Metric-Free Individual Fairness in Online Learning

Yahav Bechavod

Hebrew University of Jerusalem

November 26, 2020

Talk based on joint work with Christopher Jung and Steven Wu,
"Metric-Free Individual Fairness in Online Learning", NeurIPS 2020.

Christopher Jung
University of Pennsylvania

Steven Wu
Carnegie Mellon University

Machine Learning algorithms are becoming more potent than ever before, **altering the way tasks have "traditionally" been performed.**

As part of the change, algorithms are now intensively involved in decision making mechanisms that **crucially affect people's lives.**

PARTNER CONTENT CYNTHIA RUDIN, MIT SLOAN

PREDICTIVE POLICING: USING MACHINE LEARNING TO DETECT PATTERNS OF CRIME

Image: lydia_shiningbrightly/Flickr

Trying to detect specific patterns of crime and criminal behavior is extremely challenging. Crime analysts can spend countless hours sifting through data to determine whether a crime fits into a known pattern and to discover new patterns. Once a pattern is detected, the information can be used to predict, anticipate and prevent crime.

Artificial intelligence will help determine if you get your next job

AI is being used to attract applicants and to predict a candidate's fit for a position. But is it up to the task?

By Rebecca Heilweil | Dec 12, 2019, 8:00am EST

Berkeley News

Research ▾ People ▾ Campus & communit

POLITICS & SOCIETY, RESEARCH, TECHNOLOGY & ENGINEERING

Algorithms are better than people in predicting recidivism, study says

By Edward Lempinen | FEBRUARY 14, 2020

Tweet Share 575 Email Print

Forbes

The Algorithm That Beats Your Bank Manager

Parmy Olson Former Staff

AI

AI, robotics and the digital transformation of European business.

This article is more than 9 years old.

Image via CrunchBase

Can computers make better decisions than humans? One tech firm says they already do when it comes to lending money.

Wonga.com is a short-term lending Web site that uses an algorithm and thousands of pieces of information about its customers in the public domain, to decide in a few seconds whether to grant a short-term loan. Not to be dismissed as another loan shark dressed in Web 2.0 clothing,

In all of these settings, algorithms are now offering numerous benefits in terms of **accuracy, efficiency and cost-savings**.

However, multiple recent reports have raised **concerns of bias and unfairness**.


PRO PUBLICA
Donate
Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

This article is more than **4 years old**

A beauty contest was judged by AI and the robots didn't like dark skin

The first international beauty contest decided by an algorithm has sparked controversy after the results revealed one glaring factor linking the winners

most viewed


Arrests in Washing
Trump supporters
assemble, rejecting
victory


Judge rules Daca
suspension invalid
Homeland Security
office illegally


No-deal fears rise a
Johnson 'least willi
budge on Brexit'

HR DIVE Deep Dive Opinion Library Events Topics

BRIEF

MIT: Hiring algorithm design could impact candidate diversity, quality

White papers Books Webinars

CENTRAL BANKING

Financial Stability Fintech Economics Governance Reserves Currency Benchmarking Dire

Machine learning algorithms could increase ethnic bias – research

esk

In an attempt to better understand the problem of bias, research in **Algorithmic Fairness** has emerged in recent years.

A significant amount of work in the field is aimed at designing **highly-performing algorithms with fairness guarantees.**

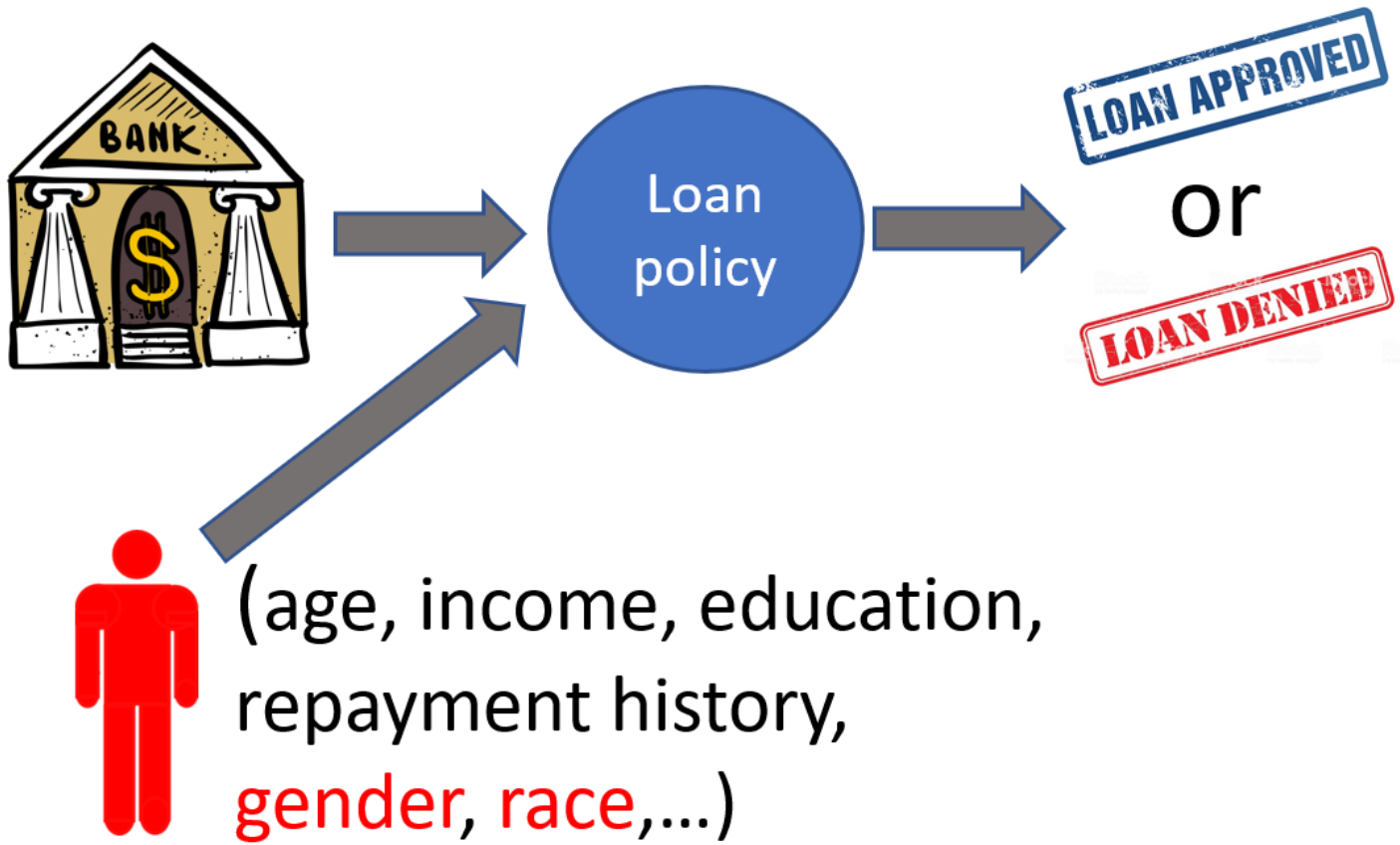
What is "Fair"?

An overwhelming majority of the work in Algorithmic Fairness considers **(statistical) Group Fairness** definitions.

(Statistical) Group Fairness

Example: **Loan Approvals**

For incoming loan applicants, predict whether each individual will **repay** or **default** on payments.



