Fair Learning with Private Demographic Data

Hussein Mozannar (MIT)

Mesrob Ohannessian (UIC)

Nati Srebro (TTIC)

ICML 2020

Summary: Problem Setting

- **Common Assumption:** Protected Attributes are fully observed when learning fair classifiers.
- **Problem:** Laws and regulations often prohibit the collection, access and use of the protected attributes in many settings.
- **This Work:** Learning fair classifiers when we have privatized samples of protected attributes and missing attributes.
- **Setting**:
  - Individuals with attributes $\boldsymbol{X}$ *(non-sensitive)*, $\boldsymbol{A}$ *(protected)*, *only* access to Locally Differentially Private $\boldsymbol{Z = Q(./A)}$
  - *Want to enforce group-fairness conditions e.g. Equalized Odds*

Summary: Results

- **Equivalence of non-discrimination**: if predictor $\hat{Y}$ is not a function of X, then non-discrimination w/r Z $\Leftrightarrow$ w/r A
- **2-step Learning Procedure with guarantees.**

Error of optimal fair predictor — Complexity of model

$$\text{err}(\tilde{Y}) \leq_\delta \overbrace{\text{err}(Y^*)} + \frac{Ce^\epsilon}{e^\epsilon - 1}\left(\overbrace{\mathfrak{R}_{n_{\min}}}(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{n_{\min}}}\right)$$

$$\text{disc}(\tilde{Y}, A) \leq_\delta \frac{Ce^{2\epsilon}}{e^{2\epsilon} - 2e^\epsilon + 1}\sqrt{\frac{\log(1/\delta)}{n_{\min}}}$$

Price of Privacy

- **Individual Choice of Reporting:** how to learn and audit when individuals retain the choice to report their attributes.

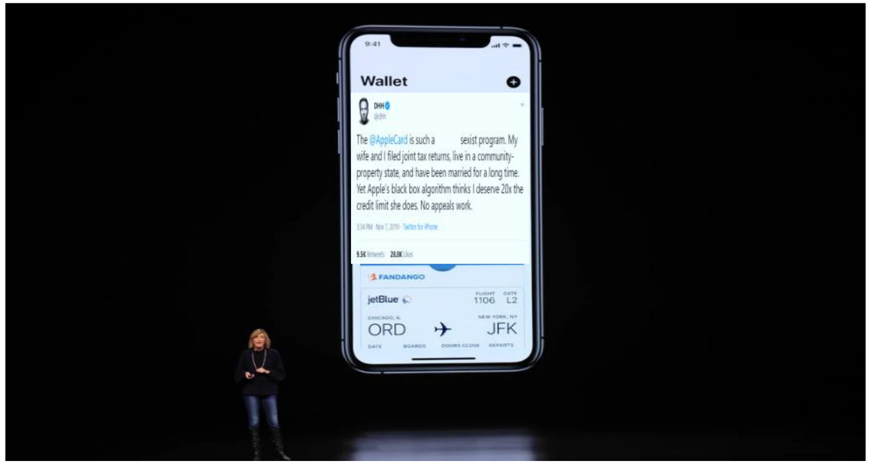
Motivating Example: Apple Card

- Apple Card was found to give wildly differing credit limits for married couples: *two individuals who deserve the same outcome but belong to different demographic groups received different treatments*

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was "sexist" against women applying for credit.



- *Spokesperson:*

"Our credit decisions are based on a customer's creditworthiness and not on factors like gender, race, age, sexual orientation or any other basis prohibited by law"

How can Apple verify and ensure this?

Access to the protected attribute (A)

- Two seemingly opposing societal concerns:

1) Apple cannot force you to disclose sensitive information (**Privacy**)
2) Apple has to prove that it is non-discriminatory (**Fairness**)

- Q1: How can Apple be **unfair** without A?

  Ans: even if *features* are independent of A, learned predictor **can be discriminatory**!

- Q2: How can Apple be **fair** without A?

  Ans: Can rely on proxies which might maybe insufficient and misleading (Kallus et al. 19)

When You Apply For Credit, Creditors May Not...

- Discourage you from applying or reject your application because of your race, color, religion, national origin, sex, marital status, age, or because you receive public assistance.
- Consider your race, sex, or national origin, although you may be asked to disclose this information if you want to. It helps federal agencies enforce anti-discrimination laws. A creditor may consider your immigration status and whether you have the right to stay in the country long enough to repay the debt.

