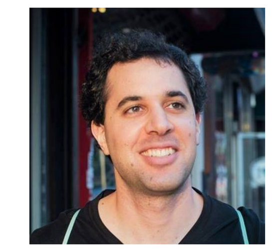
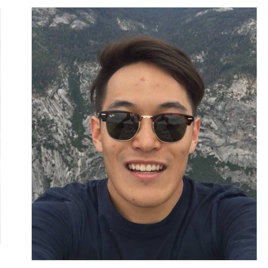
Metric-Free Individual Fairness In Online Learning

Yahav Bechavod, Hebrew University,
Christopher Jung, University of Pennsylvania,
Steven Wu, Carnegie Mellon University


NEURAL INFORMATION PROCESSING SYSTEMS

Algorithmic Fairness

- Most of previous work focuses on group fairness
- E.g. $statistic\ (group_1) = statistic(group_2)$ where statistic can be FPR, positive predictive value, etc and groups are defined according to the protected attributes
- Easy to operationalize and reason about but weak guarantees at the individual level

Individual Fairness

- "Similar Individuals should be treated similarly"

$$\underbrace{|\pi(x_1) - \pi(x_2)|}_{\text{Difference in predictions}} \leq \underbrace{d(x_1, x_2)}_{\text{Distance}}$$

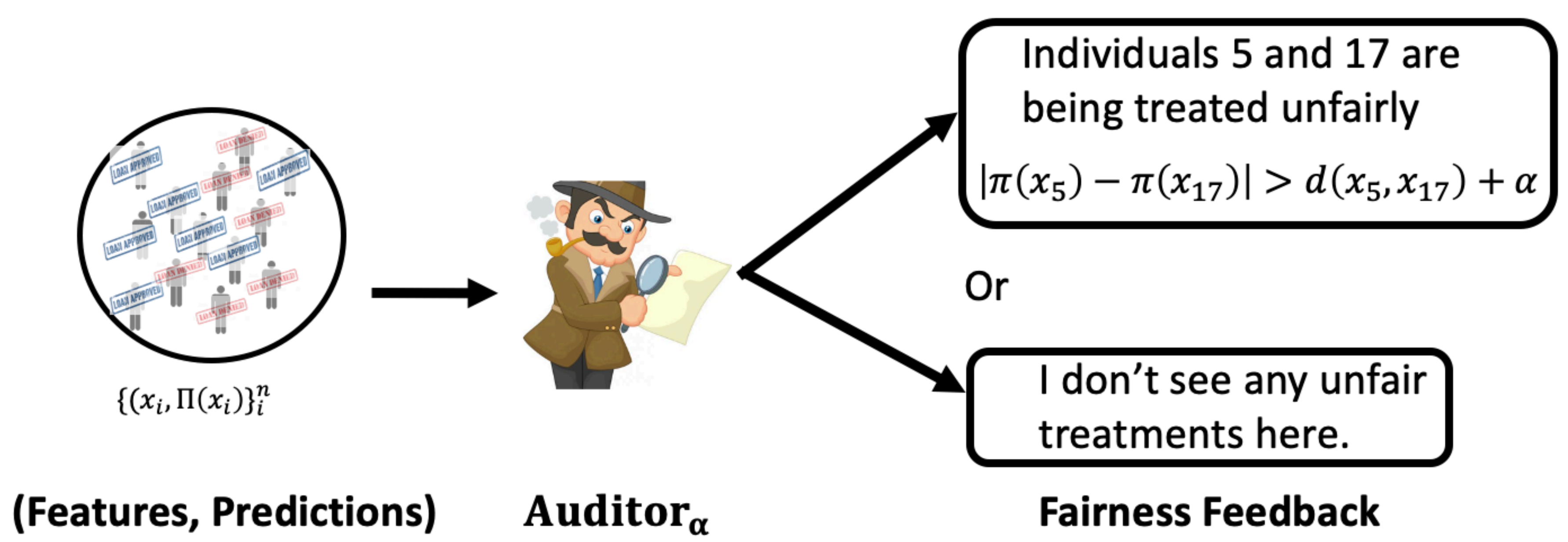
$\pi : \mathcal{X} \to [0, 1]$ "soft" predictor

