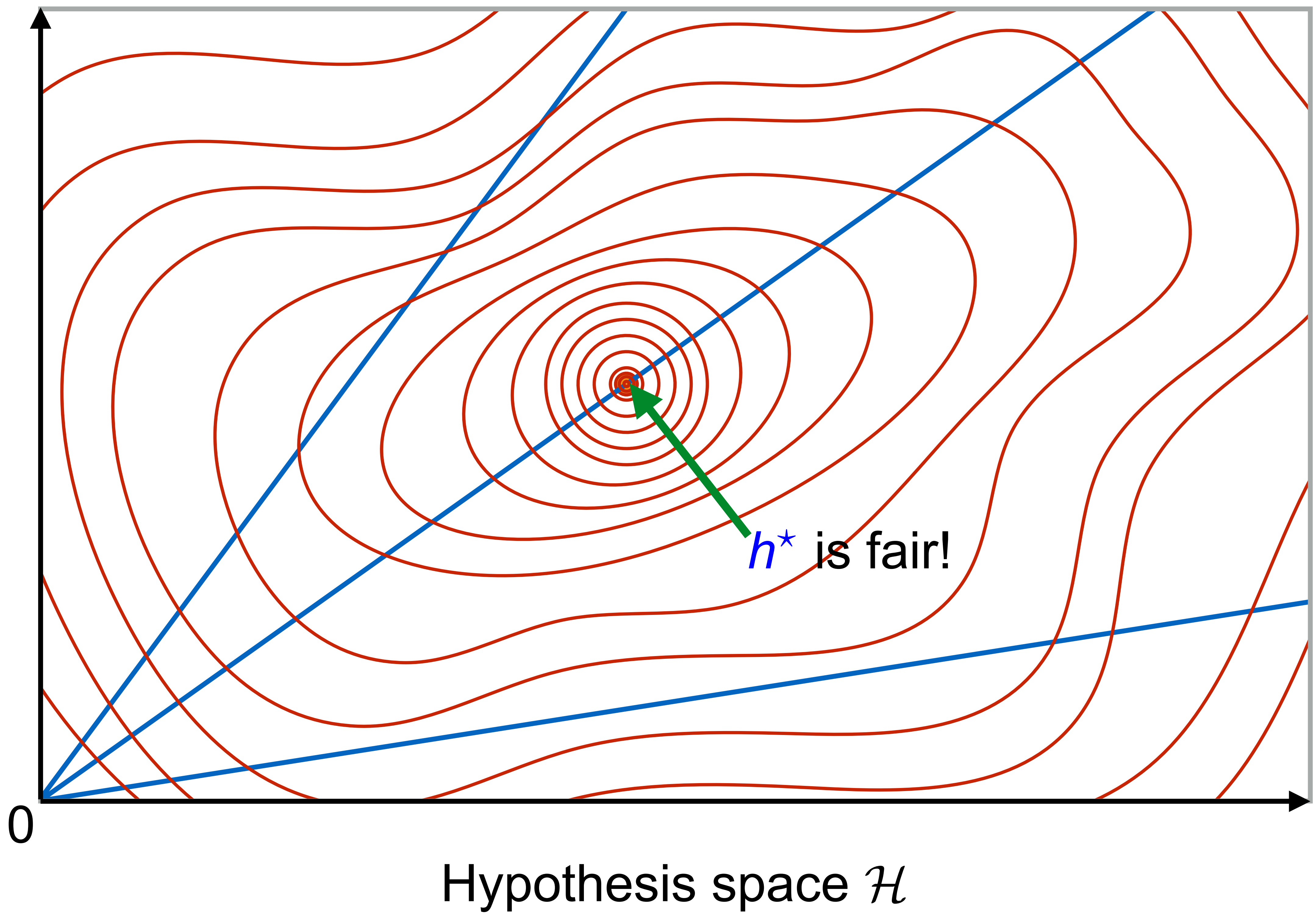
The Geometry of Statistical Parity

Assumption: $\mathbb{P}_{Y_0|X} \perp A$



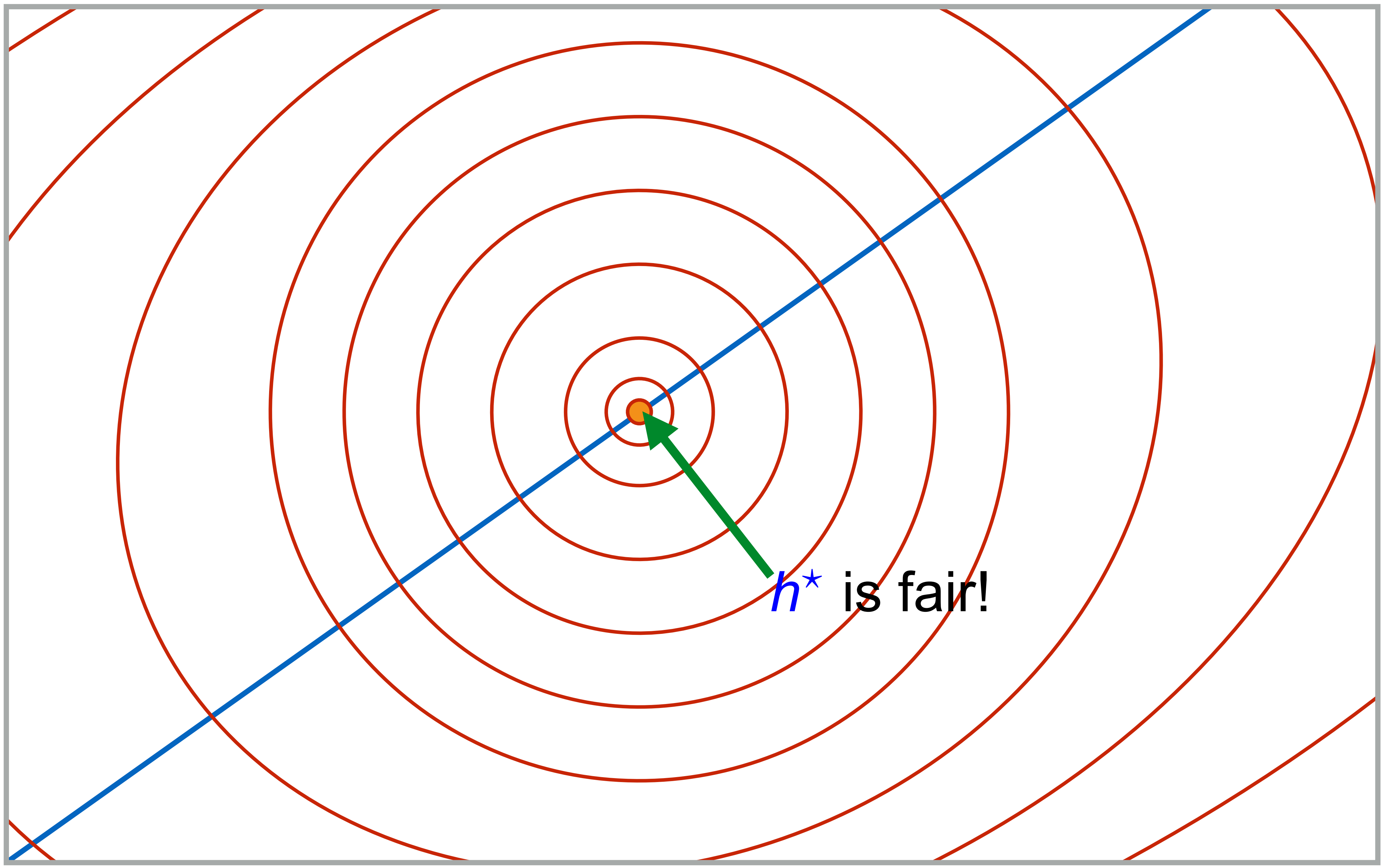
The Geometry of Statistical Parity

Prediction loss $= f(h) = \mathbb{E}[L(h(X, Y_0)]$



The Geometry of Statistical Parity

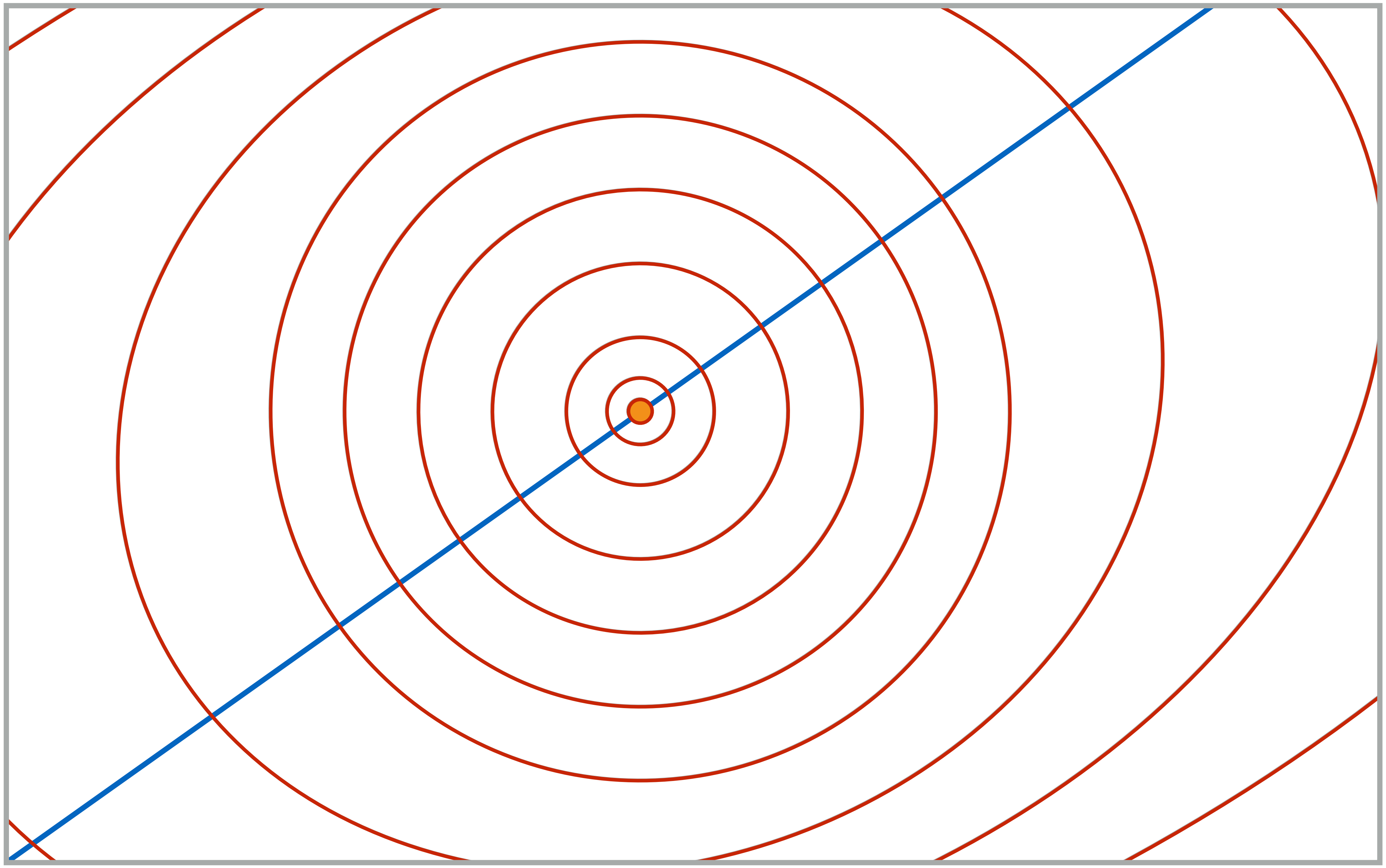
Prediction loss $= f(h) = \mathbb{E}[L(h(X, Y_0)]$