When Deciding To Grant You Credit Or When Setting The Terms Of Credit, Creditors May Not...

- Consider your race, color, religion, national origin, sex, marital status or whether you get public assistance.

Federal Trade Commission. Your equal credit opportunity rights

Fairness in Classification

- Simplified setting: get a credit limit (1) or no limit (0)
- **Goal**: Build predictor $\hat{Y}$ of target $Y \in \{0, 1\}$ based on individuals with features X and protected attribute A that is non-discriminatory.
- **Non-discrimination criteria:**

Equalized Odds [Hardt et al. 2016] :

$$\mathbb{P}(\hat{Y} = 1 | A = a, Y = y) = \mathbb{P}(\hat{Y} = 1 | A = a', Y = y) \; \forall a, a' \in \mathcal{A}, \forall y$$

Accuracy Parity :

$$\mathbb{P}(\hat{Y} \neq Y | A = a) = \mathbb{P}(\hat{Y} \neq Y | A = a) \; \forall a, a' \in \mathcal{A}$$

and many others

Local Differential Privacy

- **Objective**: Find a middle ground where we don't reveal A to Apple but better than Apple relying on proxies.
- **Potential Solution:** individuals release privatized version of A
- Formally let Z be a private version of A defined as $Z = Q(.|A)$

$$Q(z|a) = \begin{cases} \frac{e^{\varepsilon}}{|\mathcal{A}|-1+e^{\varepsilon}} & \text{if } z = a \\ \frac{1}{|\mathcal{A}|-1+e^{\varepsilon}} & \text{if } z \neq a \end{cases}$$

- Parameter ϵ controls privacy, Q is ϵ-DP.
- Data is $S = \{X_i, Y_i, Z_i\}_{i=1}^{n}$ i.i.d.

Related Work on Fairness and Privacy

- Kilbertus et al. (ICML 2018) has explored a secure multiparty computation scheme
- Jagielsky et al. (ICML 2019) notes that that model can leak information about A $\rightarrow$ learn an A-differentially private fair model (achieved in our setting)
- Learning with noisy attributes: Lamy et al. (NeurIPS 2019), Awatchi et al. (AISTATS 2020) and Wang et al. (2020)

Equivalence of non-discrimination

- *Question*: is non-discrimination with respect to *Z (privatized protected attribute)* equivalent to non-discrimination with respect to *A (protected attribute)*?

  Proposition. Yes, if a predictor $\hat{Y}$ is not an explicit function of Z, we have:

$$\begin{aligned} & \mathbb{P}(\hat{Y} = 1 | Z = a, Y = y) = \mathbb{P}(\hat{Y} = 1 | Z = a', Y = y) \\ \iff & \mathbb{P}(\hat{Y} = 1 | A = a, Y = y) = \mathbb{P}(\hat{Y} = 1 | A = a', Y = y) \end{aligned}$$

  Furthermore, this holds for a broad set of group fairness constraints.

Learning Fair Predictors: Approach 1

- **First Approach**: Learn an approximately fair predictor with respect to Z.

$$\tilde{Y} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in S} \mathbf{I}(h(x_i) \neq y_i)$$

$$\text{s.t.} \max_{a,a'} \overbrace{\left| \mathbb{P}^S(h(X) = 1 | Z = a, Y = y) - \mathbb{P}^S(h(X) = 1 | Z = a', Y = y) \right|}^{\text{disc}^S(h, Z)} \leq \alpha_n$$

Note: $\text{disc}(h, Z) = 0 \Leftrightarrow \text{disc}(h, A) = 0$

but $\text{disc}(h, Z) \leq \alpha \implies \text{disc}(h, A) \leq C \cdot \alpha$

- Practically solve using exponentiated gradient reduction for fair classification (Agarwal et al., ICML 2018)

Learning Fair Predictors: Guarantee

Proposition. Let $n_{min} = n \min_{ya} \mathbf{P}_{ya}$, w.p. $1-\delta$, the predictor $\tilde{Y}$ satisfies:

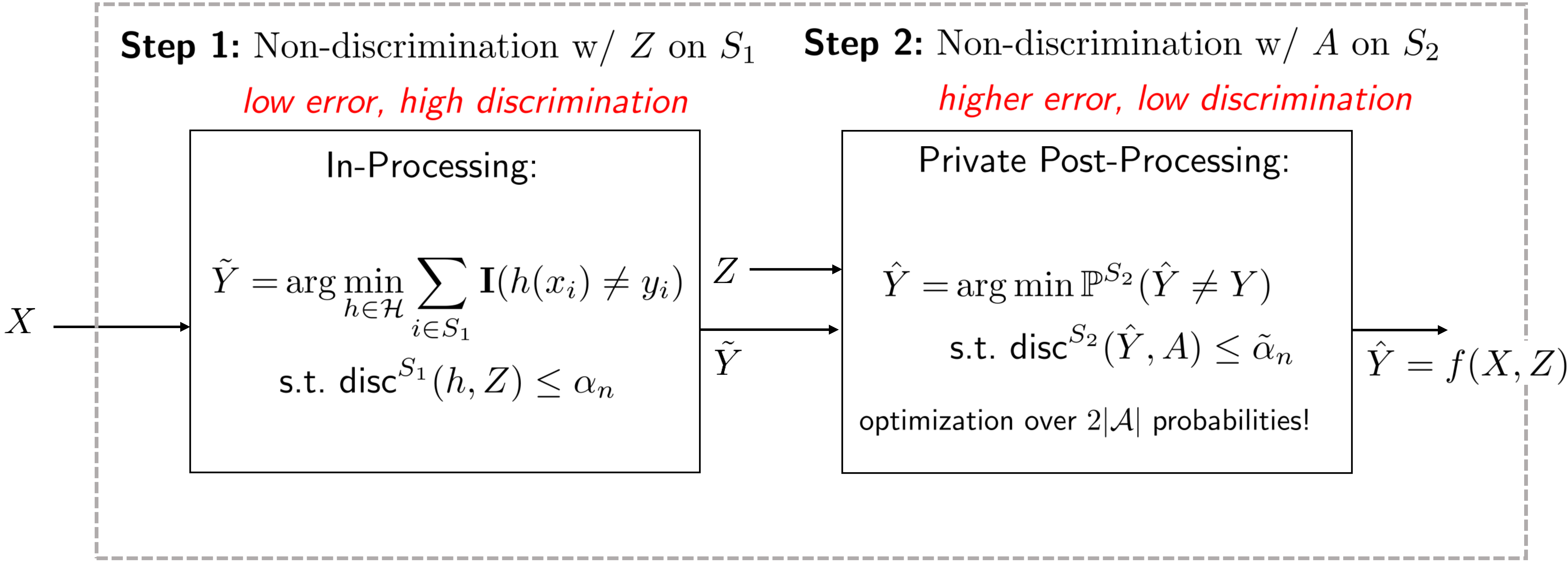
$$\mathrm{err}(\tilde{Y}) \leq_\delta \underbrace{\mathrm{err}(Y^*)}_{\text{Error of optimal fair predictor}} + \underbrace{\mathfrak{R}_n(\mathcal{H})}_{\text{Complexity of model}} + \sqrt{\frac{\log(1/\delta)}{n}}$$

$$\mathrm{disc}(\tilde{Y}, A) \leq_\delta \underbrace{\frac{Ce^\epsilon}{e^\epsilon - 1}}_{\text{Price of Privacy}} \left(\mathfrak{R}_{n_{\min}}(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{n_{\min}}} \right)$$

- Trade-off: Error is not affected by privacy! But fairness is, heavily.

Improving Fairness: Two-step procedure

- Adapt 2-step procedure of Woodworth et al. (COLT 2017), split dataset S into two sets S_1, S_2:



Fair Private Post-Processing

- Post-processing procedure of (Hardt et al., NeurIPS 2016) $\hat{Y}$ operates as follows:

$$\mathbb{P}(\hat{Y} = 1 | \tilde{Y} = \tilde{y}, Z = z)$$

Found by solving the following LP on S_2 with respect to A:

$$\hat{Y} = \arg\min \mathbb{P}^{S_2}(\hat{Y} \neq Y)$$

$$\text{s.t. } \underbrace{|\mathbb{P}^{S_2}(\hat{Y} = 1 | Y = y, A = a) - \mathbb{P}^{S_2}(\hat{Y} = 1 | Y = y, A = a')| \leq \tilde{\alpha}_n}$$

Can satisfy this without knowing A!

- How? Base predictor $\tilde{Y} = h(X)$ can recover all its statistics via inversion and randomize over actual individual's attribute.