(Statistical) Group Fairness

- Statistical Parity:
  Approve loans for same fraction of male/female applicants.
- Equal Error Rates:
  Same probability of mistake on male/female applicants.
- Equal False Positive Rate:
  Same probabilty of approving a bad applicant across male/female applicants.
- Equal False Negative Rate:
  Same probabilty of rejecting a good applicant across male/female applicants.

(Statistical) Group Fairness Definitions

- Relatively easy to operationalize.
- However: Very weak guarantees **for individuals**.

Group Fairness Breaks on Individual Level

Group Fairness Breaks on Individual Level

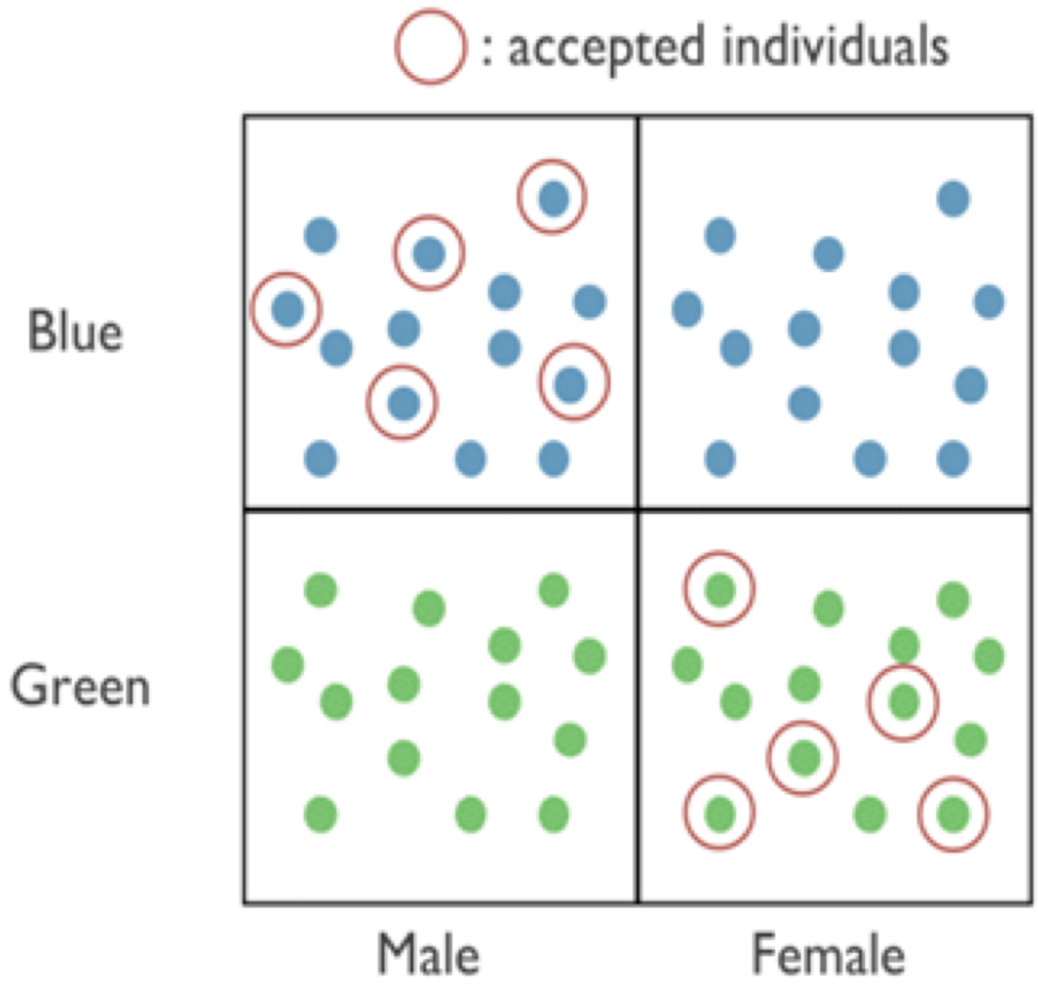
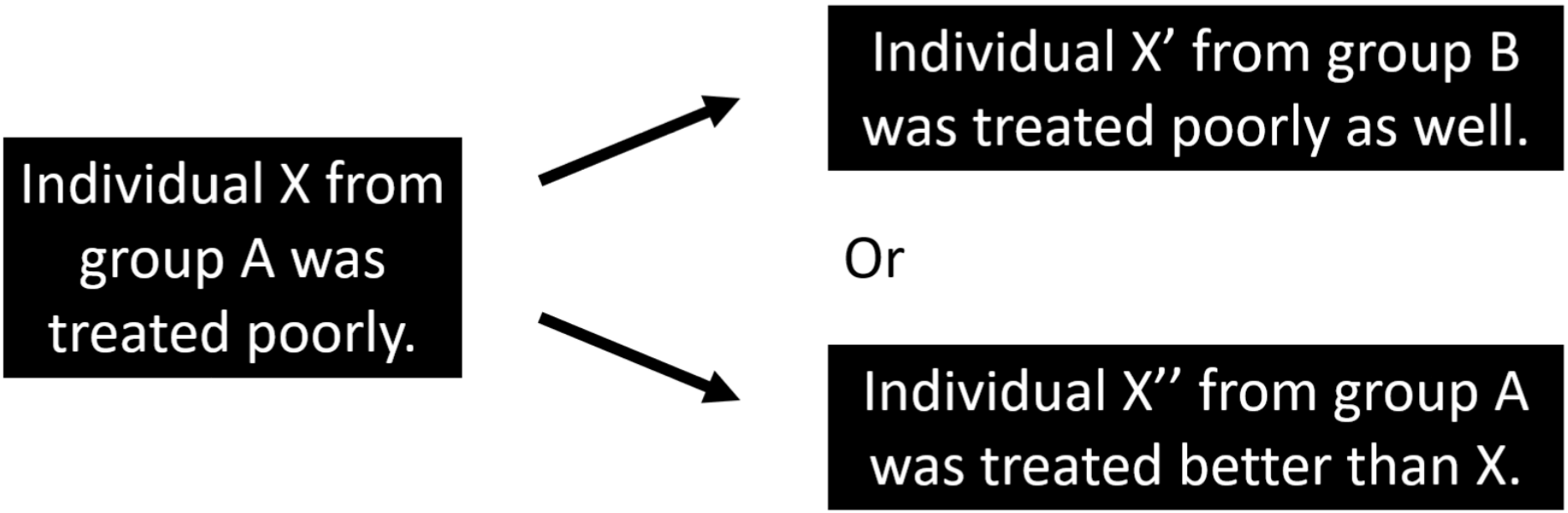


Figure: Fairness Gerrymandering: A Toy Example [Kearns et al., 2018]


LOAN APPLICATION
REJECTED

Group Fairness Breaks on Individual Level

Very weak guarantees **for individuals**:



Individual Fairness

Dwork et al. 2011: "Similar individuals should be treated similarly."

$$\underbrace{|h(x) - h(x')|}_{\text{Diff. in predictions}} \leq \underbrace{d(x, x')}_{\textit{Distance}}$$

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

Assumption: Similarity metric between individuals.

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

Example: "How similar are loan applicants x, x' ?"