Hypothesis space $\mathcal{H}$

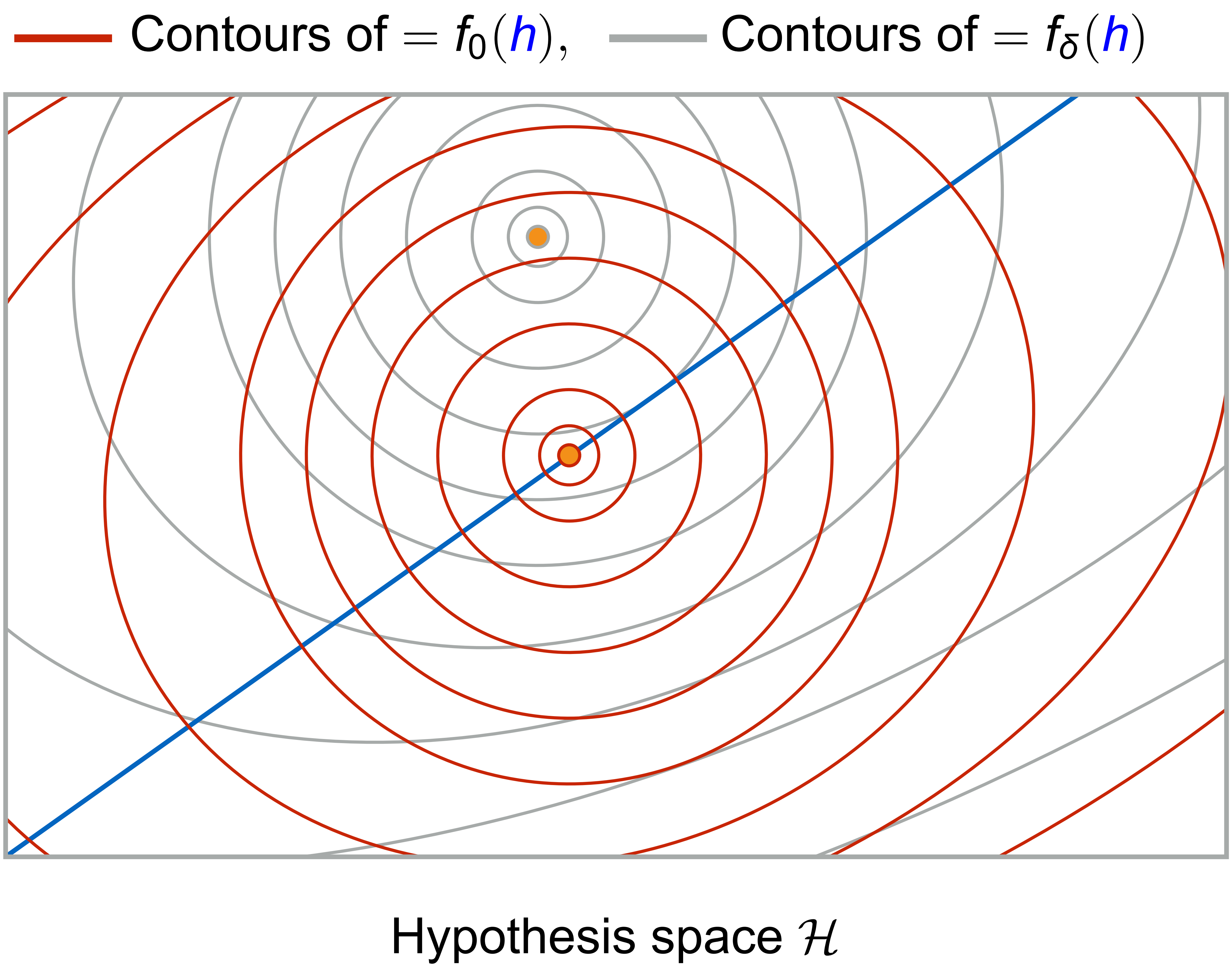
The Geometry of Statistical Parity

Biased prediction loss $= f_\delta(h) = \mathbb{E}[L(h(X, Y_\delta)]$

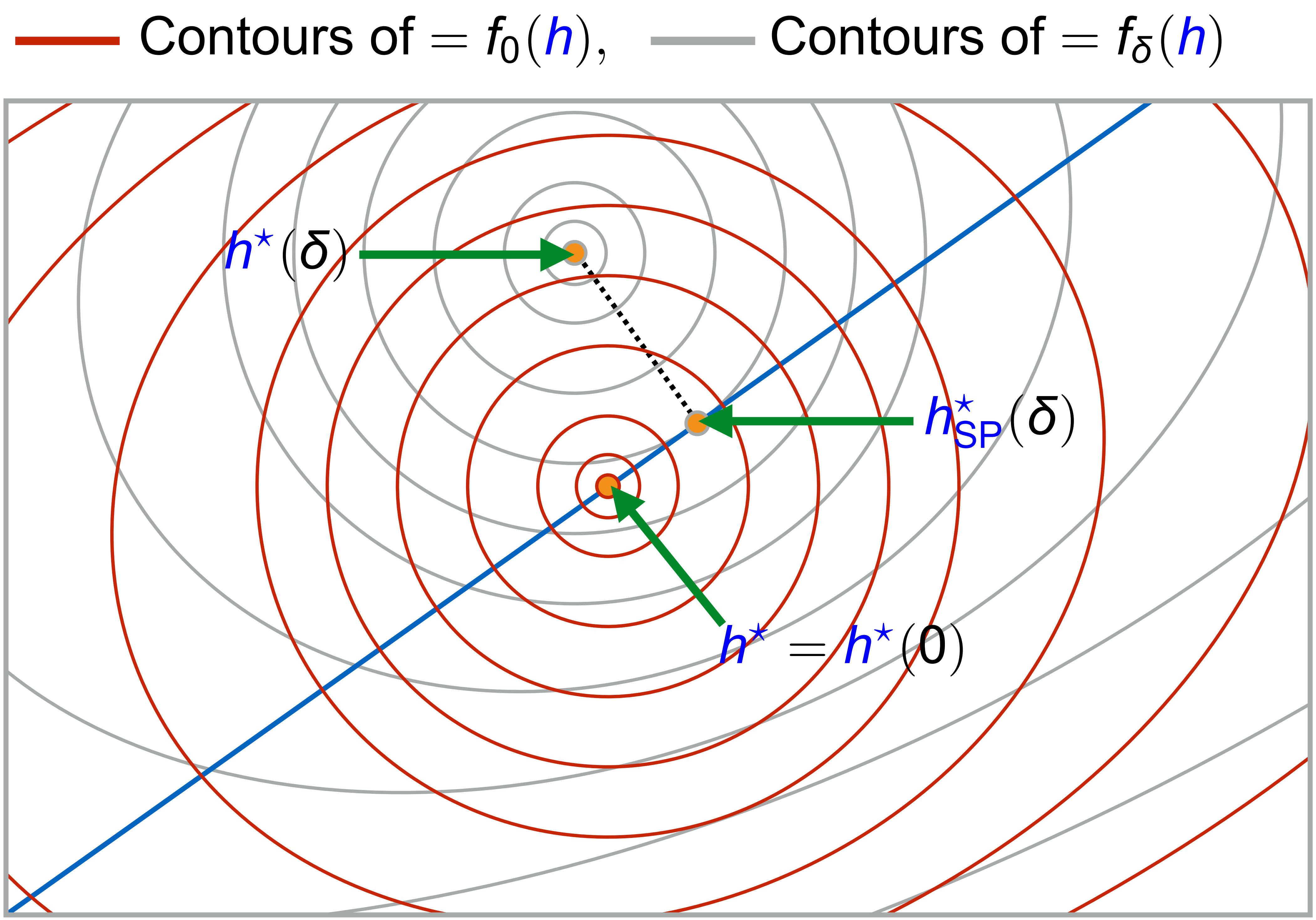


Hypothesis space $\mathcal{H}$

The Geometry of Statistical Parity



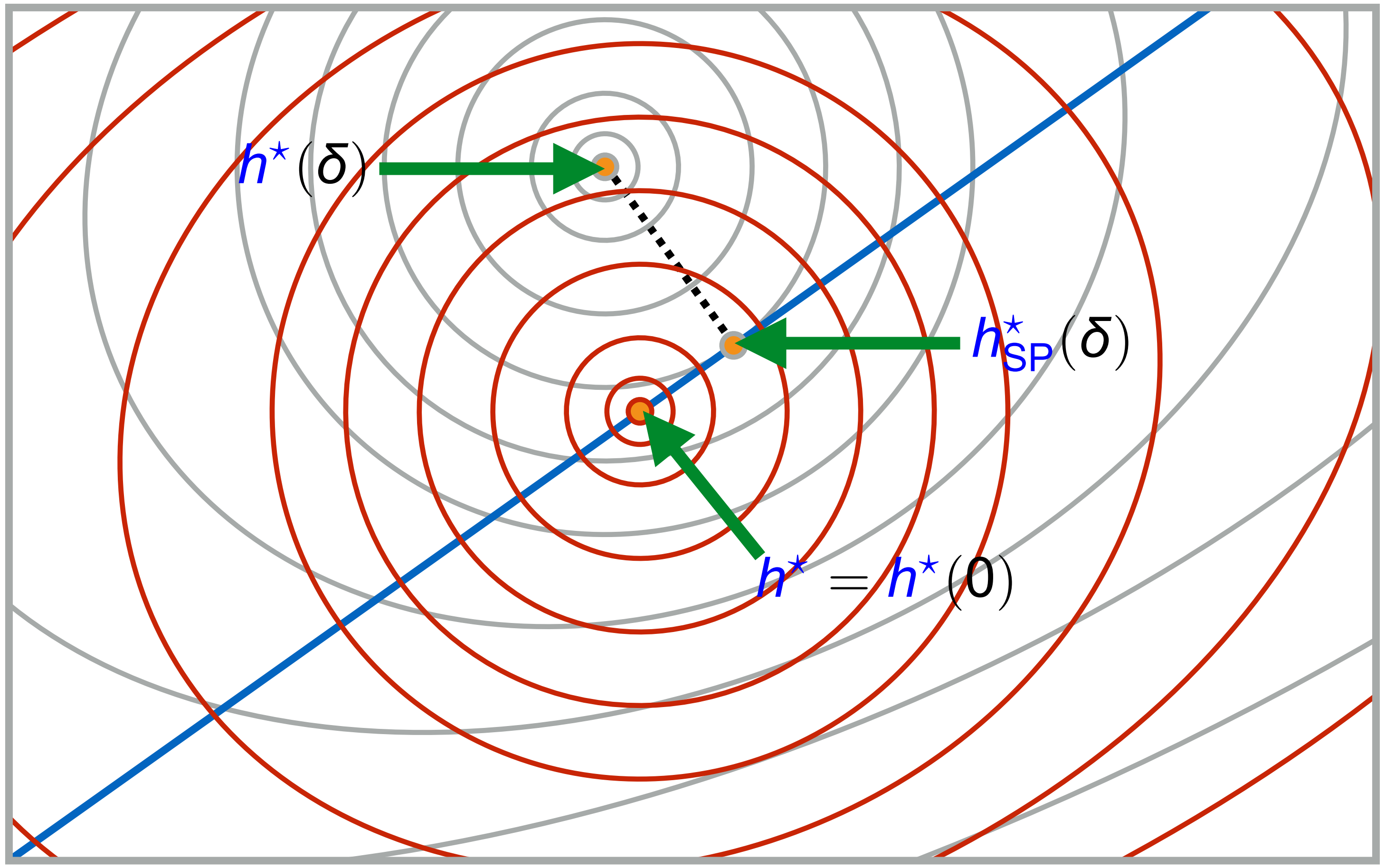
The Geometry of Statistical Parity



Hypothesis space $\mathcal{H}$

The Geometry of Statistical Parity

Theorem: If $\mathbb{P}_{Y_0|X} \perp A$ and δ is small, then $h^\star_{\text{SP}}(\delta)$ is preferable to $h^\star(\delta)$ w.r.t. the true objective $f_0(h) = \mathbb{E}[L(h(X, Y_0)]$.



