Individual Fairness in Online Classification

Yahav Bechavod

University of Pennsylvania

August 19, 2023

“Metric-Free Individual Fairness in Online Learning”
Joint with Christopher Jung and Steven Wu. NeurIPS 2020 Oral.

“Individually Fair Learning with One-Sided Feedback”
Joint with Aaron Roth. ICML 2023.

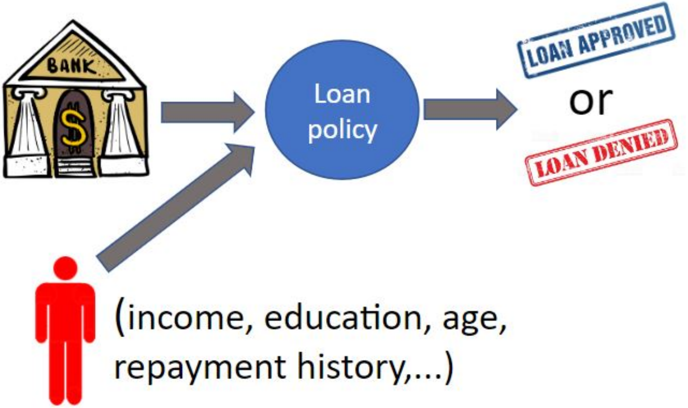
High-Level Plan

1. Re-examine commonly made assumptions regarding:
   - The level on which fairness is defined
   - The data generation process
   - The feedback model

Running Example

Example: **Loan Approvals**

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



Focus #1: Group Fairness Offers Weak Guarantees

The bulk of research in algorithmic fairness considers definitions that only bind on a **group level**.

Statistical fairness

- Select a statistic (accuracy, FPR/FNR, PPV,...).
- Define a set of groups in the population.
- (Approximately) equalize the statistic across groups.

Focus #1: Group Fairness Offers Weak Guarantees

- Advantage: relatively easy to work with.
- Disadvantage: very weak guarantees **for individuals**.

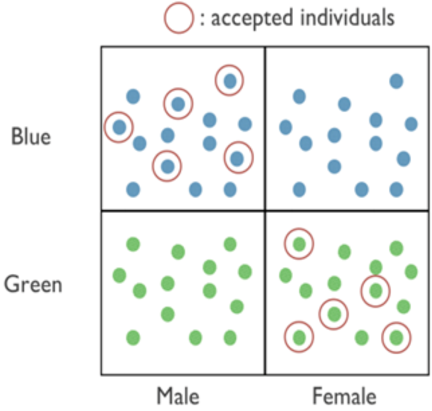


Figure: Fairness Gerrymandering: A Toy Example [Kearns et al., 2018]

Focus #2: Standard Statistical Assumptions May Not Always Apply

The majority of the work in algorithmic fairness operates under **statistical data generation assumptions**.

However: in various setting where fairness is a major concern, arriving individuals may not necessarily follow i.i.d. assumptions, due to, e.g.:

- Strategic effects (feature modifications based on knowledge/in anticipation of a specific policy, choosing whether to apply based on the policy in effect).

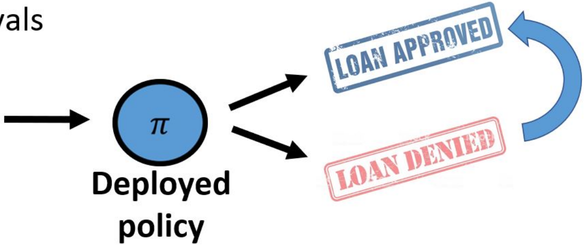
Learning in the Presence of Strategic Behavior

Individuals would like to receive **more favorable** assessments

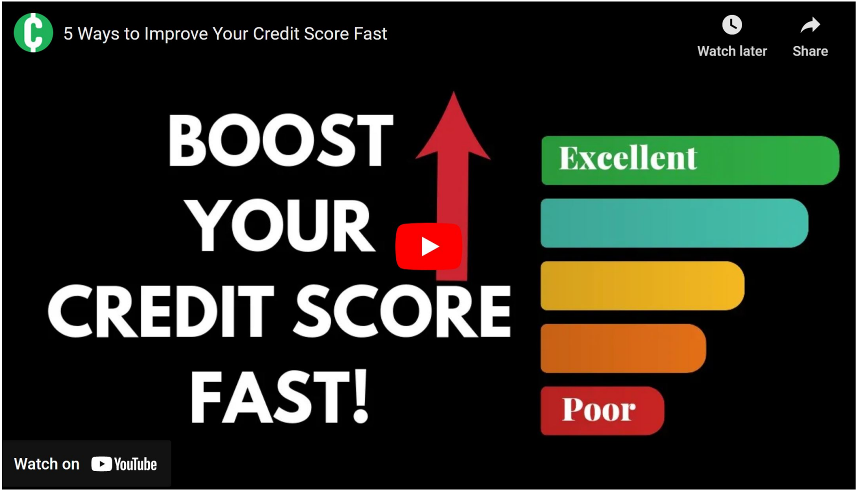
→ Act strategically

→ Strategic feature modifications

Example: loan approvals



Strategic Feature Modifications



Strategic Feature Modifications

Obtain additional credit cards
Raise your credit limits
...

Reduce your debt
Increase your income
...

OpenSCHUFA

- SCHUFA is Germany's leading credit bureau.
- SCHUFA has 943 million records on 67.7 million natural persons, and 6 million companies. Schufa processes more than 165 million credit checks each year. Of those, 2.5 million are self-checks by citizens. Schufa employs 900 people (as of 2019). In 2016 Sales amounted to approx. 190 million Euros.