Challenges in Operationalizing Individual Fairness

Although desirable in many settings, ever since it's introduction, Individual Fairness was largely not operationalizable.

Problem: Similarity metric is often **unavailable**.

- Unclear where such metric can be found.
- People have different opinions of who are similarly situated in the context of specific tasks.
- Even if an individual has a clear idea of which individuals are similarly situated, an exact mathematical formula for the metric might be **difficult to enunciate**.
- What if the given similarity function is not a metric?

Difficulty of Enunciating a Metric

"What is the **exact** formula that measures similarity for loan applicants?"

"Hard to tell..."

Difficulty of Answering Numerical Queries

"What is the distance between individuals #5 and #17?"

"Still Difficult for me to answer exactly."

Our Approach: Human Auditor for Fairness Violations

"Can you spot a pair of **similar** individuals who were treated **very differently**?"

"Yes. Individuals #5 and #17."

Auditor **"knows unfairness when he sees it."**

Auditor

Our Contributions

1. Introduce new human auditor feedback model based on reported fairness violations.
   - **Metric-Free:** Removes classical metric assumption.
   - **Easy Auditing:**
     - ★ No complex, numerical queries.
     - ★ Auditor only required to report the **existence** of fairness violations.
     - ★ Auditor only required to report **a single** violation for each batch.

Our Contributions

2. Introduce a novel online learning algorithm, which is,
   - **General:**
     - ★ No parametric assumptions on hypothesis class.
     - ★ No parametric assumptions on similarity function.
   - **Efficient:** Can be implemented efficiently using access to optimization oracle.

   and provides, using violations feedback, the following guarantees:
   1. (Adversarial Arrivals) No-Regret for Accuracy, Individual Fairness.
   2. (Stochastic Arrivals) Generalization for Accuracy, Individual Fairness.

Outline

- Human auditor model based on fairness violations.
- Online Fair Batch Classification Setting.
- No-Regret for Accuracy, Fairness,
- Generalization for Accuracy, Fairness.

Outline

- Human auditor model based on fairness violations.
- Online Fair Batch Classification Setting.
- No-Regret for Accuracy, Fairness,
- Generalization for Accuracy, Fairness.

Model and Definitions

- $\mathcal{X}$ instance space.
- $\mathcal{Y} = \{0, 1\}$ label space.
- $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$ hypothesis class.
- Assume $\mathcal{H}$ contains a constant hypothesis – i.e. h such that $h(x) = 0$ for all $x \in \mathcal{X}$.
- We allow for convex combinations of hypotheses for the purpose of randomizing the prediction and denote the simplex of hypotheses by $\Delta\mathcal{H} : \mathcal{X} \to [0, 1]$.
- For each prediction $\hat{y} \in \mathcal{Y}$ and true label $y \in \mathcal{Y}$, there is an associated misclassification loss, $\ell(\hat{y}, y) = 1(\hat{y} \neq y)$.
- We overload notation and write, for $\pi \in \Delta\mathcal{H}$:

$$\ell(\pi(x), y) = (1 - \pi(x)) \cdot y + \pi(x) \cdot (1 - y) = \mathop{\mathbb{E}}_{h \sim \pi}[\ell(h(x), y)].$$

Individual Fairness

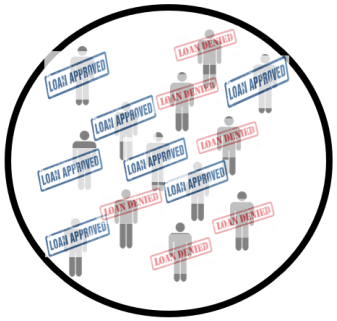
- We assume that there is a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ which captures the distance between individuals in $\mathcal{X}$.
- d need not necessarily be a **metric**:
  - d doesn't have to satisfy the triangle inequality.
  - The only two requirements on d is that it is always non-negative and symmetric.

Definition (α-fairness violation)

We say policy π has an α-fairness violation on pair (x, x') if

$$|\pi(x) - \pi(x')| > d(x, x') + \alpha.$$

Auditor



$\{(x_i, \Pi(x_i)\}_i^n$

(Features, Predictions)

Auditor$_\alpha$

Individuals 5 and 17 are being treated unfairly

$|\pi(x_5) - \pi(x_{17})| > d(x_5, x_{17}) + \alpha$

Or

I don't see any unfair treatments here.

Fairness Feedback

Auditor

Definition (Auditor $\mathcal{J}_\alpha$)

An (possibly stateful) auditor $\mathcal{J}_\alpha$ takes in a *reference set* $S = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ and a policy π. Then, it outputs ρ which is:

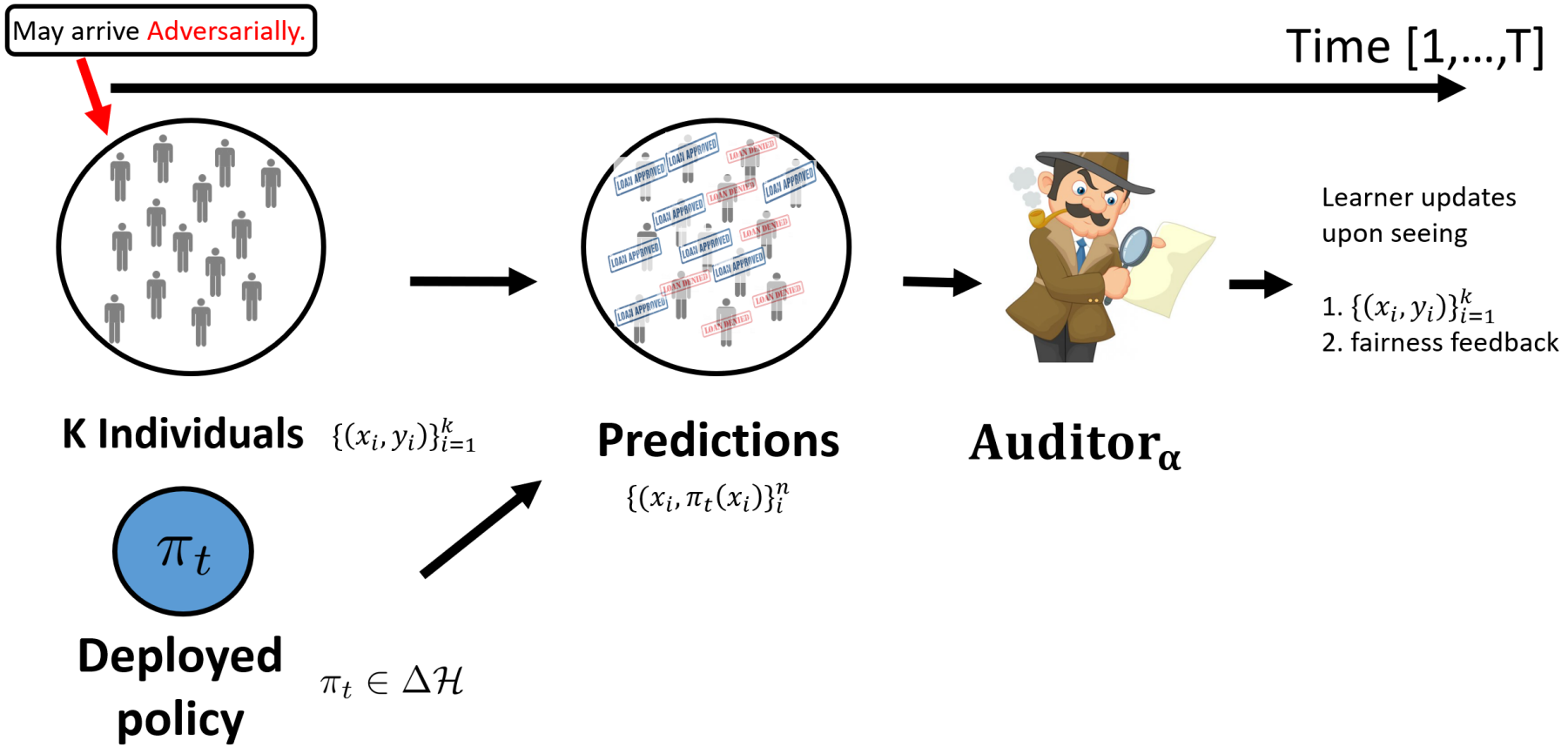
$$\mathcal{J}_\alpha(S, \pi) = \begin{cases} \rho = (\rho_1, \rho_2) & \text{if} \quad \exists \rho_1, \rho_2 \in [n] \ s.t. \\ & \qquad \pi(x_{\rho_1}) - \pi(x_{\rho_2}) - d(x_{\rho_1}, x_{\rho_2}) - \alpha > 0 \\ \rho = null & \text{otherwise} \end{cases}$$