Inversion of statistics

- TPR/FPR of 2-step predictor:

$$\mathbb{P}(\tilde{Y}=1|Y=y,A=a)=\mathbb{P}(\tilde{Y}=1|\hat{Y}=0,A=a)\cdot\mathbb{P}(\hat{Y}=0|Y=y,A=a) \quad (1)$$
$$+\mathbb{P}(\tilde{Y}=1|\hat{Y}=1,A=a)\cdot\mathbb{P}(\hat{Y}=1|Y=y,A=a)$$

- Parameters of 2-step predictor:

$$\mathbb{P}(\tilde{Y}=1|\hat{Y}=\hat{y},A=a)=\pi\cdot\mathbb{P}(\tilde{Y}=1|\hat{Y}=\hat{y},Z=a) \quad (2)$$
$$+\sum_{a'\neq a}\bar{\pi}\cdot\mathbb{P}(\tilde{Y}=1|\hat{Y}=\hat{y},Z=a') \quad (\pi=\mathbb{P}(Z=a|A=a),\ \forall a)$$

- TPR/FPR of step-1 predictor:

$$\mathbb{P}(\hat{Y}=1|Y=y,A=a)=\pi\frac{\mathbb{P}(Y=y,A=a)}{\mathbb{P}(Y=y,Z=a)}\cdot\mathbb{P}(\hat{Y}=1|Y=y,Z=a) \quad (3)$$
$$+\sum_{a'\neq a}\bar{\pi}\frac{\mathbb{P}(Y=y,A=a')}{\mathbb{P}(Y=y,Z=a)}\cdot\mathbb{P}(\hat{Y}=1|Y=y,Z=a')$$

2-step procedure: Guarantees

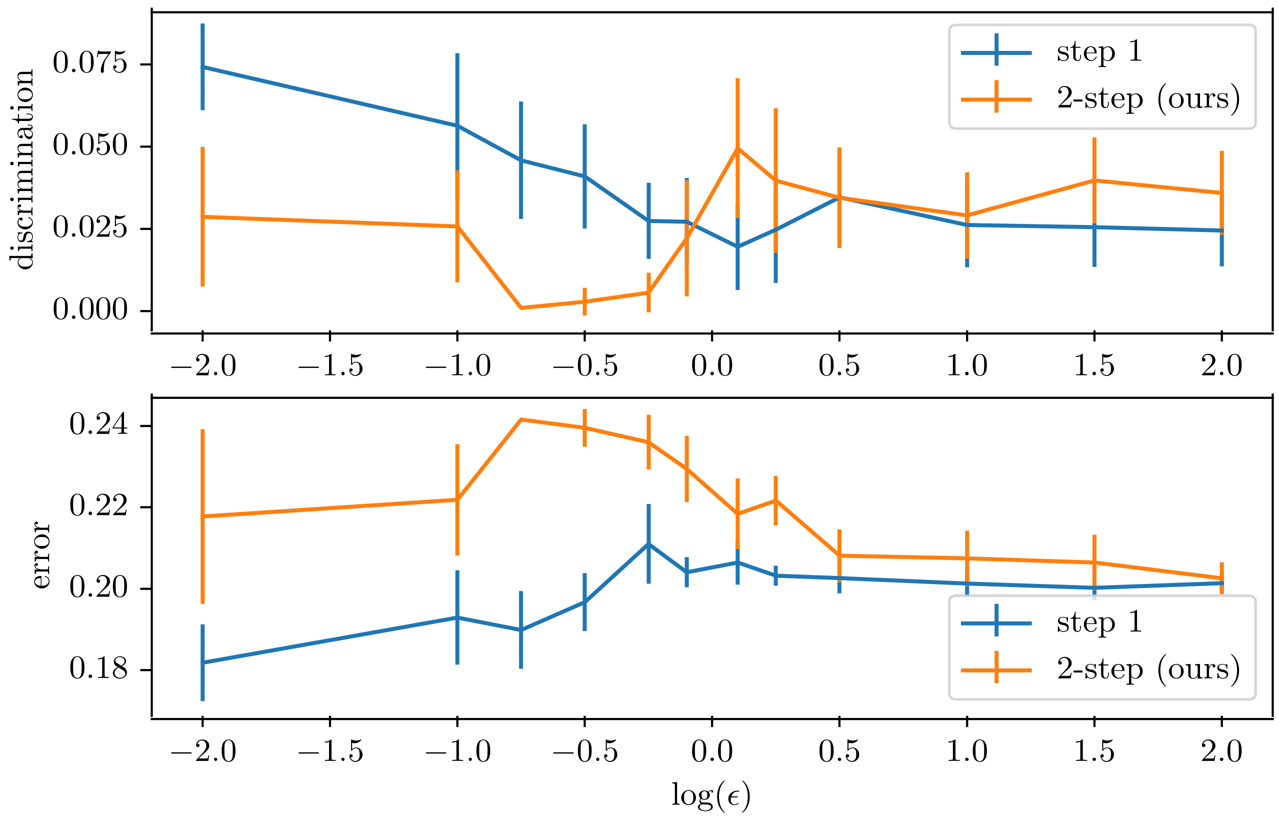
Theorem. The predictor of the 2-step procedure $\hat{Y}$ satisfies:

$$\text{err}(\tilde{Y}) \leq_\delta \underbrace{\text{err}(Y^*)}_{\text{Error of optimal fair predictor}} + \frac{Ce^\epsilon}{e^\epsilon - 1}\left(\underbrace{\mathfrak{R}_{n_{\min}}}_{\text{Complexity of model}}(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{n_{\min}}}\right)$$

$$\text{disc}(\tilde{Y}, A) \leq_\delta \underbrace{\frac{Ce^{2\epsilon}}{e^{2\epsilon} - 2e^\epsilon + 1}}_{\text{Price of Privacy}} \sqrt{\frac{\log(1/\delta)}{n_{\min}}}$$

- Complexity of model disappears from discrimination! However, privacy enters error.

Experimental Illustration



- Plot of discrimination and error of step-1 and 2-step predictors as we vary privacy on the Adult Income dataset using linear predictors.

Individual Choice of Reporting

- **New setting:** Individuals have the choice to either report or not report their protected attribute.
- Let *t(x,y,a)* (reporting probability function) be the probability that an individual *(x,y,a)* chooses to report their protected attributes.

Data: Starting from $S \sim \mathbf{P}^n$, we split into:
$S_\ell = \{(x_1, a_1, y_1), \cdots, (x_{n_\ell}, a_{n_\ell}, y_{n_\ell})\}$ (individuals who report)
and $S_u = \{(x_1, y_1), \cdots, (x_{n_u}, y_{n_u})\}$ (individuals who do not report).

How can we measure discrimination?

- **Naive Approach:** measure discrimination of predictor based on S_ℓ
- When does this work?

Proposition. Let T be a $r.v.$ s.t. $\mathbb{P}(T(x, y, a) = 1) = t(x, y, a)$ (if a person reports), then if T and $\hat{Y}$ are indepedent given (Y, A):

$$\mathrm{disc}^{S_\ell}(\hat{Y}, A) \to_p \mathrm{disc}(\hat{Y}, A) \quad \text{(Equalized Odds)}$$

example: if $t(x, y, a) := t(y, a)$ and $\hat{Y} = h(X)$, then the indepdence assumption holds.

Learning using missing data

- Modify reductions approach using a two-dataset Lagrangian:

$$L^{S_u,S_\ell}(Q,\boldsymbol{\lambda}) = \mathrm{err}^{S_u \cup S_\ell}(Q) + \boldsymbol{\lambda}^\top \underbrace{\left(M\boldsymbol{\gamma}^{S_\ell}(Q) - \alpha\mathbf{1}\right)}_{\text{discrimination violation}}.$$

Learner Q uses all the data and auditor $\boldsymbol{\lambda}$ uses only S_ℓ

Proposition. The predictor $\hat{Y}$ learned using L^{S_u,S_ℓ} satisfies:

$$\text{err}(\hat{Y}) \leq_\delta \text{err}(Y^*) + \mathfrak{R}_{n_u+n_\ell}(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{n_u + n_\ell}}$$

$$\text{disc}(\hat{Y}, A) \leq_\delta \mathfrak{R}_{n_\ell \min_{ya} \mathbf{P}_{ya} \mathbf{T}_{ya}}(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{n_\ell \min_{ya} \mathbf{P}_{ya} \mathbf{T}_{ya}}}$$

Future Work and Open Questions

- **Better Discrimination Guarantees without using A**. Can we obtain the same learning guarantees outlined here without access to Z (or A) at test time?
- **Can we leverage unlabeled data to improve discrimination?**
- **What are the limits of what we can do without any access to A?**