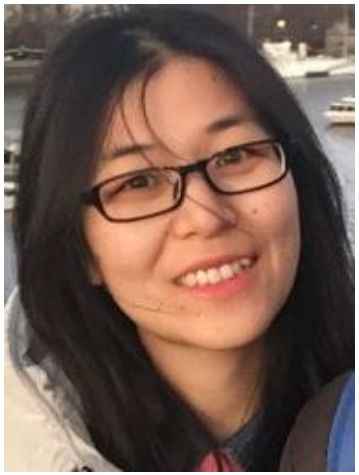
Metric-Fair Active Learning

Jie Shen[1]
Stevens Institute
of Technology

Nan Cui[1]
Stevens Institute
of Technology



Jing Wang[2]
Amazon

Active PAC Learning of homogeneous halfspaces

Active PAC Learning of homogeneous halfspaces

- **Homogeneous halfspaces class:**

$$\mathcal{H} \coloneqq \left\{ x \mapsto \operatorname{sign}(w \cdot x), w \in \mathbb{R}^d, \left|\left|w\right|\right|_2 = 1 \right\}$$

Active PAC Learning of homogeneous halfspaces

- **Homogeneous halfspaces class:**

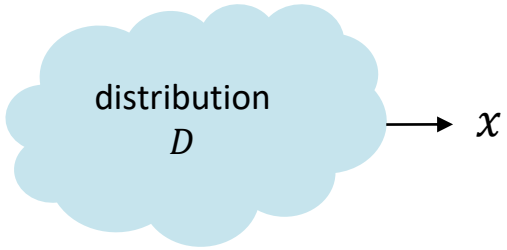
$$\mathcal{H} \coloneqq \left\{x \mapsto \ \mathrm{sign}(w \cdot x), w \in \mathbb{R}^d, \left|\left|w\right|\right|_2 = 1\right\}$$

- **PAC Learnable:**

$$\mathrm{P}_{x \sim D}\left(\mathrm{sign}(\widehat{w} \cdot x) \neq \mathrm{sign}(w^* \cdot x)\right) \leq \epsilon$$

Active PAC Learning of homogeneous halfspaces

- **Homogeneous halfspaces class:**



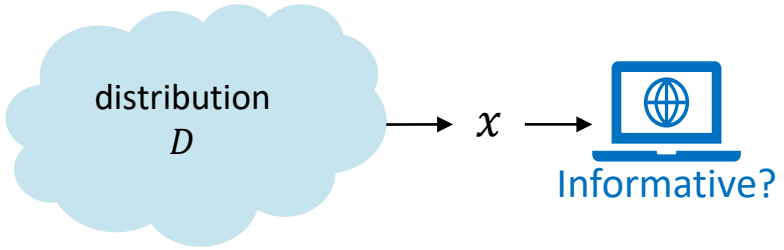
$$\mathcal{H} \coloneqq \left\{x \mapsto \ \mathrm{sign}(w \cdot x), w \in \mathbb{R}^d, \left|\left|w\right|\right|_2 = 1\right\}$$

- **PAC Learnable:**

$$\mathrm{P}_{x\sim D}\left(\mathrm{sign}(\widehat{w} \cdot x) \neq \mathrm{sign}(w^* \cdot x)\right) \leq \epsilon$$

Active PAC Learning of homogeneous halfspaces

- **Homogeneous halfspaces class:**



$$\mathcal{H} := \left\{ x \mapsto \operatorname{sign}(w \cdot x), w \in \mathbb{R}^d, \left|\left|w\right|\right|_2 = 1 \right\}$$