If there exist **multiple** pairs with an α-violation, the auditor may select one **arbitrarily**.

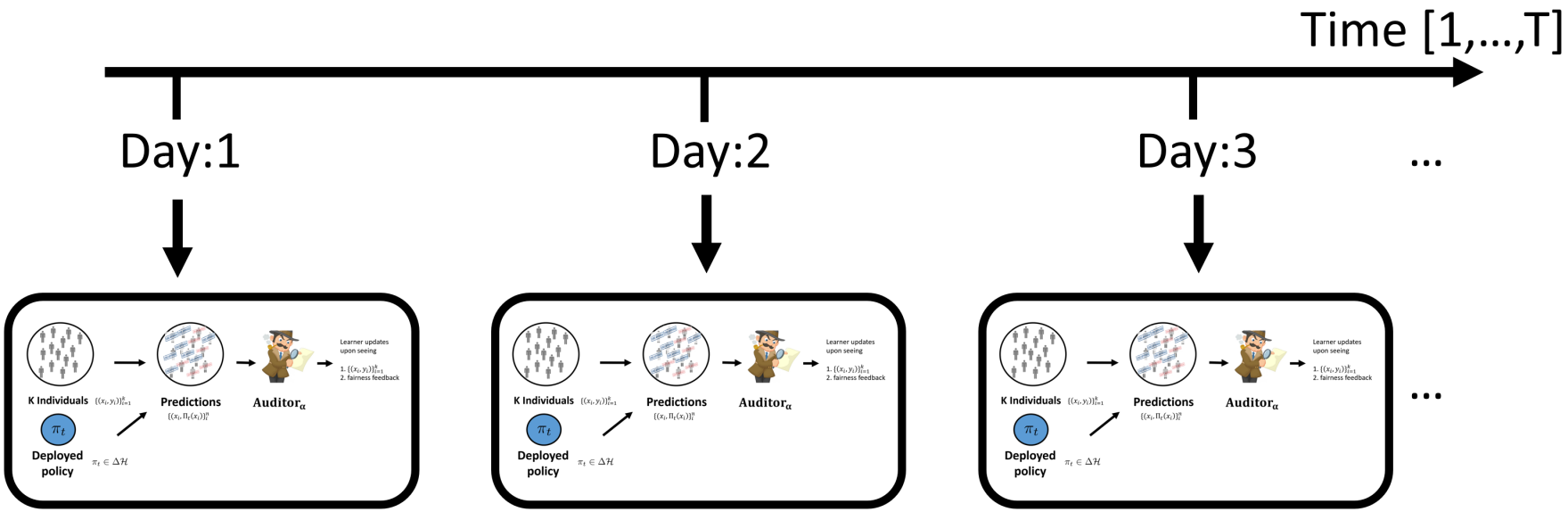
Outline

- Human auditor model based on fairness violations.
- Online Fair Batch Classification Setting.
- No-Regret for Accuracy, Fairness,
- Generalization for Accuracy, Fairness.

Setting: Online Fair Batch Classification



Setting: Online Fair Batch Classification



Online Fair Batch Classification

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 batch misclassfication loss $\mathrm{Err}(\pi^t, z^t)$

Learner incurs fairness loss $\mathrm{Unfair}(\pi^t, z^t)$

end

Where $(\bar{x}^t, \bar{y}^t) = (x_\tau^t, y_\tau^t)_{\tau=1}^k$.

Online Fair Batch Classification

Definition (Misclassification Loss)

The (batch) misclassification loss Err is

$$\mathrm{Err}(\pi, z^t) = \sum_{\tau=1}^{k} \ell(\pi(x_\tau^t), y_\tau^t).$$

Definition (Fairness Loss)

The α-fairness loss Unfair_α is

$$\mathrm{Unfair}_\alpha(\pi, z^t) = \begin{cases} 1\left(\pi(x_{\rho_1^t}^t) - \pi(x_{\rho_2^t}^t) - d(x_{\rho_1^t}^t, x_{\rho_2^t}^t) - \alpha > 0\right) & \rho^t = (\rho_1^t, \rho_2^t) \\ 0 & \text{otherwise} \end{cases}$$

Outline

- Human auditor model based on fairness violations.
- Online Fair Batch Classification Setting.
- No-Regret for Accuracy, Fairness,
- Generalization for Accuracy, Fairness.

Regret

Definition (Algorithm $\mathcal{A}$)

An algorithm $\mathcal{A} : (\Delta\mathcal{H} \times \mathcal{Z})^* \to \Delta\mathcal{H}$ takes in its past history $(\pi^\tau, z^\tau)_{\tau=1}^{t-1}$ and deploys a policy $\pi^t \in \Delta\mathcal{H}$ at every round $t \in [T]$.

Definition (Regret)

For some $Q \subseteq \Delta\mathcal{H}$, the regret of algorithm $\mathcal{A}$ with respect to some loss $L : \Delta\mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ is denoted as $\text{Regret}^L(\mathcal{A}, Q, T)$, if for any $(z_t)_{t=1}^T$,