$d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$

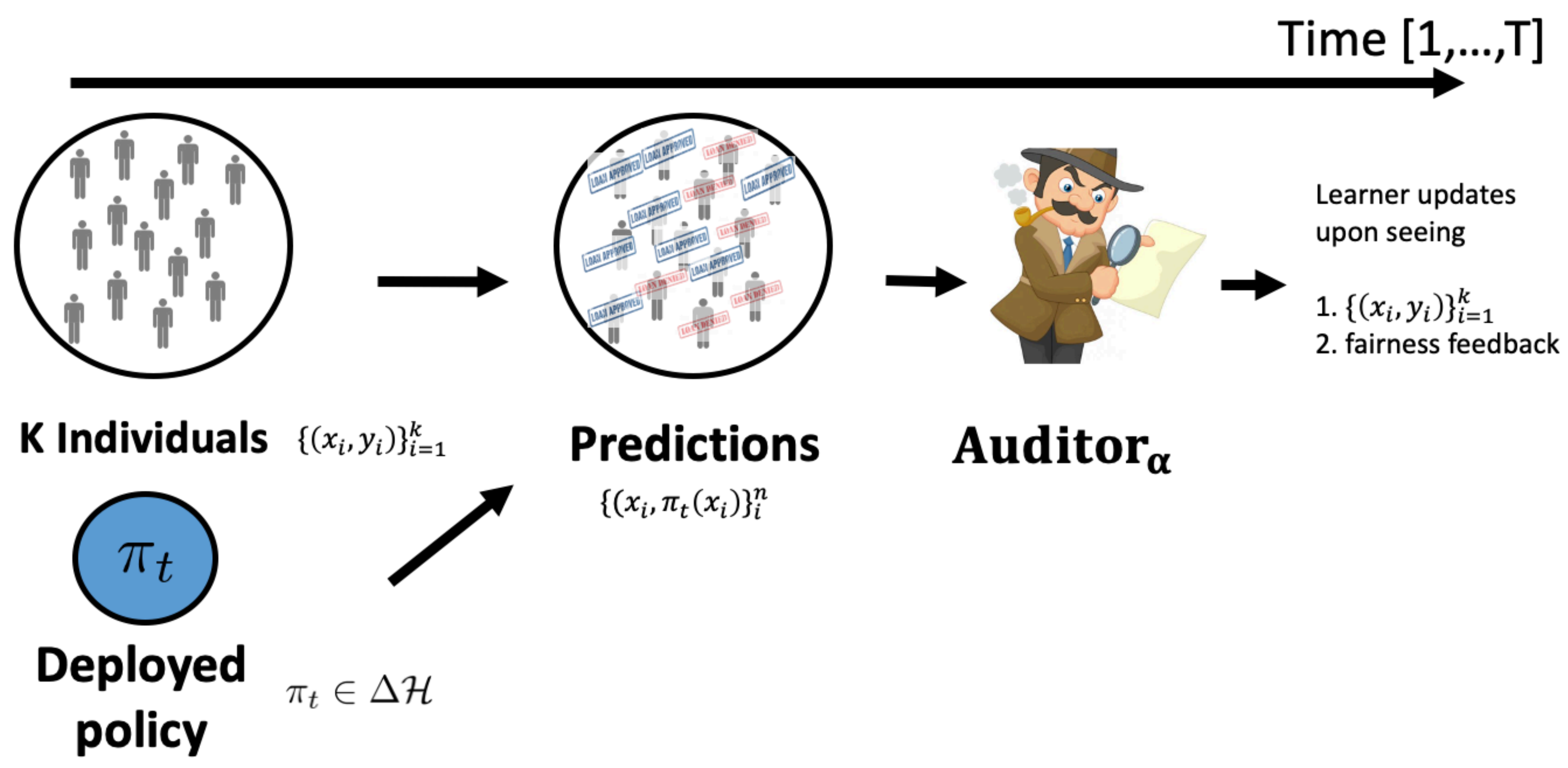
- Hard to enunciate what the metric d should be exactly even for domain experts

Fairness Auditor

- Rely on an auditor who can detect violations of individual fairness



Online learning



Comparison to previous work

1. No parametric assumption on the underlying metric of the auditor
   d doesn't need to satisfy triangle inequality.
2. No need for numerical distance queries.
   Ilvento (2018) suggests learning through distance queries between individuals.
3. Single fairness feedback
   Gillen et al. (2018) requires **all fairness violations** to be reported by the auditor.
   We require only **one fairness violation** to be reported by the auditor.