The Geometry of Statistical Parity

Theorem: If $\mathbb{P}_{Y_0|X} \perp A$ and δ is small, then $h^\star_{\text{SP}}(\delta)$ is preferable to $h^\star(\delta)$ w.r.t. the true objective $f_0(h) = \mathbb{E}[L(h(X, Y_0)]$.

Win-win situation:
Statistical parity improves both fairness and predictive power!

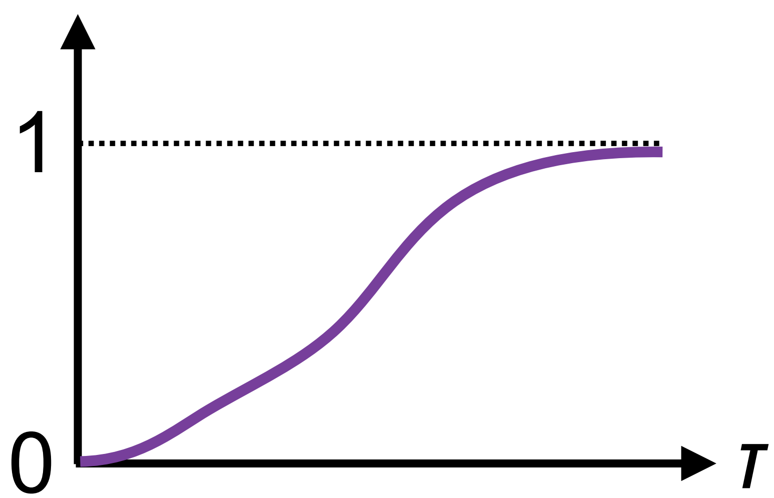
Unfairness Measures and Integral Probability Metrics

Relaxing Statistical Parity

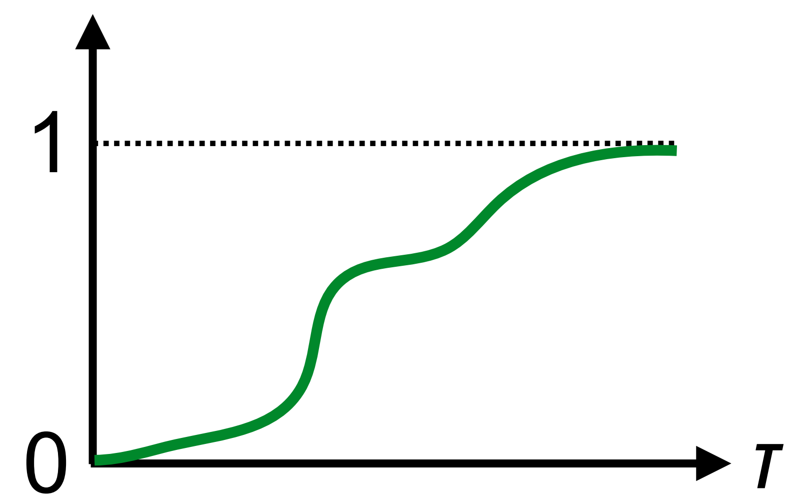
Statistical parity at level ε:

$$|\mathbb{P}(h(X) \leq \tau | A = 0) - \mathbb{P}(h(X) \leq \tau | A = 1)| \leq \varepsilon \quad \forall \tau \in \mathbb{R}$$

CDF of predictor:



$$\mathbb{P}(h(X) \leq \tau | A = 0)$$



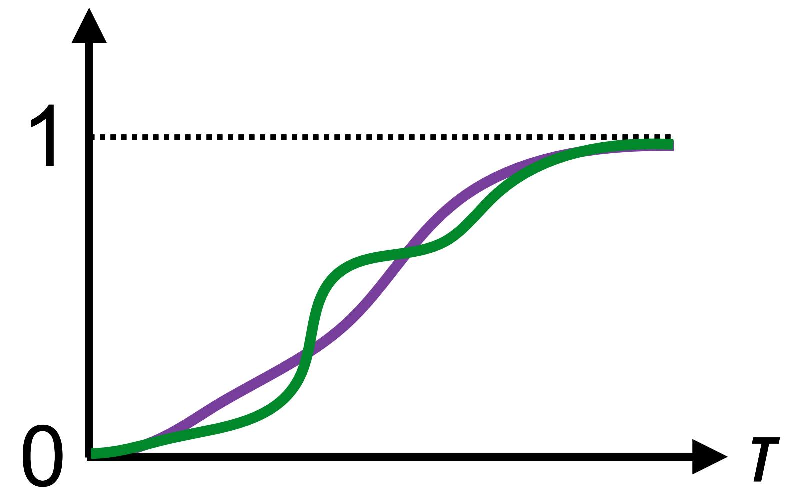
$$\mathbb{P}(h(X) \leq \tau | A = 1)$$

Relaxing Statistical Parity

Statistical parity at level ε:

$$|\mathbb{P}(h(X) \leq \tau | A = 0) - \mathbb{P}(h(X) \leq \tau | A = 1)| \leq \varepsilon \quad \forall \tau \in \mathbb{R}$$

CDF of predictor:

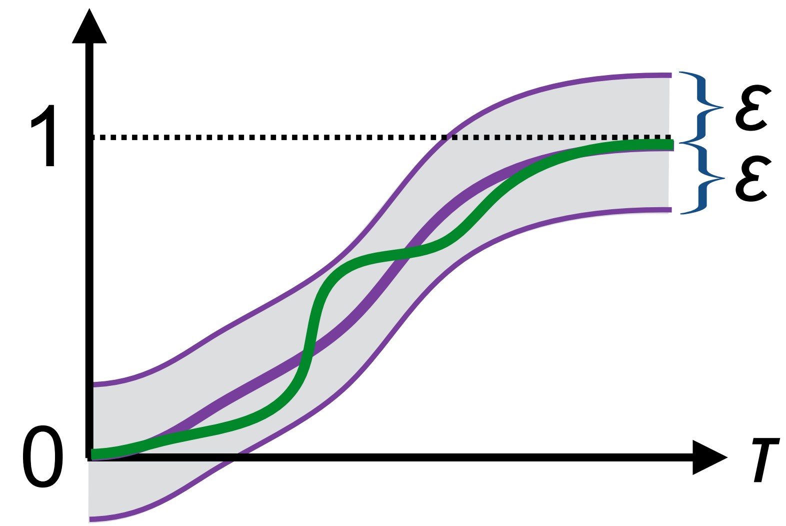


Relaxing Statistical Parity

Statistical parity at level ε:

$$|\mathbb{P}(h(X) \leq \tau | A = 0) - \mathbb{P}(h(X) \leq \tau | A = 1)| \leq \varepsilon \quad \forall \tau \in \mathbb{R}$$

CDF of predictor:



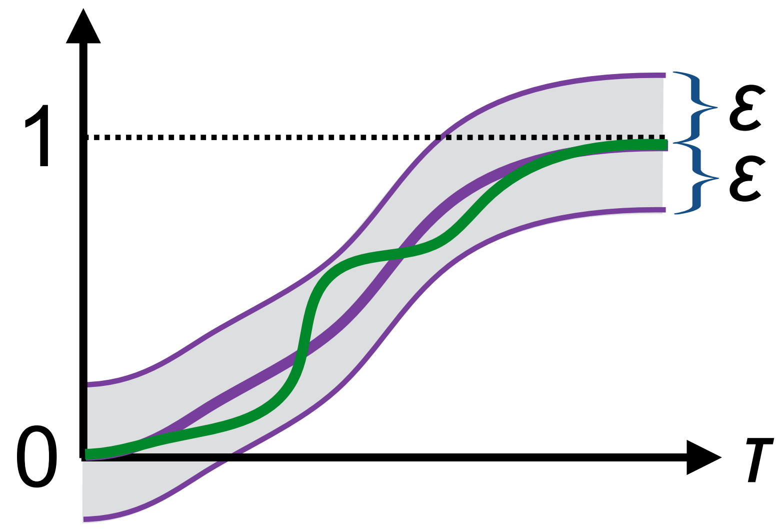
Relaxing Statistical Parity

Statistical parity at level ε:

$$\mathcal{D}\left(\mathbb{P}_{h(X)|A=0}, \mathbb{P}_{h(X)|A=1}\right) \leq \varepsilon$$

Kolmogorov distance

CDF of predictor:



Integral Probability Metrics (IPMs)

$$\mathcal{D}_\Psi(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{\psi \in \Psi} \int_{\mathbb{R}} \psi(y)\, \mathbb{Q}_1(\mathrm{d}y) - \int_{\mathbb{R}} \psi(y)\, \mathbb{Q}_2(\mathrm{d}y)$$

Integral Probability Metrics (IPMs)

$$\mathcal{D}_\Psi(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{\psi \in \Psi} \int_{\mathbb{R}} \psi(y)\, \mathbb{Q}_1(dy) - \int_{\mathbb{R}} \psi(y)\, \mathbb{Q}_2(dy)$$

↑

generator

Integral Probability Metrics (IPMs)

$$\mathcal{D}_\Psi(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{\psi \in \Psi} \int_{\mathbb{R}} \psi(y)\, \mathbb{Q}_1(\mathrm{d}y) - \int_{\mathbb{R}} \psi(y)\, \mathbb{Q}_2(\mathrm{d}y)$$

Examples:

IPM	Ψ
Kolmogorov distance	$\{\psi : \exists \tau \in \mathbb{R} \text{ with } \psi(y) = \pm 1_{y \le \tau}\}$
Wasserstein distance	$\{\psi : \mathrm{Lip}(\psi) \le 1\}$
$\mathcal{L}^p$-distance ($\frac{1}{p} + \frac{1}{q} = 1$)	$\{\psi : \|\psi'\|_{\mathcal{L}^q} \le 1\}$
Kernel distance	$\{\psi : \|\psi\|_{\mathbb{H}_K} \le 1\}$
Total variation distance	$\{\psi : \|\psi\|_{\mathcal{L}^\infty} \le 1\}$

Relaxing Statistical Parity

Statistical parity at level ε:

$$\mathcal{D}_\Psi \left(\mathbb{P}_{h(X)|A=0}, \mathbb{P}_{h(X)|A=1}\right) \leq \varepsilon$$

Relaxing Statistical Parity

Statistical parity at level ε:

$$\mathcal{D}_\Psi \left(\mathbb{P}_{h(X)|A=0}, \mathbb{P}_{h(X)|A=1}\right) \leq \varepsilon$$

any IPM

Fair learning problem:

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h(X)|A=0}, \mathbb{P}_{h(X)|A=1}))$$

Fair Statistical Learning

Fair learning problem:

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)] + \underbrace{\rho(\mathcal{D}_\Psi(\mathbb{P}_{h(X)|A=0}, \mathbb{P}_{h(X)|A=1}))}_{\text{unfairness penalty}}$$

unfairness penalty

Numerical Solution of Fair Learning Problems

Fair Statistical Learning

Fair learning problem:

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h(X)|A=0}, \mathbb{P}_{h(X)|A=1}))$$

Fair Statistical Learning

Fair learning problem:

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h(X)|A=0}, \mathbb{P}_{h(X)|A=1}))$$

$$\mathcal{H} = \{h_\theta : \theta \in \Theta\}$$

Fair learning problem:

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h(X)|A=0}, \mathbb{P}_{h(X)|A=1}))$$

$$\mathcal{H} = \{h_\theta : \theta \in \Theta\}$$

- all linear hypotheses
- all neural networks with a fixed architecture

Fair learning problem:

$$\min_{\theta \in \Theta} \mathbb{E}[L(h_\theta(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h_\theta(X)|A=0}, \mathbb{P}_{h_\theta(X)|A=1}))$$

Fair learning problem:

$$\min_{\theta \in \Theta} \mathbb{E}[L(h_\theta(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h_\theta(X)|A=0}, \mathbb{P}_{h_\theta(X)|A=1}))$$