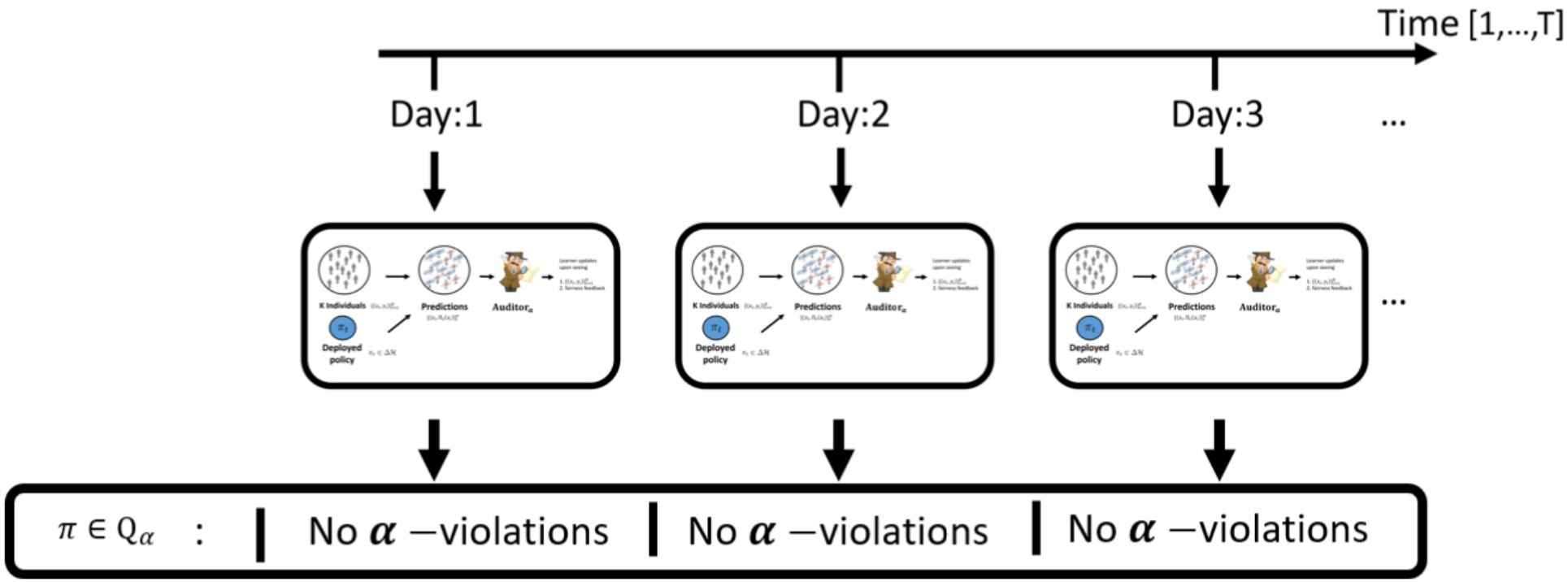
$$\sum_{t=1}^{T} L\left(\pi^t, z^t\right) - \inf_{\pi^* \in Q} \sum_{t=1}^{T} L\left(\pi^*, z^t\right) \leq \text{Regret}^L(\mathcal{A}, Q, T),$$

where $\pi^t = \mathcal{A}((\pi^j, z^j)_{j=1}^{t-1})$.

What Should We Compare to?

We wish to compare performance with the **highest-performing** policy which is also **individually fair**.

Baseline: $Q_\alpha = \{\pi \in \Delta\mathcal{H} : \pi \text{ is } \alpha\text{-fair on } \bar{x}^t \text{ for all } t \in [T]\}$.



What Should We Compare to?

- **Baseline**: All policies Q_α that are α-fair on $\bar{x}^t$ for all $t \in [T]$.
- **Algorithm's Fairness Level:** Since environment/auditor may report violations of magnitude arbitrarily close to α, we will allow additional slack ϵ.
  - Algorithm will only be penalized for $(\alpha + \epsilon)$-fairness violations.
  - For human sensitivity levels, we can think of α, ϵ as constants.

In the Paper

An algorithm $\mathcal{A}$ such that for any $(z^t)_{t=1}^T$,

1. $\text{Regret}^{\text{Err}}_{\text{FAIR-BATCH}}(\mathcal{A}, Q_\alpha, T) = o(T)$.
2. $\sum_{t=1}^T \text{Unfair}_{\alpha+\epsilon}(\pi^t, z^t) = o(T)$.

Proof Idea:

1. Reduce Online Fair Batch Classification to "standard" Adversarial Online Learning, by creating a "fake" stream of examples, according to original stream + auditor's feedback.
2. Upper bound $\text{Regret}^{\text{Err}}_{\text{FAIR-BATCH}}(\mathcal{A}, Q_\alpha, T)$, $\sum_{t=1}^T \text{Unfair}_{\alpha+\epsilon}(\pi^t, z^t)$ by the regret of an algorithm for Adversarial Online Learning.

No-Regret Guarantees

Theorem

If the separator set S for $\mathcal{H}$ is of size s, then CONTEXT-FTPL *achieves the following misclassification and fairness regret in the online fair batch classification setting:*

$$\text{Regret}^{Err}_{\text{FAIR-BATCH}}(\mathcal{A}, Q_\alpha, T) \leq O\left(\left(\frac{sk}{\epsilon}\right)^{\frac{3}{4}} \sqrt{T \log(|\mathcal{H}|)}\right)$$

$$\sum_{t=1}^{T} Unfair_{\alpha+\epsilon}(\pi^t, z^t) \leq O\left(\left(\frac{sk}{\epsilon}\right)^{\frac{3}{4}} \sqrt{T \log(|\mathcal{H}|)}\right)$$

Outline

- Human auditor model based on fairness violations.
- Online Fair Batch Classification Setting.
- No-Regret for Accuracy, Fairness,
- Generalization for Accuracy, Fairness.

So far...

- Adversarial arrivals setting.
- What if we wish to, at some point, deploy the learned policy?
- Realistically, individuals are not expected to show up in adversarial fashion...

Generalization

- We will assume the existence of an (unknown) data distribution from which individual arrivals are drawn:

$$\{\{(x_\tau^t, y_\tau^t)\}_{\tau=1}^k\}_{t=1}^T \sim_{i.i.d.} \mathcal{D}^{Tk}$$

- We wish to output a policy for which we can prove generalization guarantees over $\mathcal{D}$, for **both Accuracy and Fairness**.

Generalization

- **Accuracy Generalization:** Relatively straightforward, since we receive full, unbiased feedback.

- **Fairness Generalization:** More challenging.
  - **Limited Feedback #1:** Auditor is only required to report a **single** fairness violation, though multiple ones may exist.
  - **Limited Feedback #2:** In the worst case, auditor's fairness feedback applies **only to the deployed policy**.
  - **Adaptivity:** Even in stochastic arrivals setting, auditor is adaptive, and may point to **any violating pair**.

Generalization

- We **cannot** use a uniform convergence argument.
- Instead, we will consider the **average policy over time**, and rely on regret guarantees to upper bound Accuracy, Fairness generalization error.