Objectives

1. Fairness loss

$$\text{Fair}Loss = \sum_t 1[\text{Auditor}_\alpha \text{ complains on day t}]$$

2. Classification error against other α —fair policies

$$Regret_{Misclassification} = \sum_t E_{f\sim\pi_t}\,[[1[f(x_t) \neq y_t]\,] - \min_{\pi^* \in \Pi_{\alpha-fair}} \sum_t E_{f\sim\pi^*}\,[[1[f(x_t) \neq y_t$$

Results

(1) Adversarial arrival

Algorithm 1: Online Fair Batch Classification FAIR-BATCH

for $t = 1, \ldots, T$ **do**
  Learner deploys π^t
  Environment chooses $(\bar{x}^t, \bar{y}^t)$
  Environment chooses the pair ρ^t
  $z^t = (\bar{x}^t, \bar{y}^t) \times \rho^t$
  Learner incurs misclassfication loss $\text{Err}(\pi^t, z^t)$
  Learner incurs fairness loss $\text{Unfair}(\pi^t, z^t)$
end

→

Algorithm 2: Online Batch Classification BATCH

for $t = 1, \ldots, T$ **do**
  Learner deploys π^t
  Environment chooses $z^t = (\bar{x}^t, \bar{y}^t)$
  Learner incurs misclassification loss $\text{Err}(\pi^t, z^t)$
end

Regret-preserving reduction by adding in some carefully chosen examples to the batch.

Inherit the regret of the algorithm in online contextual learning without fairness constraints.

$$Regret_{Misclassification}, \text{Fair}Loss \leq Regret(online\ algorithm)$$

$$\Downarrow$$

1. No-regret with respect to classification error

$$Regret_{Misclassification} = o(T)$$

2. Sublinear Fairness Loss

$$\text{Fair}Loss = o(T)$$

(2) Stochastic arrival

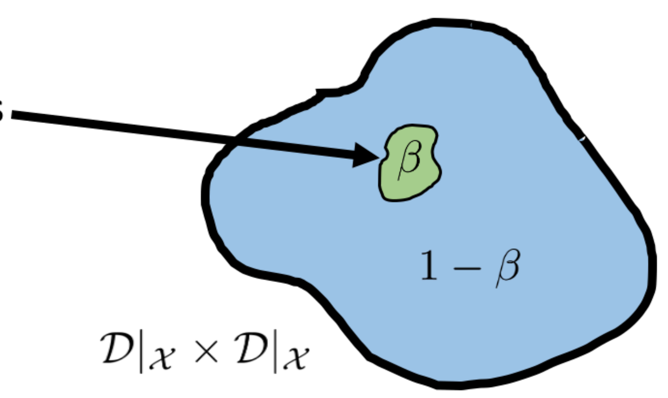
We consider the **average policy** deployed by the algorithm over time.

1. Misclassification error generalization
Through vanilla online-to-batch conversion

2. Fairness generalization

(α, β)-Fairness of π:

$$\Pr_{(x,x')\sim\mathcal{D}|_\mathcal{X}\times\mathcal{D}|_\mathcal{X}}[|\pi(x) - \pi(x')| > d(x, x') + \alpha] \leq \beta.$$



Probability of drawing a pair on which π has an α-violation is smaller than β.

Thm. (simplified):
Average policy over time is $(\underbrace{\alpha + \mathcal{O}(T^{-\frac{1}{4}})}_{\text{Violation Size}}, \underbrace{\mathcal{O}(T^{-\frac{1}{4}})}_{\text{Violation Frequency}})$-fair.

Conclusion

1. **Metric-Free**: removed classical metric assumption
2. **Easy Auditing**: No complex, numerical queries / existence of fairness violations /single fairness violation reported
3. **General**: no parametric assumption on hypothesis class / metric
4. **Efficient**: oracle-efficient