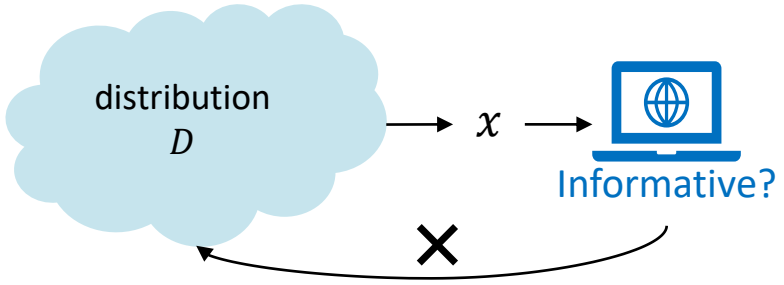
- **PAC Learnable:**

$$\mathrm{P}_{x \sim D}\left(\operatorname{sign}(\widehat{w} \cdot x) \neq \operatorname{sign}(w^* \cdot x)\right) \leq \epsilon$$

Active PAC Learning of homogeneous halfspaces

- **Homogeneous halfspaces class:**

$$\mathcal{H} \coloneqq \left\{x \mapsto \ \mathrm{sign}(w \cdot x), w \in \mathbb{R}^d, \left|\left|w\right|\right|_2 = 1\right\}$$



- **PAC Learnable:**

$$\mathrm{P}_{x\sim D}\left(\mathrm{sign}(\widehat{w} \cdot x) \neq \mathrm{sign}(w^* \cdot x)\right) \leq \epsilon$$

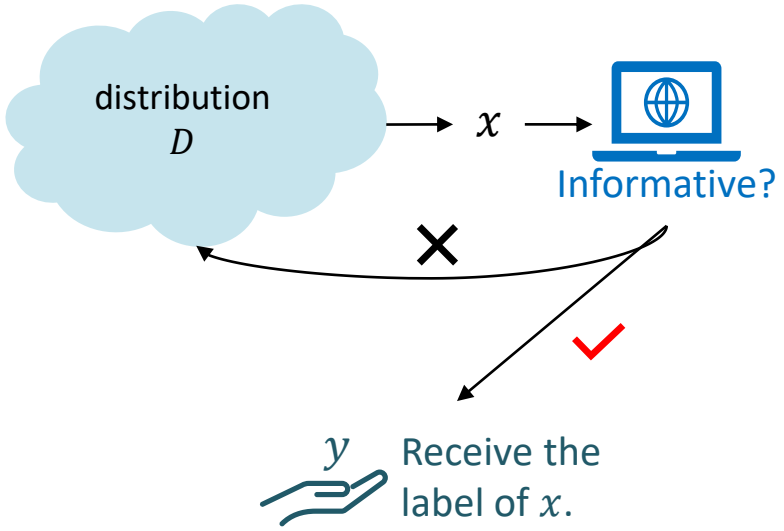
Active PAC Learning of homogeneous halfspaces

- **Homogeneous halfspaces class:**

$$\mathcal{H} := \left\{ x \mapsto \text{sign}(w \cdot x), w \in \mathbb{R}^d, \left|\left|w\right|\right|_2 = 1 \right\}$$

- **PAC Learnable:**

$$\mathrm{P}_{x \sim D}\left(\text{sign}(\widehat{w} \cdot x) \neq \text{sign}(w^* \cdot x)\right) \leq \epsilon$$



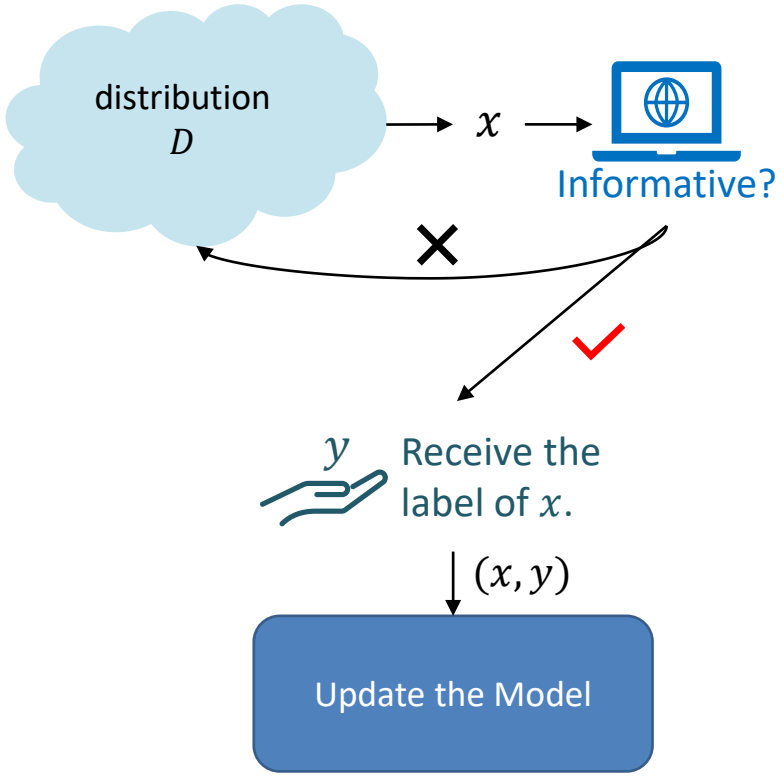
Active PAC Learning of homogeneous halfspaces