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i),\ i \in \mathbb{N},$ i.i.d. samples

Fair learning problem:

$$\min_{\theta \in \Theta} \mathbb{E}[L(h_\theta(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h_\theta(X)|A=0}, \mathbb{P}_{h_\theta(X)|A=1}))$$

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i),\ i \in \mathbb{N},$ i.i.d. samples

SGD: $\theta_{k+1} = \theta_k - \gamma \cdot g_k(\theta_k)$

Fair Statistical Learning

Fair learning problem:

$$\min_{\theta \in \Theta} \mathbb{E}[L(h_\theta(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h_\theta(X)|A=0}, \mathbb{P}_{h_\theta(X)|A=1}))$$

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i),\ i \in \mathbb{N},$ i.i.d. samples

SGD: $\theta_{k+1} = \theta_k - \gamma \cdot g_k(\theta_k)$

unbiased stochastic gradient
constructed from batch of N samples

Empirical Risk Minimization

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, i.i.d. samples

✓ $\frac{1}{N}\sum_{i=1}^{N} L(h_\theta(\hat{X}_i), \hat{Y}_i)$ unbiased estimator for $\mathbb{E}[L(h(X), Y)]$

Empirical Risk Minimization

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i),\ i \in \mathbb{N},$ i.i.d. samples

✓ $\frac{1}{N}\sum_{i=1}^{N} L(h_\theta(\hat{X}_i), \hat{Y}_i)$ unbiased estimator for $\mathbb{E}[L(h(X), Y)]$

✗ difficult to find unbiased estimator for unfairness penalty

Towards an Unbiased Estimator

Fair learning problem:

$$\min_{\theta \in \Theta} \mathbb{E}[L(h_\theta(X), Y)] + \rho(\mathcal{D}_\Psi(\underbrace{\mathbb{P}_{h_\theta(X)|A=0}}_{\mathbb{Q}_0}, \underbrace{\mathbb{P}_{h_\theta(X)|A=1}}_{\mathbb{Q}_1}))$$

Unfairness Penalty: Squared Kernel Distance

$\rho(\mathcal{D}_{\Psi}(\mathbb{Q}_0, \mathbb{Q}_1))$

Unfairness Penalty: Squared Kernel Distance

$\mathcal{D}_{\Psi}(\mathbb{Q}_0, \mathbb{Q}_1)^2$

Unfairness Penalty: Squared Kernel Distance

$$\mathcal{D}_\Psi(\mathbb{Q}_0, \mathbb{Q}_1)^2 = \left(\sup_{\|\psi\|_{\mathbb{H}} \leq 1} \int_{\mathbb{R}} \psi(y)\, \mathbb{Q}_0(\mathrm{d}y) - \int_{\mathbb{R}} \psi(y)\, \mathbb{Q}_1(\mathrm{d}y) \right)^2$$

Unfairness Penalty: Squared Kernel Distance

Theorem:[1] $\mathcal{D}_\Psi(\mathbb{Q}_0, \mathbb{Q}_1)^2 = \int_{\mathbb{R}\times\mathbb{R}} K(y, y')\, \mathbb{Q}_0(\mathrm{d}y)\, \mathbb{Q}_0(\mathrm{d}y') + \int_{\mathbb{R}\times\mathbb{R}} K(y, y')\, \mathbb{Q}_1(\mathrm{d}y)\, \mathbb{Q}_1(\mathrm{d}y') - 2\int_{\mathbb{R}\times\mathbb{R}} K(y, y')\, \mathbb{Q}_0(\mathrm{d}y)\, \mathbb{Q}_1(\mathrm{d}y')$

admits unbiased estimator!

[1] Sriperumbudur et al., *Electron. J. Stat.*, 2012.

Unfairness Penalty: Squared Kernel Distance

Theorem:[1)]
$$\mathcal{D}_\Psi(\mathbb{Q}_0, \mathbb{Q}_1)^2 = \int_{\mathbb{R}\times\mathbb{R}} K(y, y')\, \mathbb{Q}_0(\mathrm{d}y)\, \mathbb{Q}_0(\mathrm{d}y') + \int_{\mathbb{R}\times\mathbb{R}} K(y, y')\, \mathbb{Q}_1(\mathrm{d}y)\, \mathbb{Q}_1(\mathrm{d}y') - 2 \int_{\mathbb{R}\times\mathbb{R}} K(y, y')\, \mathbb{Q}_0(\mathrm{d}y)\, \mathbb{Q}_1(\mathrm{d}y')$$

Data: $\hat{Y}_i^0, \ldots, \hat{Y}_{N^0}^0 \sim \mathbb{Q}_0$ i.i.d., $\hat{Y}_i^1, \ldots, \hat{Y}_{N^1}^1 \sim \mathbb{Q}_1$ i.i.d.

Unfairness Penalty: Squared Kernel Distance

Theorem:[1)] $$\mathcal{D}_\Psi(\mathbb{Q}_0,\mathbb{Q}_1)^2 = \int_{\mathbb{R}\times\mathbb{R}} K(y,y')\,\mathbb{Q}_0(dy)\,\mathbb{Q}_0(dy') + \int_{\mathbb{R}\times\mathbb{R}} K(y,y')\,\mathbb{Q}_1(dy)\,\mathbb{Q}_1(dy') - 2\int_{\mathbb{R}\times\mathbb{R}} K(y,y')\,\mathbb{Q}_0(dy)\,\mathbb{Q}_1(dy')$$

Data: $\hat{Y}^0_i, \ldots, \hat{Y}^0_{N^0} \sim \mathbb{Q}_0$ i.i.d., $\hat{Y}^1_i, \ldots, \hat{Y}^1_{N^1} \sim \mathbb{Q}_1$ i.i.d.

Unbiased estimator for $\mathcal{D}_\Psi(\mathbb{Q}_0,\mathbb{Q}_1)^2$**:**

$$\sum_{a\in\{0,1\}} \frac{1}{N^a(N^a-1)} \sum_{\substack{i,j=1\\ i\neq j}}^{N^a} K(\hat{Y}^a_i, \hat{Y}^a_j) - 2\frac{1}{N^0 N^1}\sum_{i=1}^{N^0}\sum_{j=1}^{N^1} K(\hat{Y}^0_i, \hat{Y}^1_j)$$

Random Batches

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i),\ i \in \mathbb{N},$ i.i.d. samples (● : $A_i = 0$, ● : $A_i = 1$)

Random Batches

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i),\ i \in \mathbb{N},$ i.i.d. samples (●: $A_i = 0$, ●: $A_i = 1$)

Batch: At least 2 samples from each class and $\overline{N}$ samples in total

Random Batches

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i),\ i \in \mathbb{N},$ i.i.d. samples (blue dot: $A_i = 0$, yellow dot: $A_i = 1$)

Batch: At least 2 samples from each class and $\bar{N}$ samples in total

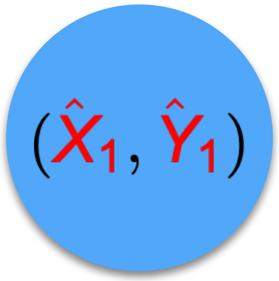
Example: $\bar{N} = 5$

Random Batches

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, i.i.d. samples (blue: $A_i = 0$, yellow: $A_i = 1$)

Batch: At least 2 samples from each class and $\bar{N}$ samples in total

Example: $\bar{N} = 5$

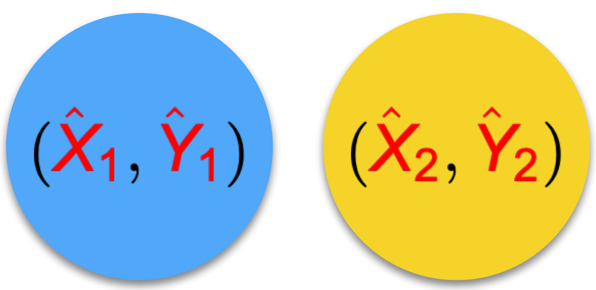


Random Batches

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, i.i.d. samples (● : $A_i = 0$, ● : $A_i = 1$)

Batch: At least 2 samples from each class and $\bar{N}$ samples in total

Example: $\bar{N} = 5$

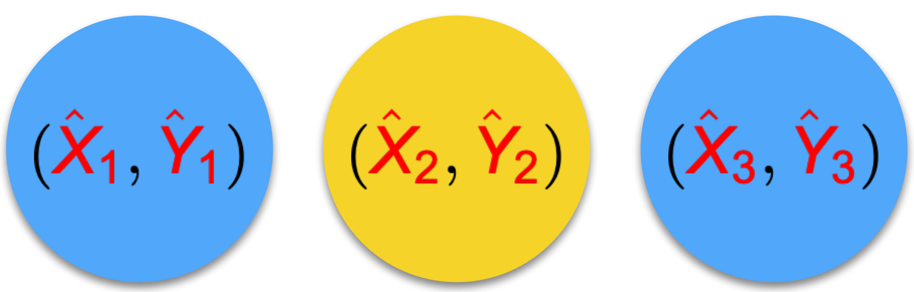


Random Batches

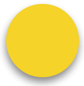
Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, i.i.d. samples (● : $A_i = 0$, ● : $A_i = 1$)

Batch: At least 2 samples from each class and $\bar{N}$ samples in total

Example: $\bar{N} = 5$

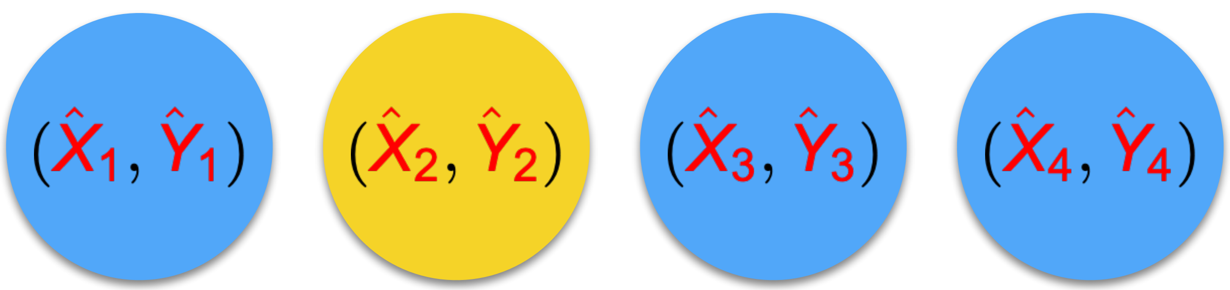


Random Batches

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, i.i.d. samples (: $A_i = 0$, : $A_i = 1$)

Batch: At least 2 samples from each class and $\bar{N}$ samples in total

Example: $\bar{N} = 5$

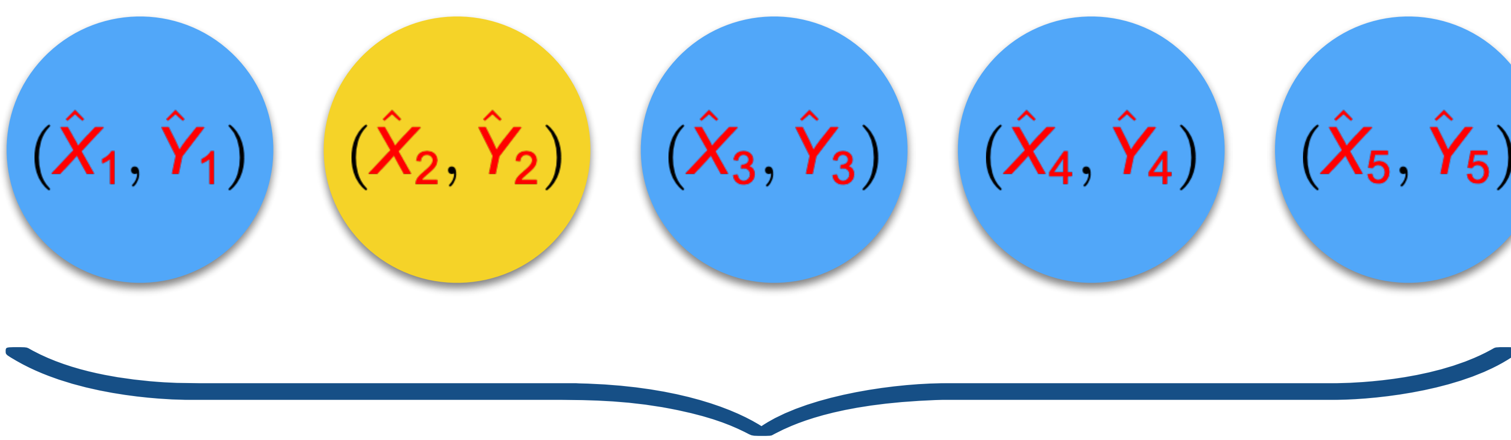


Random Batches

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, i.i.d. samples (: $A_i = 0$, : $A_i = 1$)