Definition (Average Policy)

Let π^t be the policy deployed by the algorithm at round t. The average policy π^{avg} is defined by:

$$\forall x : \pi^{avg}(x) = \frac{1}{T}\sum_{t=1}^{T}\pi^t(x)$$

Accuracy Generalization

Theorem (Accuracy Generalization)

With probabilty $1-\delta$*, the misclassification loss of* π^{avg} *is upper bounded by*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\pi^{avg}(x),y)\right] \leq \inf_{\pi\in Q_\alpha} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\pi(x),y)\right] + \ldots$$

$$\ldots + \frac{1}{kT}\text{Regret}^{C,\alpha,\mathcal{J}_{\alpha+\epsilon}}(\mathcal{A}, Q_\alpha, T) + \sqrt{\frac{8\ln\left(\frac{4}{\delta}\right)}{T}}$$

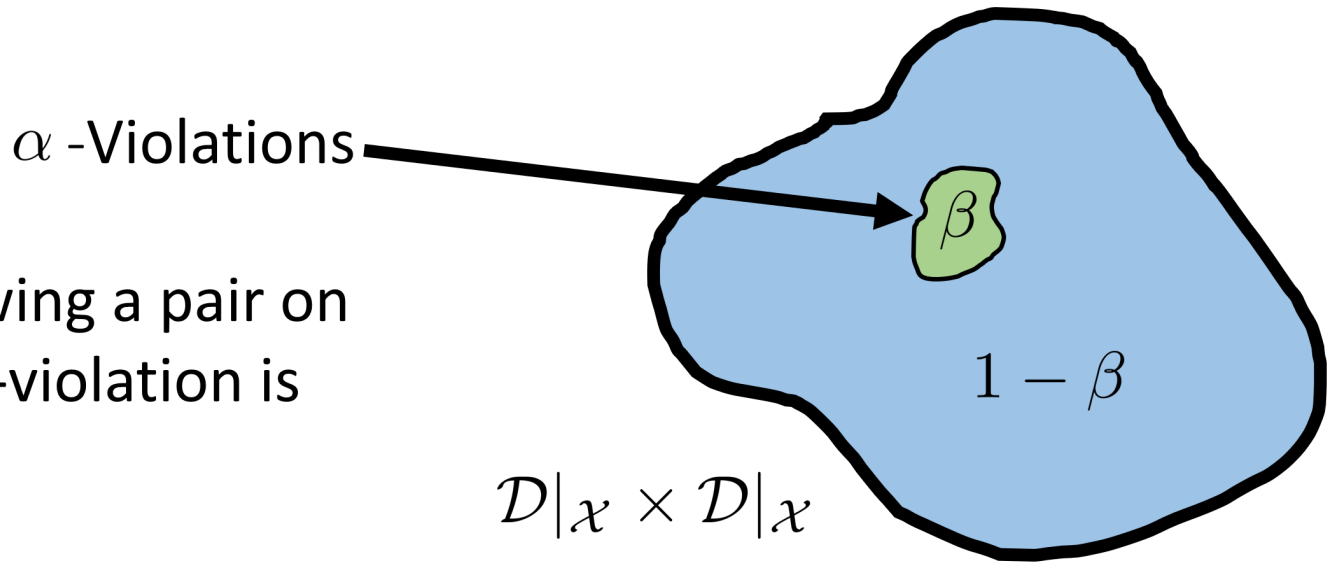
Fairness Generalization

Definition ((α, β)-fairness)

Assume $\alpha, \beta > 0$. A policy is π is said to be (α, β)-fair on distribution $\mathcal{D}$, if

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

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



Fairness Generalization

Theorem (Fairness Generalization)

Assume that for all t, π^t is (α, β^t)-fair $(0 \leq \beta^t \leq 1)$. With probability $1-\delta$, for any integer $q \leq T$, π^{avg} is $(\alpha + \frac{q}{T}, \beta^*)$-fair where

$$\beta^* = \frac{1}{q}\left(\text{Regret}^{C,\alpha,\mathcal{J}_{\alpha+\epsilon}}(\mathcal{A}, Q_\alpha, T) + \sqrt{2T \ln\left(\frac{2}{\delta}\right)}\right).$$

Proof Sketch:

1. Upper bound the probability of an $(\alpha + \frac{q}{T})$-fairness violation using $\sum_{t=1}^{T} \beta^t$.
2. Upper bound $\sum_{t=1}^{T} \beta^t$ using the regret guarantee of the algorithm.

Fairness Level of Convex Combination of Policies

1. Upper bound the probability of an $(\alpha + \frac{q}{T})$-fairness violation using $\sum_{t=1}^{T} \beta^t$.

Question: Why not aim to upper bound probability of an α-violation?

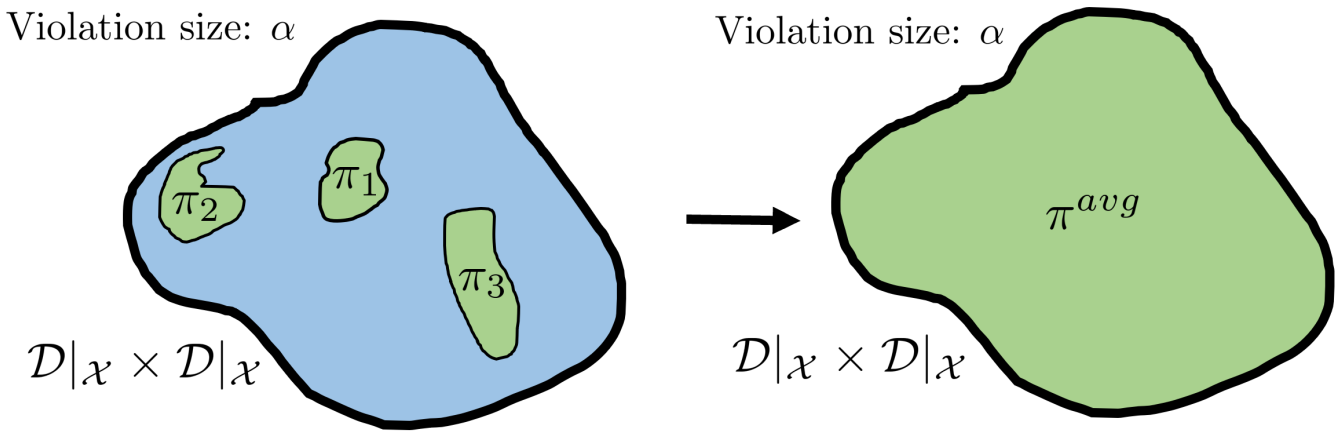
Bounding Probability of α-Violation

Observation

Suppose for all t, π^t is (α, β^t)-fair. Then, π^{avg} is $\left(\alpha, \sum_{t=1}^{T} \beta^t\right)$-fair.

Proof:

- Consider all policies deployed along the run: $\pi_1, \pi_2, \pi_3, \ldots, \pi_T$.
- Worst case: All violations are of magnitude 1, violations do not cancel out, all non-violations are arbitrarily close to α.



Dissatisfying, vacuous when $\sum_{t=1}^{T} \beta^t \geq 1$.

Interpolating α and β

1. Upper bound the probability of an $(\alpha + \frac{q}{T})$-fairness violation using $\sum_{t=1}^{T} \beta^t$.

Lemma

Assume that for all t, π^t is (α', β^t)-fair $(0 \leq \beta^t \leq 1)$. For any integer $q \leq T$, π^{avg} is $\left(\alpha' + \frac{q}{T}, \frac{1}{q}\sum_{t=1}^{T} \beta^t\right)$-fair.