- **Homogeneous halfspaces class:**

$$\mathcal{H} := \left\{ x \mapsto \text{sign}(w \cdot x), w \in \mathbb{R}^d, ||w||_2 = 1 \right\}$$

- **PAC Learnable:**

$$\mathrm{P}_{x \sim D}\left(\text{sign}(\hat{w} \cdot x) \neq \text{sign}(w^* \cdot x)\right) \leq \epsilon$$



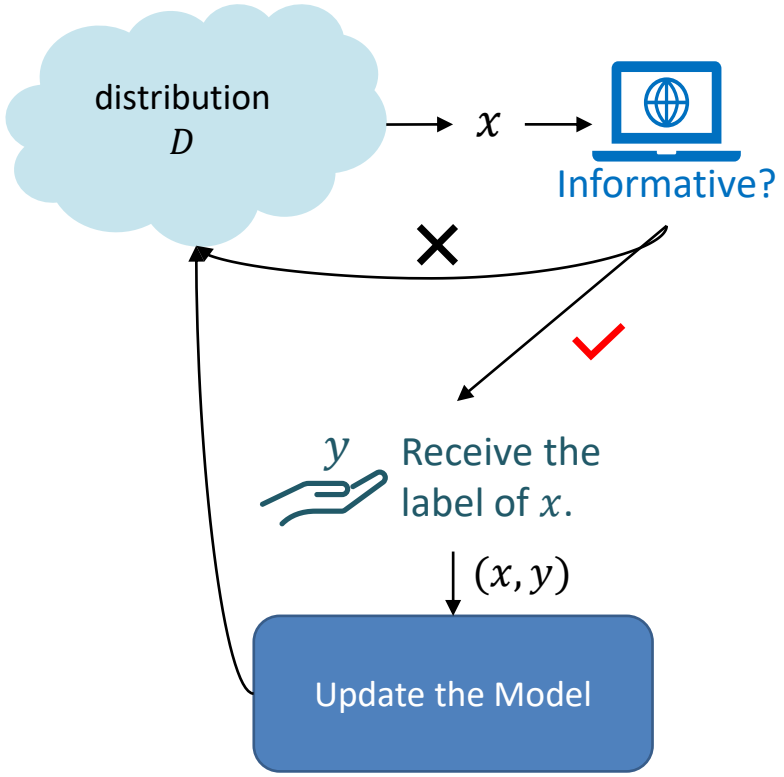
Active PAC Learning of homogeneous halfspaces

- **Homogeneous halfspaces class:**

$$\mathcal{H} := \left\{ x \mapsto \text{sign}(w \cdot x), w \in \mathbb{R}^d, ||w||_2 = 1 \right\}$$

- **PAC Learnable:**

$$\mathrm{P}_{x \sim D}\left(\text{sign}(\hat{w} \cdot x) \neq \text{sign}(w^* \cdot x)\right) \leq \epsilon$$