Batch: At least 2 samples from each class and $\overline{N}$ samples in total

Example: $\bar{N} = 5$



first 5 samples contain only
one sample from class 1

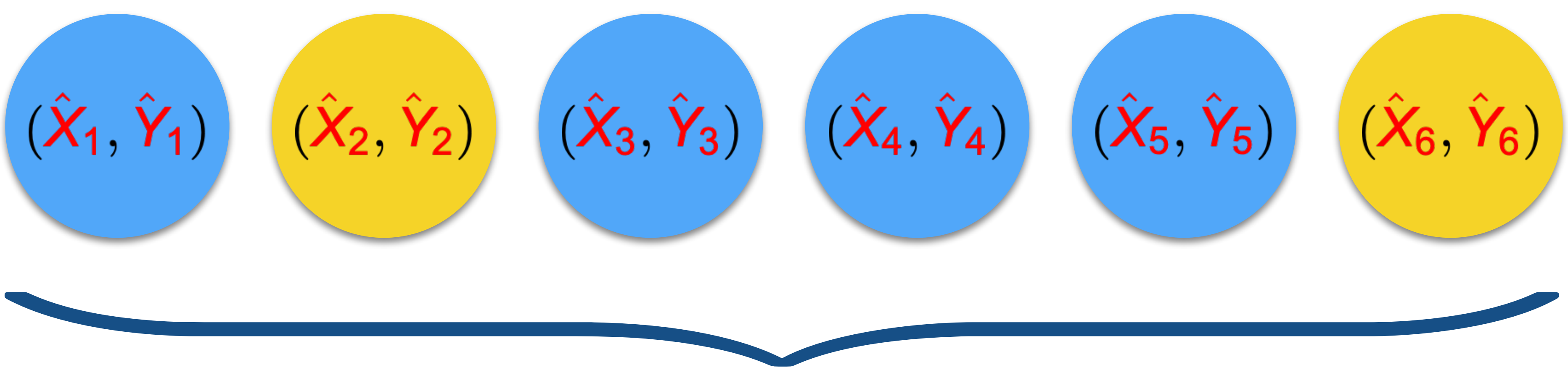
$\implies$ estimator for unfairness penalty undefined!

Random Batches

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, i.i.d. samples (: $A_i = 0$, : $A_i = 1$)

Batch: At least 2 samples from each class and $\bar{N}$ samples in total

Example: $\bar{N} = 5$



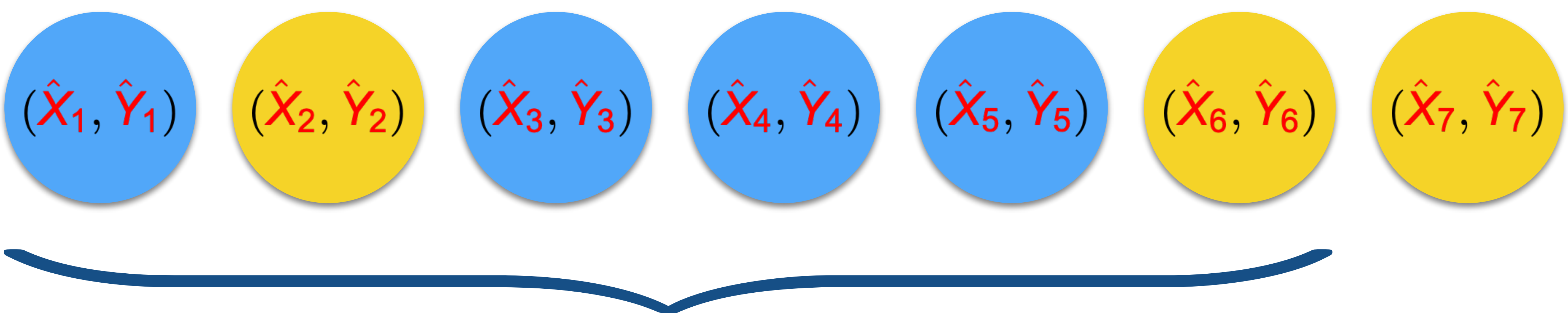
complete batch with 6 samples
(4 from class 0, 2 from class 1)

Random Batches

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, i.i.d. samples (: $A_i = 0$, : $A_i = 1$)

Batch: At least 2 samples from each class and $\bar{N}$ samples in total

Example: $\bar{N} = 5$



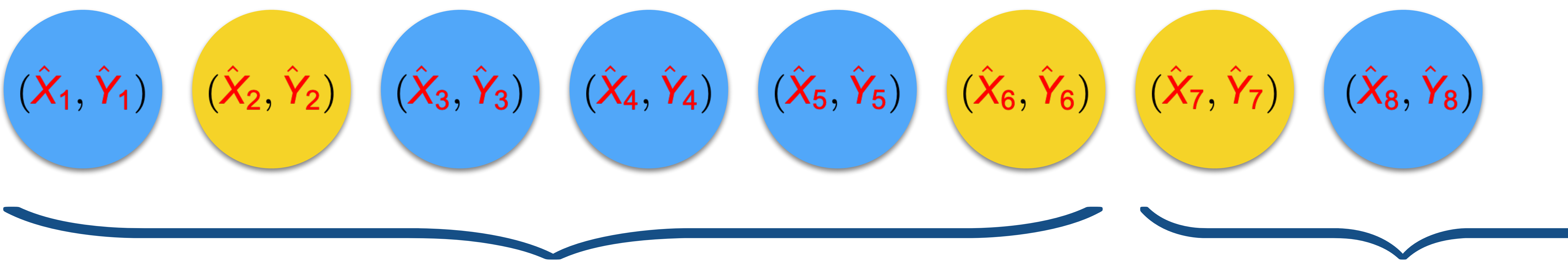
complete batch with 6 samples
(4 from class 0, 2 from class 1)

Random Batches

Data: $(\hat{X}_i, \hat{Y}_i, \hat{A}_i)$, $i \in \mathbb{N}$, i.i.d. samples (: $A_i = 0$, : $A_i = 1$)

Batch: At least 2 samples from each class and $\overline{N}$ samples in total

Example: $\bar{N} = 5$



complete batch with 6 samples
(4 from class 0, 2 from class 1)

next batch

Unbiased Estimators: Unfairness Penalty

Notation:

- $\mathcal{I}_b \subseteq \mathbb{N}$ *b*-th batch
- $\mathcal{I}_b^a =$ class *a* samples in $\mathcal{I}_b, \; a \in \{0, 1\}$

Unbiased Estimators: Unfairness Penalty

Notation:

- $\mathcal{I}_b \subseteq \mathbb{N}$ b-th batch
- $\mathcal{I}_b^a$ = class a samples in $\mathcal{I}_b$, $a \in \{0,1\}$

Lemma: The following estimator of the unfairness penalty is unbiased for every batch b.

$$\hat{U}_b(\theta) = \begin{cases} \displaystyle\sum_{a\in\{0,1\}} \frac{1}{|\mathcal{I}_b^a|(|\mathcal{I}_b^a|-1)} \sum_{i\neq j\in\mathcal{I}_b^a} K(h_\theta(\hat{X}_i), h_\theta(\hat{X}_j)) \\ \displaystyle -2\frac{1}{|\mathcal{I}_b^0|\cdot|\mathcal{I}_b^1|} \sum_{i\in\mathcal{I}_b^0}\sum_{j\in\mathcal{I}_b^1} K(h_\theta(\hat{X}_i), h_\theta(\hat{X}_j)) \end{cases}$$

Unbiased Estimators: Unfairness Penalty

Notation:
- $\mathcal{I}_b \subseteq \mathbb{N}$ b-th batch
- $\mathcal{I}_b^a =$ class a samples in $\mathcal{I}_b$, $a \in \{0, 1\}$

Lemma: The following estimator of the unfairness penalty is unbiased for every batch b.

$$\hat{U}_b(\theta) = \begin{cases} \displaystyle\sum_{a\in\{0,1\}} \frac{1}{|\mathcal{I}_b^a|(|\mathcal{I}_b^a|-1)} \sum_{i\neq j\in\mathcal{I}_b^a} K(h_\theta(\hat{X}_i), h_\theta(\hat{X}_j)) \\ \displaystyle -2\frac{1}{|\mathcal{I}_b^0|\cdot|\mathcal{I}_b^1|} \sum_{i\in\mathcal{I}_b^0}\sum_{j\in\mathcal{I}_b^1} K(h_\theta(\hat{X}_i), h_\theta(\hat{X}_j)) \end{cases}$$

Note: All index sets are random!

Unbiased Estimators: Unfairness Penalty

Notation:
- $\mathcal{I}_b \subseteq \mathbb{N}$ b-th batch
- $\mathcal{I}_b^a$ = class a samples in $\mathcal{I}_b$, $a \in \{0, 1\}$

Lemma: The following estimator of the unfairness penalty is unbiased for every batch b.

$$\hat{U}_b(\theta) = \begin{cases} \displaystyle\sum_{a\in\{0,1\}} \frac{1}{|\mathcal{I}_b^a|(|\mathcal{I}_b^a| - 1)} \sum_{i\neq j\in\mathcal{I}_b^a} K(h_\theta(\hat{X}_i), h_\theta(\hat{X}_j)) \\ \displaystyle -2\frac{1}{|\mathcal{I}_b^0| \cdot |\mathcal{I}_b^1|} \sum_{i\in\mathcal{I}_b^0} \sum_{j\in\mathcal{I}_b^1} K(h_\theta(\hat{X}_i), h_\theta(\hat{X}_j)) \end{cases}$$

$\Longrightarrow$ $\nabla_\theta \hat{U}_b(\theta)$ is an unbiased stochastic gradient

Unbiased Estimators: Prediction Loss

Empirical prediction loss:

$$\frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} L(h_{\theta}(\hat{X}_i), \hat{Y}_i)$$

Unbiased Estimators: Prediction Loss

Empirical prediction loss:

$$\frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} L(h_\theta(\hat{X}_i), \hat{Y}_i)$$

biased because $\mathcal{I}_b$ is random!