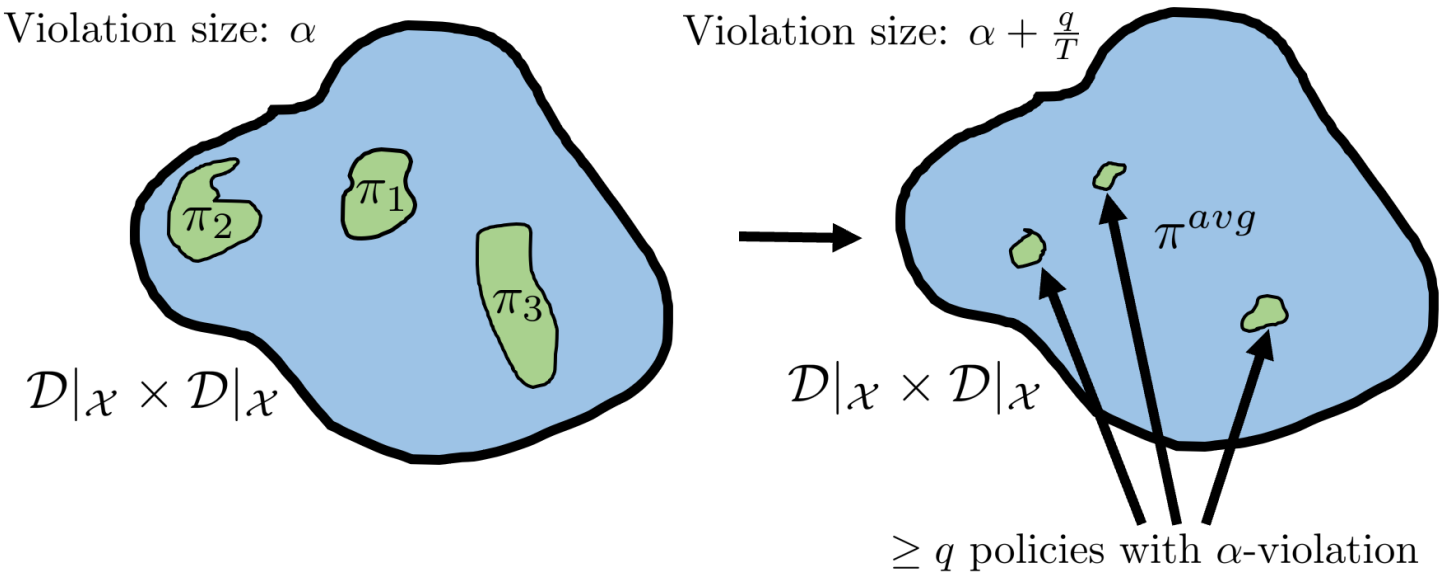
Interpolating α and β

Proof Sketch:

1. Considering $\alpha + \frac{q}{T}$, we have the following:

$$\forall x, x' : \pi^{avg} \text{ has } (\alpha + \frac{q}{T}) - \text{fairness violation on } x, x' \implies$$
$$\exists i_1, \ldots, i_q \in [T] \text{ s.t. } \forall j : |\pi_j(x) - \pi_j(x')| > d(x, x') + \alpha.$$

2. $\Pr_{x,x'}[\pi^{avg} \text{ has } (\alpha + \frac{q}{T}) - \text{fairness violation on } x, x'] \leq \frac{1}{q} \sum_{t=1}^{T} \beta^t.$



Linking $\sum_{t=1}^{T} \beta^t$ and Regret

2. Upper bound $\sum_{t=1}^{T} \beta^t$ using the regret guarantee of the algorithm.

Lemma

With probability $1-\delta$, *we have*

$$\sum_{t=1}^{T} \beta^t \leq \text{Regret}^{C,\alpha,\mathcal{J}_{\alpha+\epsilon}}(\mathcal{A}, Q_\alpha, T) + \sqrt{2T \ln\left(\frac{2}{\delta}\right)}$$

Accuracy + Fairness Generalization Guarantees

Corollary

Using CONTEXT-FTPL *from Syrgkanis et al. (2016) with a separator set of size s, with probability $1-\delta$, the average policy π^{avg} has the following guarantees:*

Accuracy:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\pi^{avg}(x),y)\right] \leq \inf_{\pi\in Q_\alpha} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(\pi(x),y)\right] + \ldots$$

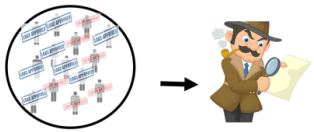
$$\ldots + O\left(\frac{1}{k^{\frac{1}{4}}}\left(\frac{s}{\epsilon}\right)^{\frac{3}{4}}\sqrt{\frac{\ln(|\mathcal{H}|)+\ln\left(\frac{1}{\delta}\right)}{T}}\right).$$

Fairness: *π^{avg} is $(\alpha'+\lambda,\lambda)$-fair where*

$$\lambda = O\left(\left(\frac{sk}{\epsilon}\right)^{\frac{3}{4}}\left(\frac{\ln\left(|\mathcal{H}|\right)+\ln\left(\frac{1}{\delta}\right)}{T}\right)^{\frac{1}{4}}\right).$$