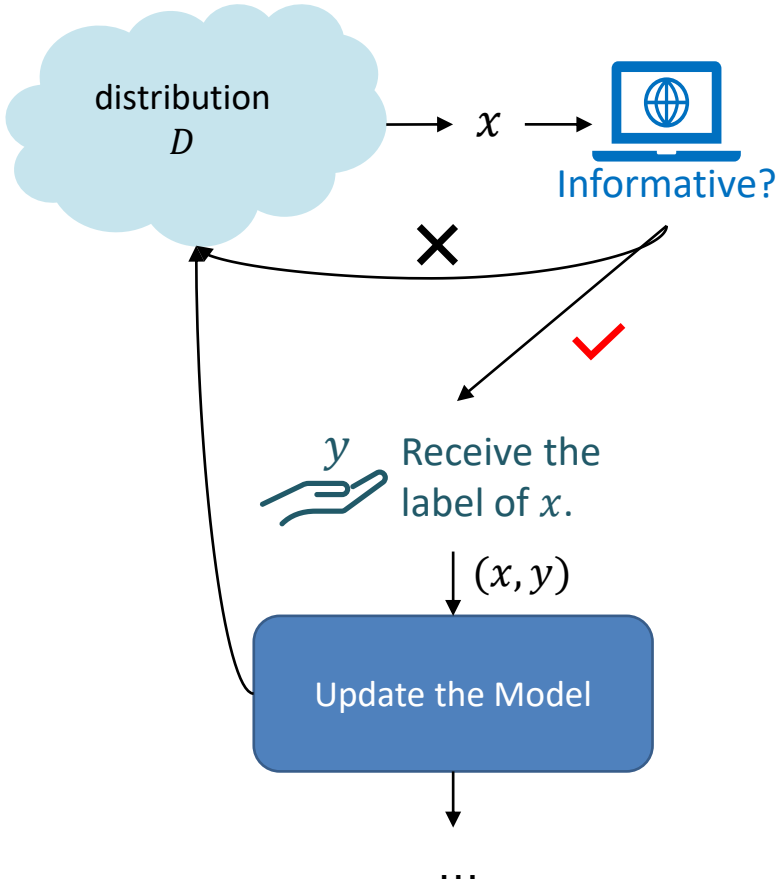
Active PAC Learning of homogeneous halfspaces

- **Homogeneous halfspaces class:**

$$\mathcal{H} := \left\{x \mapsto \text{sign}(w \cdot x), w \in \mathbb{R}^d, \left||w|\right|_2 = 1\right\}$$

- **PAC Learnable:**

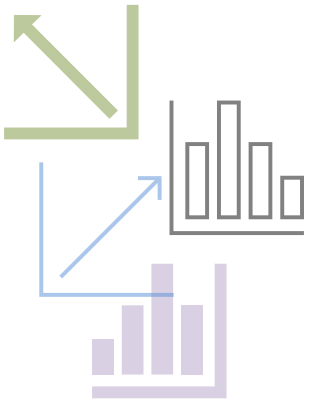
$$\mathrm{P}_{x\sim D}\left(\text{sign}(\hat{w} \cdot x) \neq \text{sign}(w^* \cdot x)\right) \leq \epsilon$$



Main Goal

We study the two properties that seemingly are odd with each other in machine learning:

Prediction Accuracy



Try your best to minimize the model's loss.

To avoid treating vulnerable populations unfairly.