Unbiased Estimators: Prediction Loss

Empirical prediction loss:

$$\frac{1}{|\mathcal{I}_b|} \sum_{a \in \{0,1\}} \sum_{i \in \mathcal{I}_b^a} L(h_\theta(\hat{X}_i), \hat{Y}_i)$$

Unbiased Estimators: Prediction Loss

Empirical prediction loss:

$$\frac{1}{|\mathcal{I}_b|} \sum_{a \in \{0,1\}} \sum_{i \in \mathcal{I}_b^a} \underbrace{\Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|)} \cdot L(h_\theta(\hat{X}_i), \hat{Y}_i)$$

bias correction term

Definition: For $N \in \{\bar{N}, \bar{N}+1, \ldots\}$ and $n \in \{2, \ldots, N-2\}$, set

$$\Delta(N, n) = 1_{N=\bar{N}} + \frac{N}{2(N-1)} 1_{(N>\bar{N}) \wedge (n=2)} + \frac{N}{N-1} 1_{(N>\bar{N}) \wedge (n=N-2)}$$

Unbiased Estimators: Prediction Loss

Empirical prediction loss:

$$\frac{1}{|\mathcal{I}_b|} \sum_{a \in \{0,1\}} \sum_{i \in \mathcal{I}_b^a} \underbrace{\Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|)} \cdot L(h_\theta(\hat{X}_i), \hat{Y}_i)$$

bias correction term

Definition: For $N \in \{\bar{N}, \bar{N}+1, \ldots\}$ and $n \in \{2, \ldots, N-2\}$, set

$$\Delta(N, n) = 1_{N=\bar{N}} + \frac{N}{2(N-1)} 1_{(N>\bar{N}) \wedge (n=2)} + \frac{N}{N-1} 1_{(N>\bar{N}) \wedge (n=N-2)}$$

Unbiased Estimators: Prediction Loss

Lemma: The following estimator of the prediction loss is unbiased for every batch *b*.

$$\hat{R}_b(\theta) = \frac{1}{|\mathcal{I}_b|} \sum_{a \in \{0,1\}} \sum_{i \in \mathcal{I}_b^a} \Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|) \cdot L(h_\theta(\hat{X}_i), \hat{Y}_i)$$

Unbiased Estimators: Prediction Loss

Lemma: The following estimator of the prediction loss is unbiased for every batch b.

$$\hat{R}_b(\theta) = \frac{1}{|\mathcal{I}_b|} \sum_{a \in \{0,1\}} \sum_{i \in \mathcal{I}_b^a} \Delta(|\mathcal{I}_b|, |\mathcal{I}_b^a|) \cdot L(h_\theta(\hat{X}_i), \hat{Y}_i)$$

$\implies \nabla_\theta \hat{R}_b(\theta)$ is an unbiased stochastic gradient

SGD Convergence

Fair learning problem:

$$\min_{\theta \in \Theta} \mathbb{E}[L(h_\theta(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h_\theta(X)|A=0}, \mathbb{P}_{h_\theta(X)|A=1})) \qquad (*)$$

SGD Convergence

Fair learning problem:

$$\min_{\theta \in \Theta} \mathbb{E}[L(h_\theta(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h_\theta(X)|A=0}, \mathbb{P}_{h_\theta(X)|A=1})) \qquad (*)$$

Theorem: If $\mathcal{D}_\Psi$ is a kernel distance and $\rho(z) = \lambda z^2$ with $\lambda \geq 0$, then $\nabla_\theta \hat{R}_b(\theta) + \lambda \cdot \nabla_\theta \hat{U}_b(\theta)$ is an unbiased gradient estimator for $(*)$.

SGD Convergence

Fair learning problem:

$$\min_{\theta \in \Theta} \mathbb{E}[L(h_\theta(X), Y)] + \rho(\mathcal{D}_\Psi(\mathbb{P}_{h_\theta(X)|A=0}, \mathbb{P}_{h_\theta(X)|A=1})) \qquad (*)$$

Theorem: If $\mathcal{D}_\Psi$ is a kernel distance and $\rho(z) = \lambda z^2$ with $\lambda \geq 0$, then $\nabla_\theta \hat{R}_b(\theta) + \lambda \cdot \nabla_\theta \hat{U}_b(\theta)$ is an unbiased gradient estimator for $(*)$.

$\implies$ SGD converges (in expectation) to a stationary point of $(*)$

Numerical Experiments

Regression

Synthetic data:

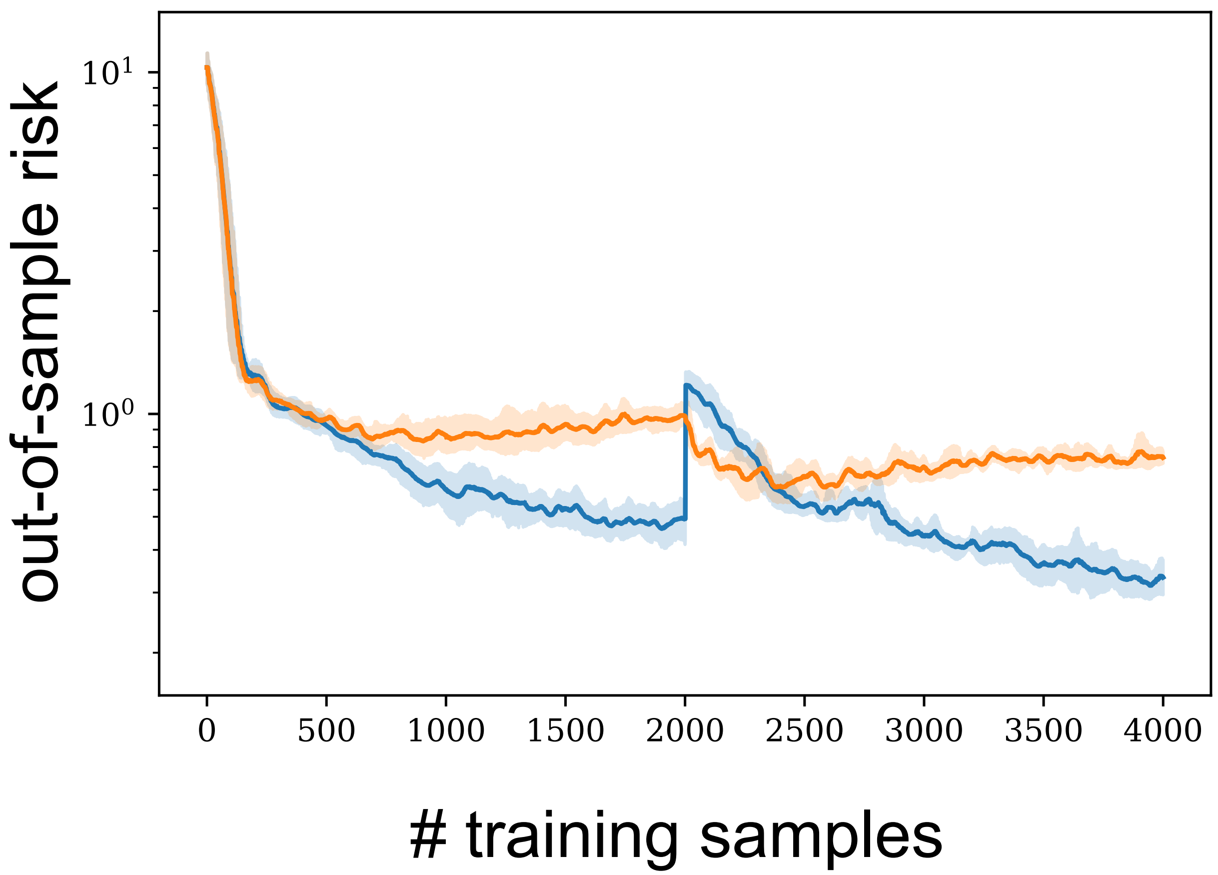
- Input: $X \sim \mathcal{U}([0,1]^9 \times \{0,1\})$
- Sensitive attribute: $A = X_{10}$
- Output: $Y = \max\{s_1^\top X, \ldots, s_5^\top X\}$

Regression model:

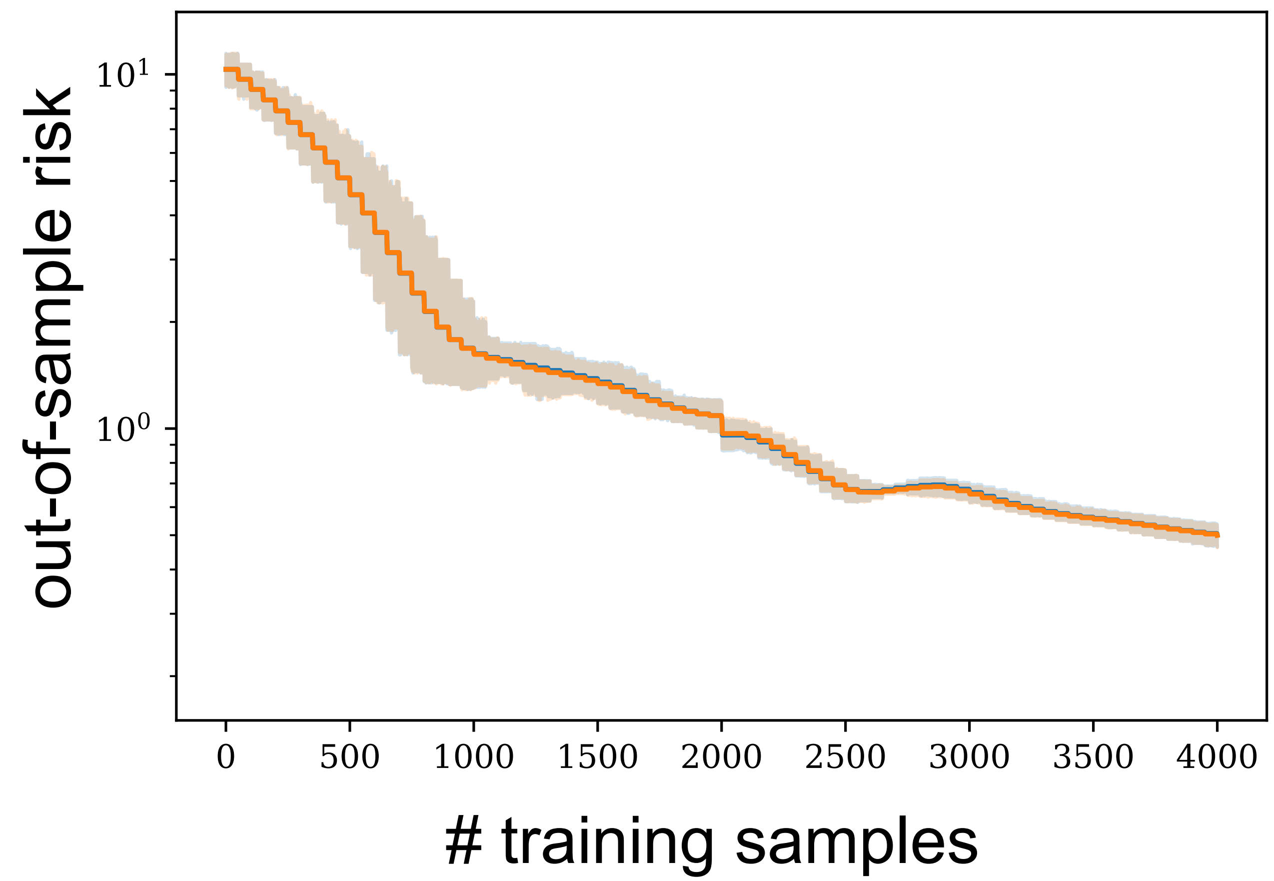
- Square loss: $L(\widehat{y}, y) = (\widehat{y} - y)^2$
- Predictor: 3-layer NN with 20 hidden nodes