Recap

Human Auditor For Fairness Violations



Predictions

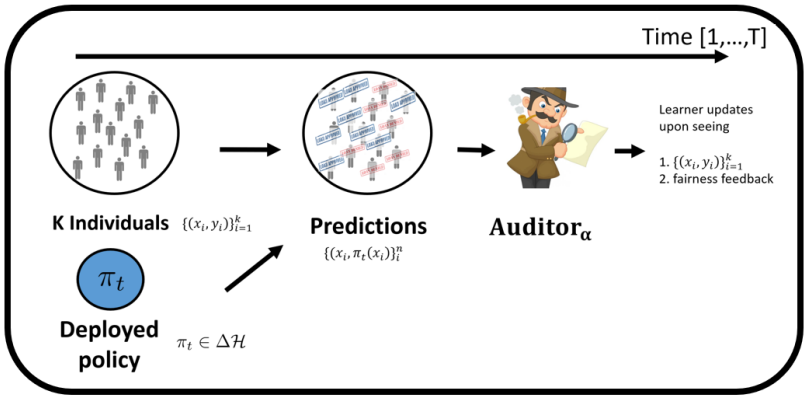
$\{(x_i, \pi_t(x_i)\}_i^n$

Auditor_α

Fairness Feedback

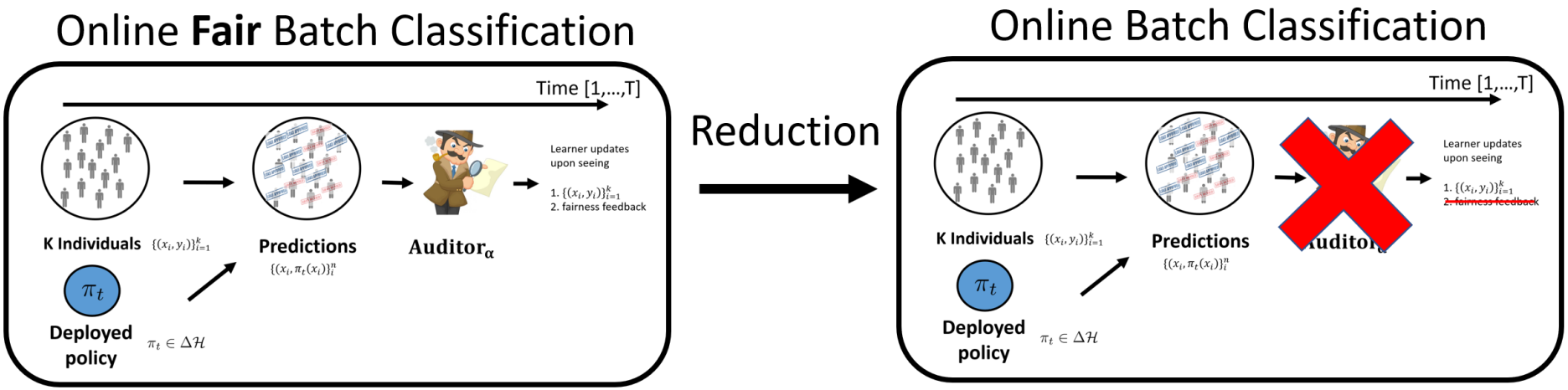
Recap

Online **Fair** Batch Classification



1. Individuals arrive **adversarially**.
2. Auditor reports a **single** violation.

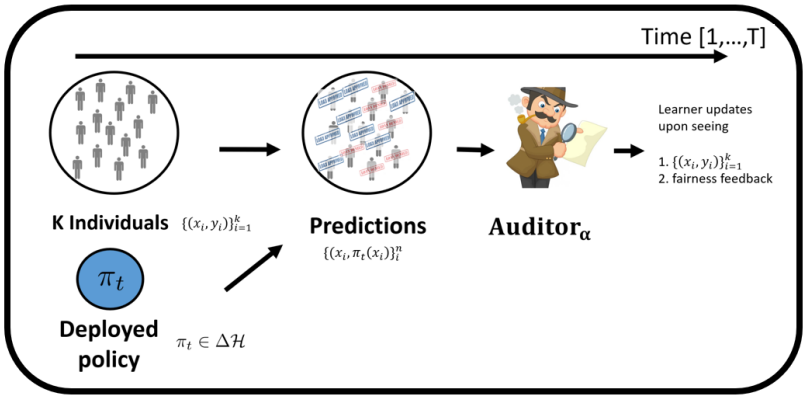
Recap



1. Individuals arrive **adversarially**.
2. Auditor reports a **single** violation.

Recap

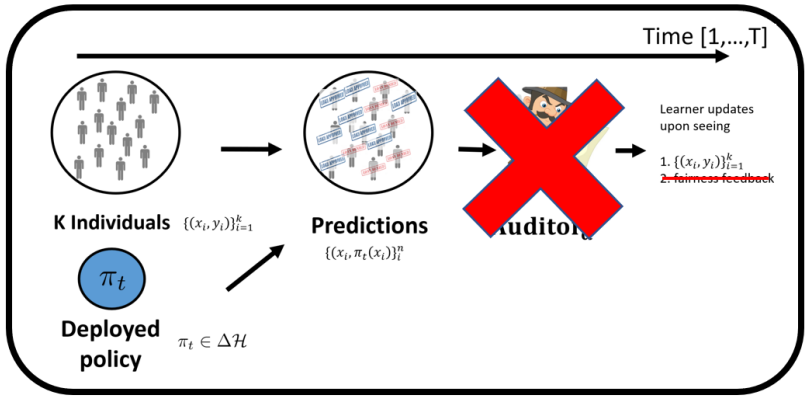
Online **Fair** Batch Classification



1. Individuals arrive **adversarially**.
2. Auditor reports a **single** violation.

Reduction

Online Batch Classification

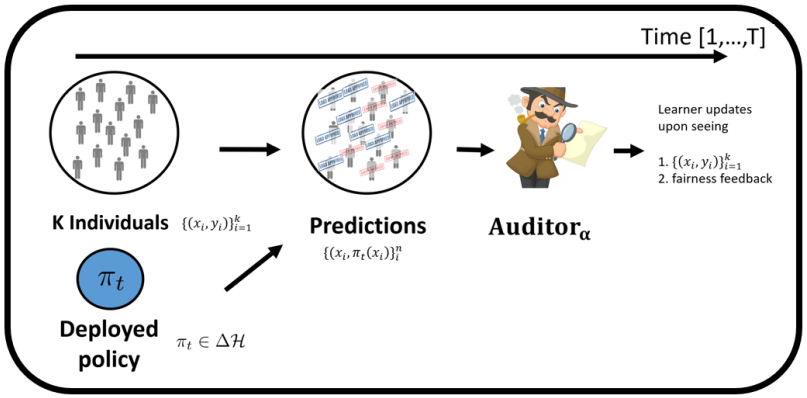


(Any) No-Regret Algorithm

No-Regret for both accuracy, fairness w.r.t. most accurate α-fair policy.

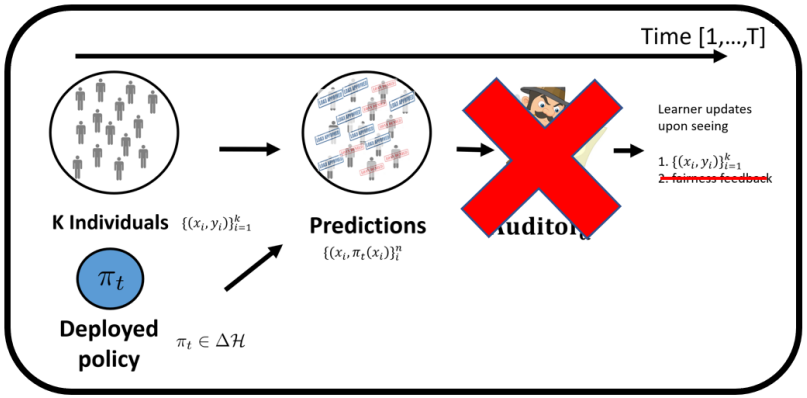
Recap

Online **Fair** Batch Classification



Reduction

Online Batch Classification



1. ~~Individuals arrive **adversarially**.~~
2. Auditor reports a **single** violation.

(Any) No-Regret Algorithm

No-Regret for both accuracy, fairness w.r.t. most accurate α-fair policy.

Distributional Assumption

1. Composition Argument.
2. Upper bound using regret.

Average policy is (W.H.P.):

1. **Accurate:**

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\pi^{avg}(x), y)] \leq \inf_{\pi\in Q_\alpha} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\pi(x), y)] + O(\tfrac{1}{\sqrt{T}}).$$

2. $(\alpha' + O(T^{-\frac{1}{4}}), O(T^{-\frac{1}{4}}))$-**Fair.**

Individual Vs. Group Fairness

Say I apply for a loan. What am I guaranteed?

Group Fairness

A group that contains me is treated similarly to a different group.

Individual Fairness

All individuals who are similar to me are treated the way I am treated.

Comparison with Prior Work on Individual Fairness

- Dwork et al., ITCS 2011: Conceptual introduction of Individual Fairness, relying on the availability of similarity metric.
- Rothblum and Yona, ICML 2018: Assume metric is given, provide generalization results for batch setting.
- Ilvento, FORC 2020: Attempts to learn the metric via distance and numerical comparison queries.
- Gillen et al., NeurIPS 2018: Assume specific structure of metric, linear bandit setting, auditor must report all violations.

Conclusions

- Statistical Fairness notions - Weak guarantees for Individuals.
- Individual Fairness - Was not operationalizable due to classical similarity metric assumption.
- We suggest a human auditor-based approach, which is:
  - **Easy to implement**: Instead of multiple numerical queries, only require auditor to report a single fairness violation.
  - **Metric-Free**: Auditor's judgements need not be consistent with any metric.
- We suggest an algorithm, which is:
  - **General**: No parametric assumptions on hypothesis class or similarity function.
  - **Efficient**: Can be implemented efficiently given access to optimization oracle.
- (Adversarial Arrivals) No-Regret guarantees for Accuracy, Fairness.
- (Stochastic Arrivals) Generalization for Accuracy, Fairness.
- **Actionable** notions of Algorithmic Fairness.