Fairness

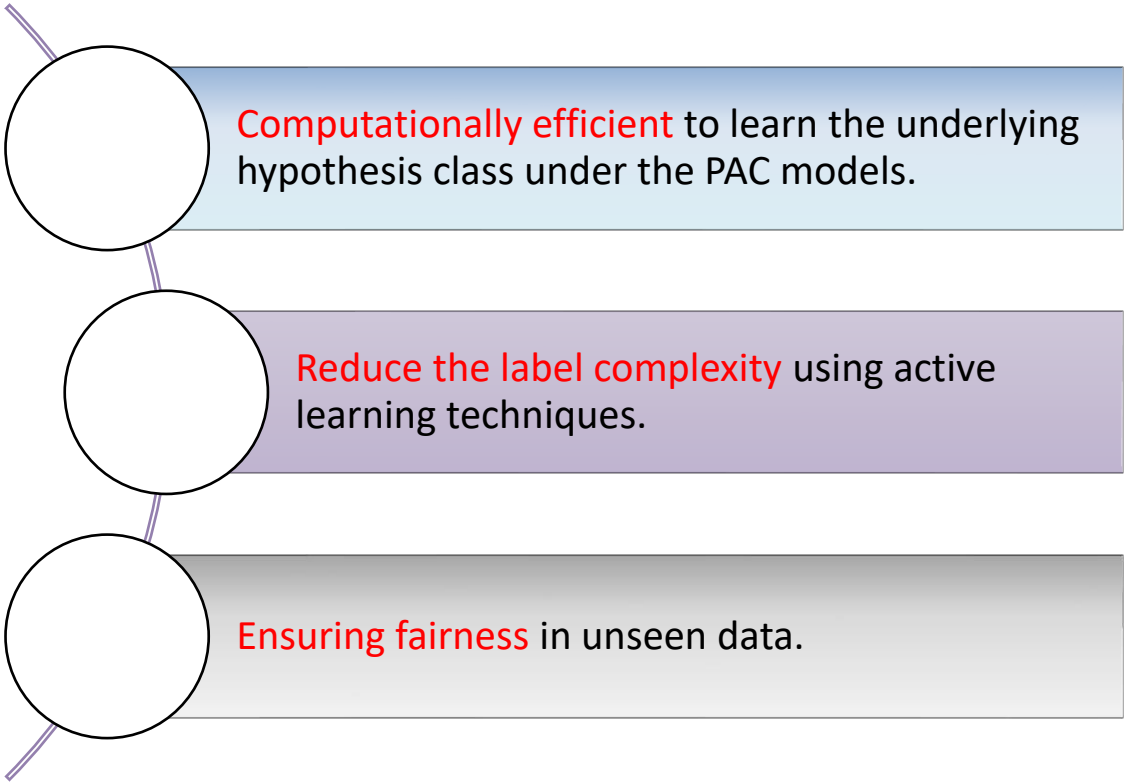
Only ensuring fairness is simple but considering both perspectives is difficult.

Main Goal

Our Main Goals:

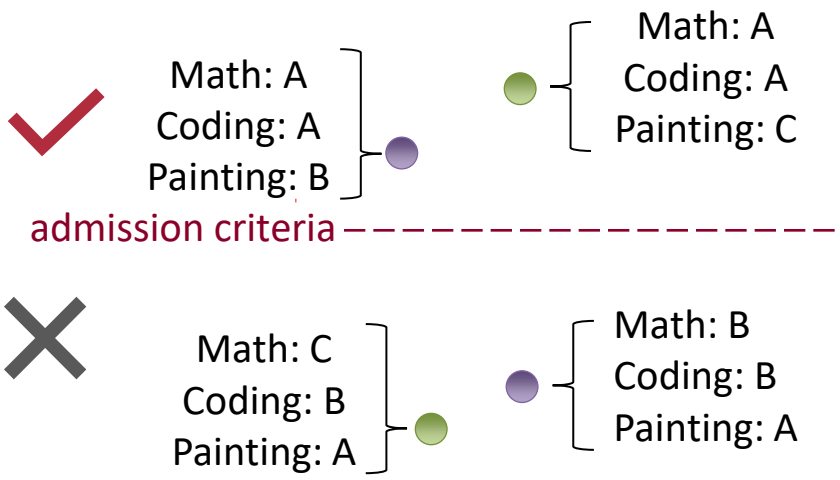
Main Goal

Our Main Goals:



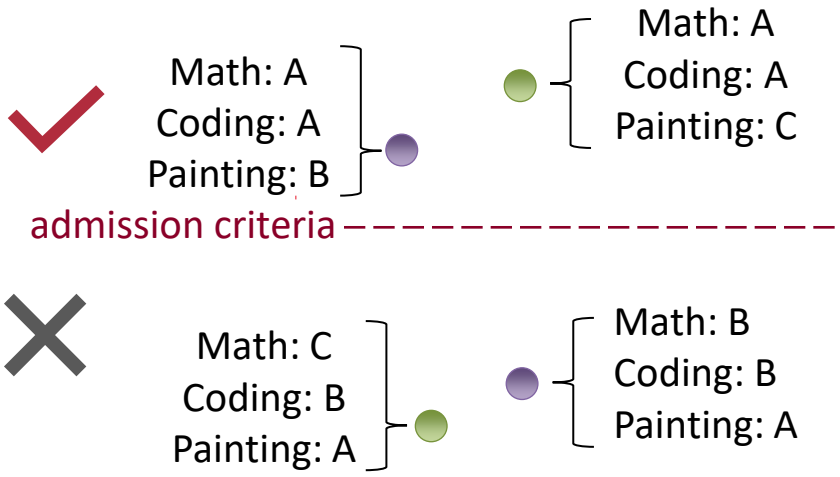
Fairness

- Individual Fairness [Dwork et al. 2011]:
  “similar people should be treated similarly”.



Fairness

- Individual Fairness [Dwork et al. 2011]: “similar people should be treated similarly”.



- **α-Approximate metric-fairness** [Yona & Rothblum 2018]:

  Given a metric $\zeta: \mathcal{X} \times \mathcal{X} \to [0,1]$ and an $\alpha \in (0,1)$, let the fairness error be

$$\Pr_{D_X \times D_X} \left(|w \cdot x - w' \cdot x| > \zeta(x, x') \right) \leq \alpha.$$

Techniques

Metric-fair Learning via Convex Fairness Loss

Techniques

Metric-fair Learning via Convex Fairness Loss

- We have a new Rademacher analysis for both PAC and fairness.

Techniques

Metric-fair Learning via Convex Fairness Loss

- We have a new Rademacher analysis for both PAC and fairness.
- Indicator Function Metric-Fair Loss:

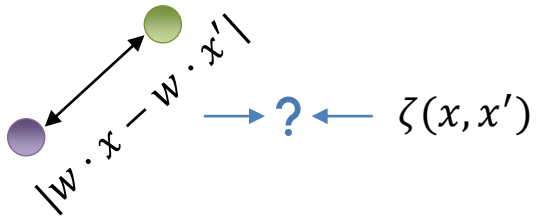
$$f_\zeta\big(w;(x,x')\big) = \begin{cases} 1 & \text{if } |w\cdot x - w\cdot x'| > \zeta(x,x'), \\ 0 & \text{otherwise.} \end{cases}$$

Techniques

Metric-fair Learning via Convex Fairness Loss

- We have a new Rademacher analysis for both PAC and fairness.
- Indicator Function Metric-Fair Loss:

$$f_\zeta\big(w;(x,x')\big) = \begin{cases} 1 & \text{if } |w\cdot x - w\cdot x'| > \zeta(x,x'), \\ 0 & \text{otherwise.} \end{cases}$$



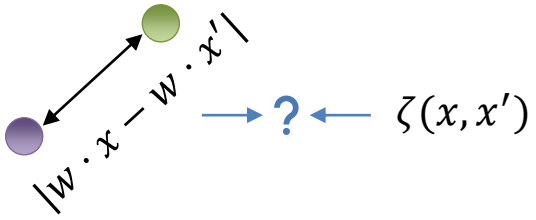
Techniques

Metric-fair Learning via Convex Fairness Loss

- We have a new Rademacher analysis for both PAC and fairness.
- Indicator Function Metric-Fair Loss:

$$f_\zeta\big(w;(x,x')\big) = \begin{cases} 1 & \text{if } |w\cdot x - w\cdot x'| > \zeta(x,x'), \\ 0 & \text{otherwise.} \end{cases}$$

Non-convex and non-Lipschitz!