Regression

Batch size $\overline{N} = 4$

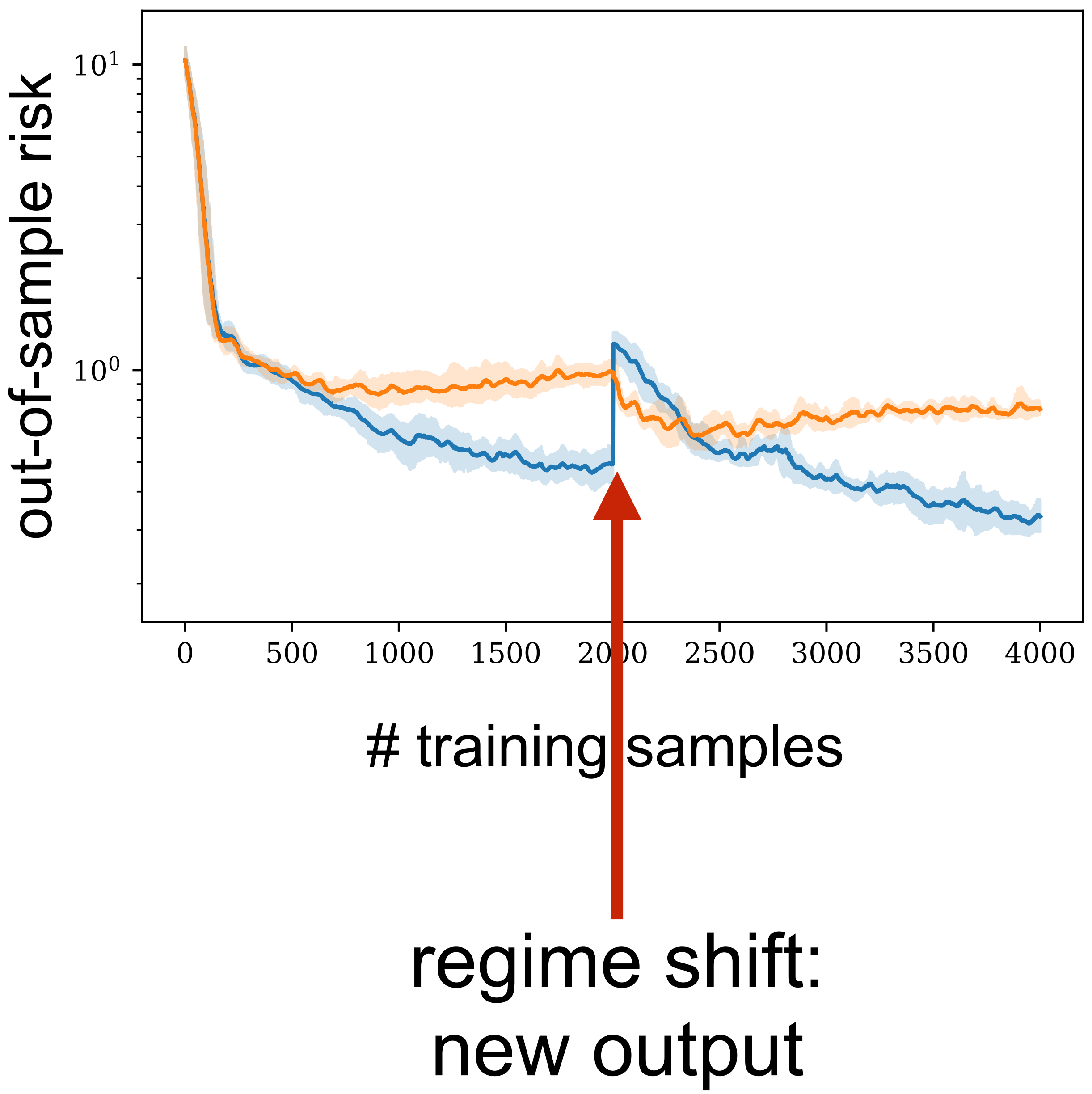


Batch size $\overline{N} = 50$

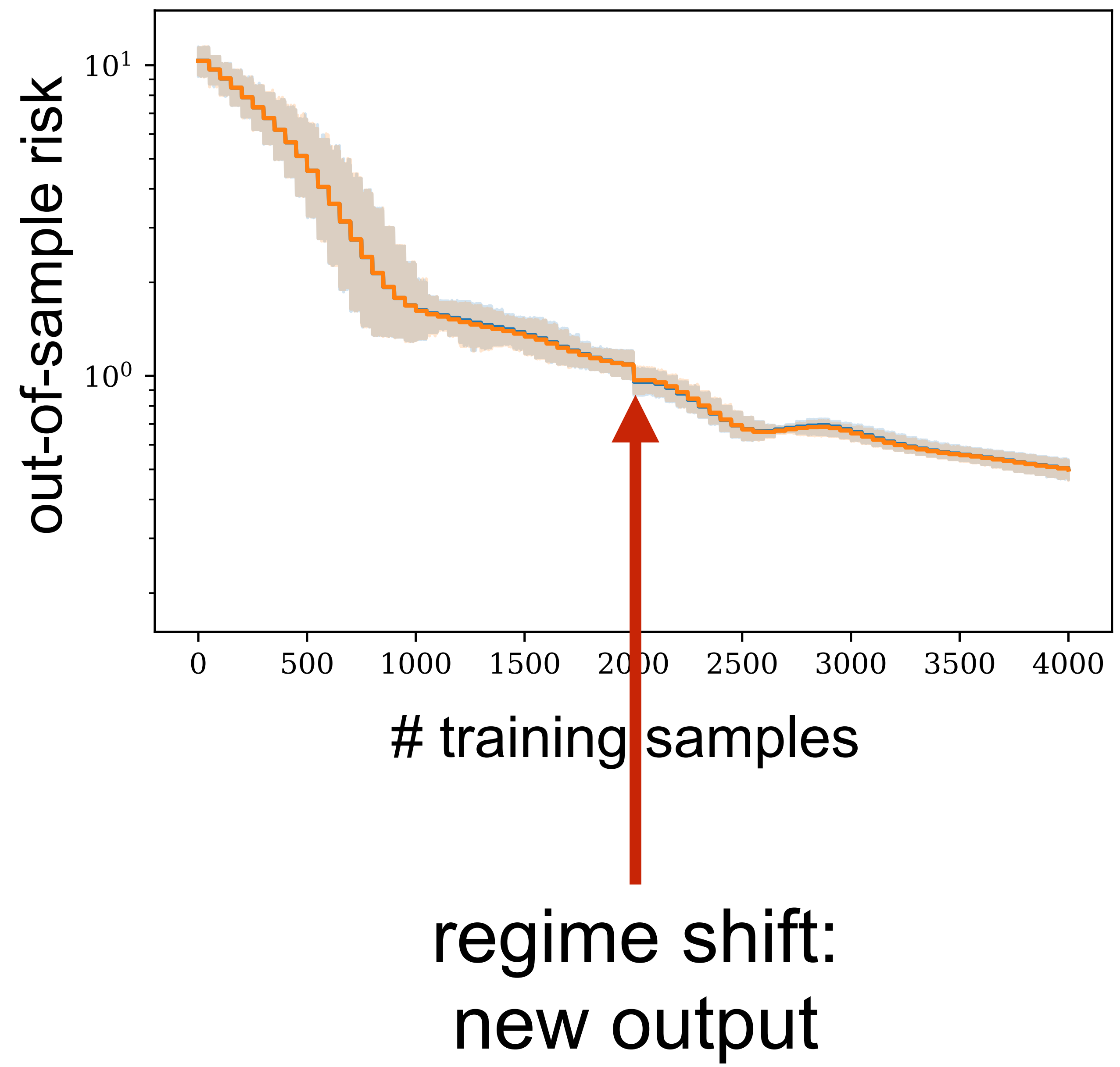


Regression

Batch size $\overline{N} = 4$

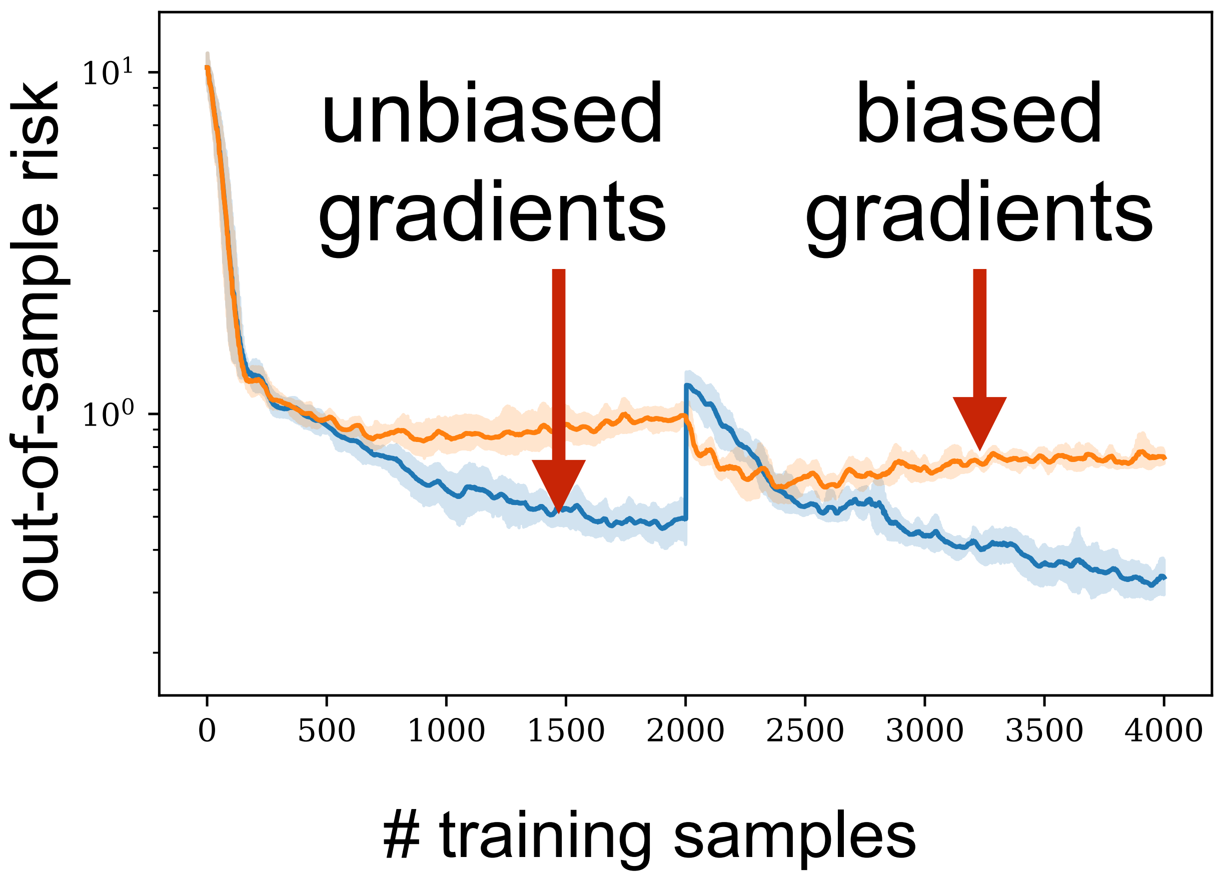


Batch size $\overline{N} = 50$

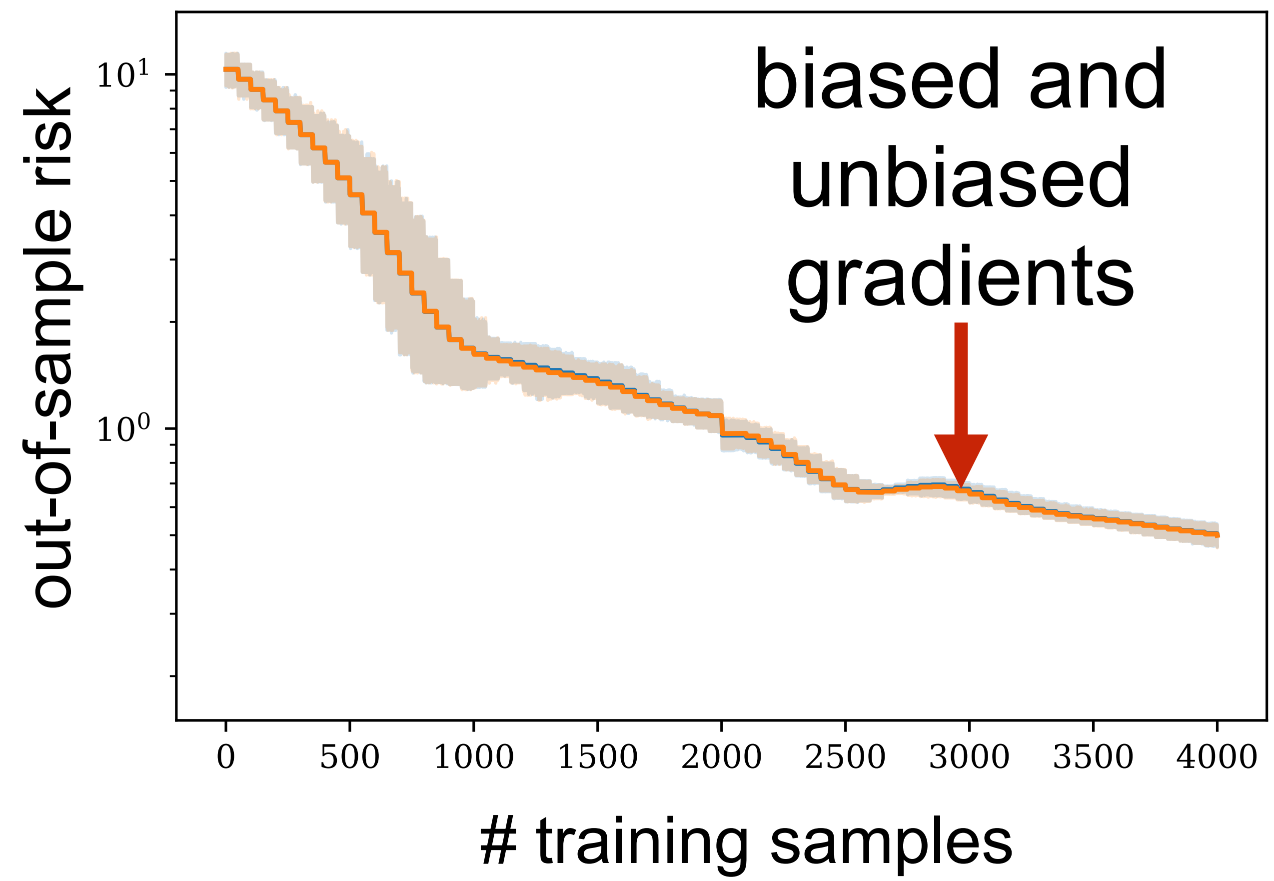


Regression

Batch size $\overline{N} = 4$



Batch size $\overline{N} = 50$



Merits of Unbiased Gradient Estimators

		batch size	
		small	large
gradients	biased		
	unbiased		

Merits of Unbiased Gradient Estimators

		batch size	
		small	large
gradients	biased	fast convergence to bad solution	
	unbiased		

Merits of Unbiased Gradient Estimators

		batch size: small	batch size: large
gradients	biased	fast convergence to bad solution	slow convergence to good solution
	unbiased		

Merits of Unbiased Gradient Estimators

		batch size: small	batch size: large
gradients	biased	fast convergence to bad solution	slow convergence to good solution
gradients	unbiased		slow convergence to good solution

Merits of Unbiased Gradient Estimators

		batch size	
		small	large
gradients	biased	fast convergence to bad solution	slow convergence to good solution
	unbiased	fast convergence to good solution	slow convergence to good solution

Drug dataset:[1]

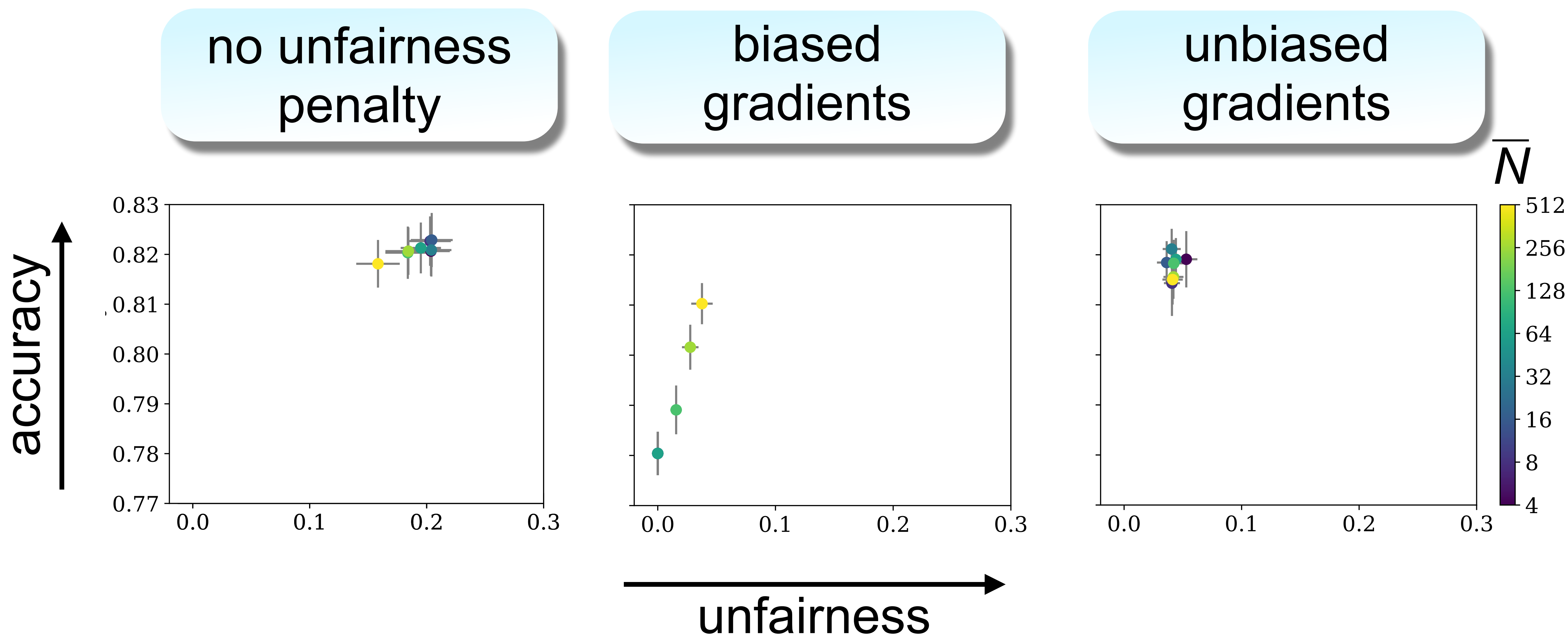
- Input: personality type, level of education, age etc.