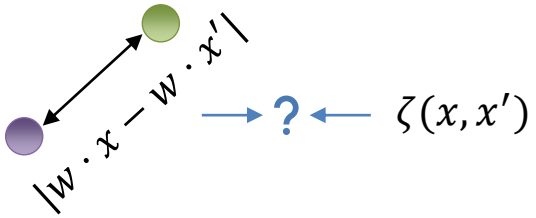
Techniques

Metric-fair Learning via Convex Fairness Loss

- We have a new Rademacher analysis for both PAC and fairness.
- Indicator Function Metric-Fair Loss:

$$f_\zeta(w;(x,x')) = \begin{cases} 1 & \text{if } |w\cdot x - w\cdot x'| > \zeta(x,x'), \\ 0 & \text{otherwise.} \end{cases}$$

Non-convex and non-Lipschitz!



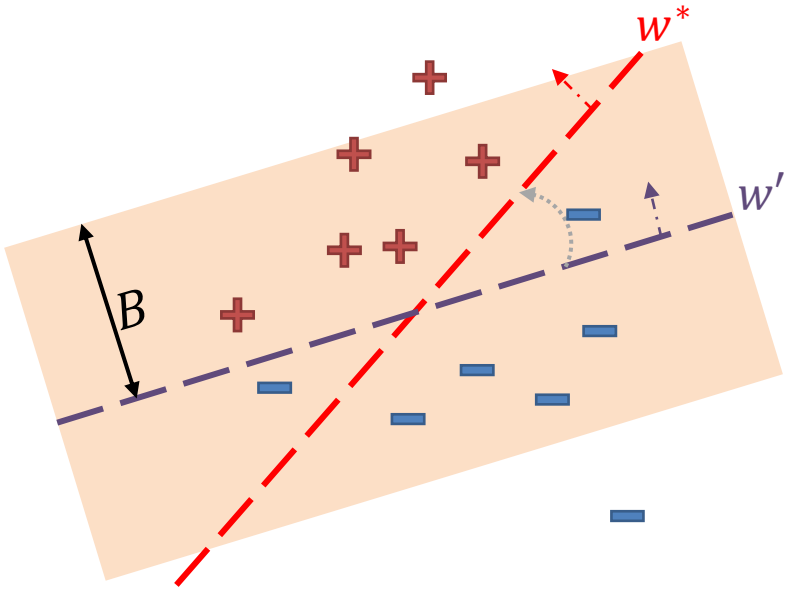
Make it convex and G-Lipschitz!

- Surrogate Metric-Fair Loss:

$$f_\zeta^G(w;(x,x')) \coloneqq \max\{0, G(|w\cdot x - w\cdot x'| - \zeta(x,x')) + 1\}$$

Techniques

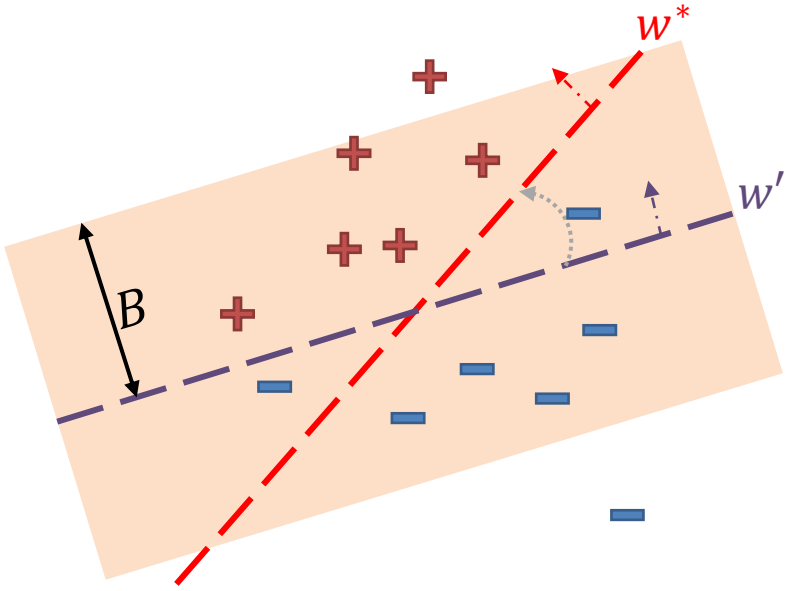
Margin-Based Active Learning [Balcan et al. 2007]



Techniques

Margin-Based Active Learning [Balcan et al. 2007]

- Localized Sampling: Pick the instances only residing in the band B.

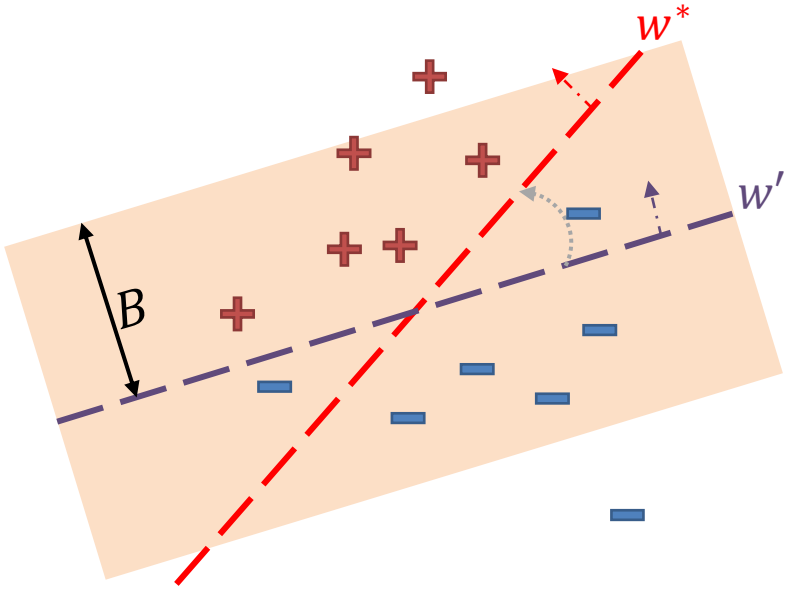


Techniques

Margin-Based Active Learning [Balcan et al. 2007]

- Localized Sampling: Pick the instances only residing in the band B.
- Minimizing a hinge loss with a sample set S under a certain constraint set W:

$$w' \leftarrow \text{argmin}_{w \in W} \sum_{(x,y)\sim S} \max\left\{1 - \frac{y\, w \cdot x}{\tau}\right\}.$$



Main Results

Main Results

- **Main Theorem:** There is a polynomial-time algorithm that PACF learns $\mathcal{H}$ with the label complexity is $\text{polylog}\left(\frac{1}{\epsilon}, \frac{1}{\alpha}\right) \cdot O(d \cdot \text{polylog}(d))$.

Main Results

- **Main Theorem:** There is a polynomial-time algorithm that PACF learns $\mathcal{H}$ with the label complexity is $\text{polylog}\left(\frac{1}{\epsilon},\frac{1}{\alpha}\right)\cdot O(d\cdot\text{polylog}(d))$.

Work	Label Complexity
[Yona & Rothblum 2018]	$\frac{1}{\epsilon^2}\cdot\frac{1}{\alpha^2}\cdot O(d)$
Our Work	$\text{polylog}\left(\frac{1}{\epsilon},\frac{1}{\alpha}\right)\cdot O(d)$

Main Results

- **Main Theorem:** There is a polynomial-time algorithm that PACF learns $\mathcal{H}$ with the label complexity is $\text{polylog}\left(\frac{1}{\epsilon},\frac{1}{\alpha}\right)\cdot O(d\cdot\text{polylog}(d))$.

Work	Label Complexity
[Yona & Rothblum 2018]	$\frac{1}{\epsilon^2}\cdot\frac{1}{\alpha^2}\cdot O(d)$
Our Work	$\text{polylog}\left(\frac{1}{\epsilon},\frac{1}{\alpha}\right)\cdot O(d)$

- Our work can work with η-adversarial noises and t-sparse halfspaces.

Thank you!