- Sensitive attribute: race
- Output: “used” vs. “never used” for heroin

Classification model:

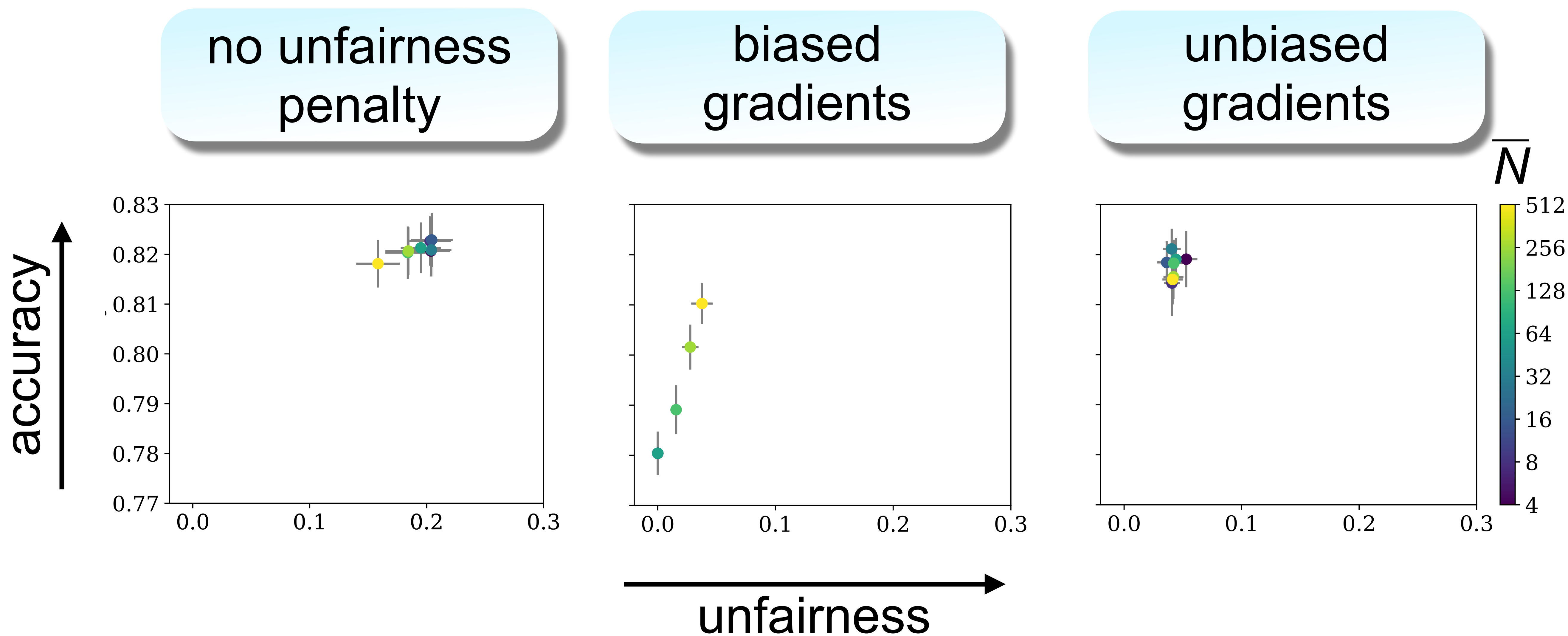
- Cross-entropy loss: $L(\widehat{y}, y) = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$
- Predictor: 3-layer NN with 16 hidden nodes

1) https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29

Classification



Classification



Classical ERM:
- high accuracy
- high unfairness

Fair ERM:
- accuracy sensitive to $\overline{N}$
- low unfairness

Our method:
- high accuracy
- low unfairness

Conclusions

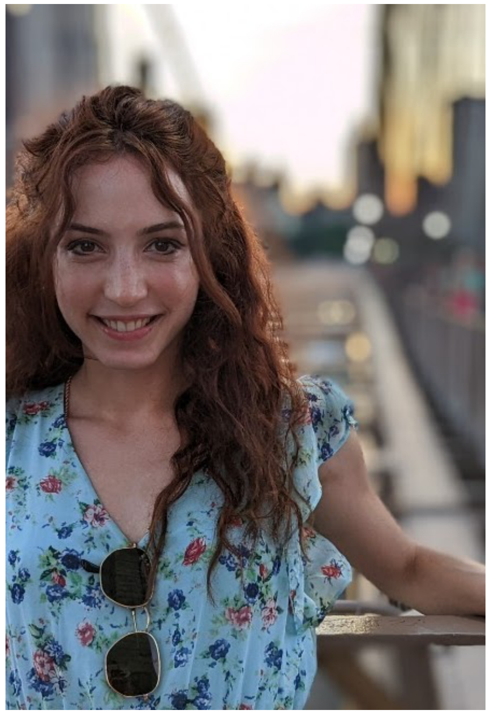
- **Impact of SP constraints**
  - Y has no bias in training $\Longrightarrow$ SP increases test error
  - Y has small bias in training & A is irrelevant for predicting Y $\Longrightarrow$ SP decreases test error
  - Good sensitive attribute: Any feature A with $\mathbb{P}_{Y|X} \perp A$

- **Learning problems with unfairness penalties**
  - Any IPM provides an unfairness measure
  - Empirical estimator of unfairness penalty is biased
  - Moore Aronszajn theorem $\Longrightarrow$ squared kernel distance admits unbiased estimator
  - Fair learning problems can be solved with SGD

This Talk is Based on…

Y. Rychener, B. Taşkesen, D. Kuhn. **Metrizing Fairness**. arXiv. 2024.

Yves Rychener



Bahar Taşkesen

SWISS NATIONAL SCIENCE FOUNDATION

NCCR Automation

EPFL