OpenSCHUFA

"We were able to motivate more than 4,000 people to provide us with their SCHUFA information – very sensitive information that people usually keep to themselves."

Beyond Standard Statistical Assumptions

Arriving individuals may not necessarily follow i.i.d. assumptions:

- Strategic effects (feature modifications based on knowledge/in anticipation of a specific policy, choosing whether to apply based on the policy in effect).
- distribution shifts over time (e.g. ability to repay a loan may be affected by changes to the economy or recent events).
- Adaptivity to previous decisions (e.g. if an individuals receives a loan, that may affect the ability to repay future loans by this individual or his/her vicinity).

Focus #3: Feedback May Not Be Fully Observable

The bulk of the literature on algorithmic fairness operates in either:

- Batch setting
- Online setting with full information
- Bandit setting

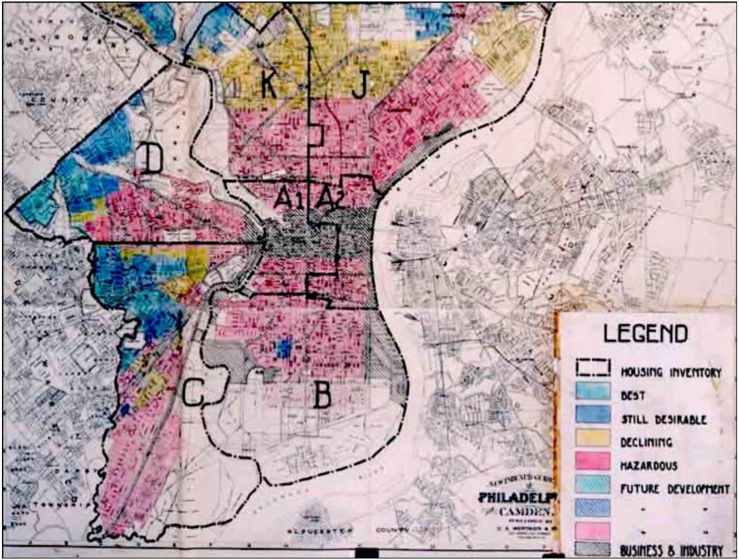
Focus #3: Feedback May Not Be Fully Observable

However, in many domains where fairness is a major concern, feedback may arrive for **positively predicted** individuals only. Cannot observe counterfactuals.

- Loan approvals
- College admissions
- Hiring for jobs
- Online advertising
- ...

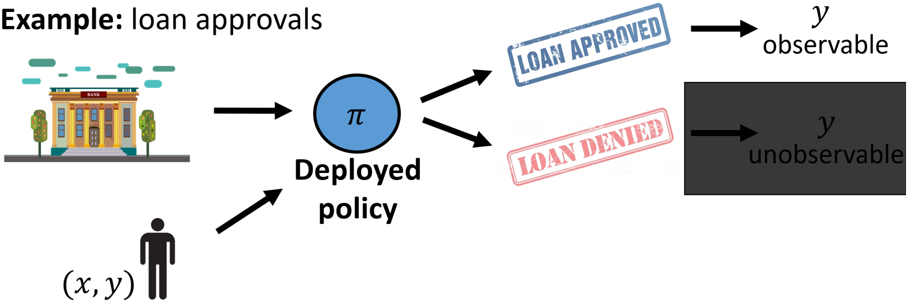
$\Longrightarrow$ Batch setting - data could be "skewed" to only include individuals accepted by past policy. In particular, if not careful, could inherit biases of historical discriminatory policies.

Redlining


K
J
D
A1
A2
C
B
PHILADELP
CAMDEN
LEGEND
HOUSING INVENTORY
BEST
STILL DESIRABLE
DECLINING
HAZARDOUS
FUTURE DEVELOPMENT
BUSINESS & INDUSTRY

One-Sided Feedback

Example: loan approvals



This is **not** a bandit setting!

High-Level Plan

1. Re-examine the assumptions commonly made regarding:
   - The level on which fairness is defined
   - The data generation process
   - The feedback model
2. Design efficient algorithms that:
   - Offer meaningful guarantees to individuals
   - Operate beyond standard statistical assumptions
   - Can handle limited feedback

Outline

- Fairness Framework: Metric-Free Individual Fairness via Panels
- Individually Fair Online Batch Classification
- Reduction to Contextual Combinatorial Semi-Bandit
- Multi-Criteria No Regret Guarantees for Accuracy, Fairness
- Oracle-Efficient Algorithm

Outline

- Fairness Framework: Metric-Free Individual Fairness via Panels
- Individually Fair Online Batch Classification
- Reduction to Contextual Combinatorial Semi-Bandit
- Multi-Criteria No Regret Guarantees for Accuracy, Fairness
- Oracle-Efficient Algorithm

Individual Fairness

Dwork et al. 2011: "Fairness Through Awareness"

"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: Access to similarity metric between individuals:

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

Challenges in Operationalizing Individual Fairness

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**.

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."

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

Prior Work on Individual Fairness

- Dwork, Hardt, Pitassi, Reingold, Zemel, 2011: Conceptual introduction of individual fairness, relying on the availability of a similarity metric.
- Rothblum and Yona 2018: Assume metric is given, provide generalization results for accuracy and fairness in batch setting.
- Ilvento 2020: Learning the metric via distance and numerical comparison queries to human arbiters.
- Kim, Reingold, Rothblum, 2018: Group-based relaxation of individual fairness, relying on access to an auditor returning unbiased estimates of distances between pairs of individuals
- Gillen, Jung, Kearns, Roth, 2018: Auditor "knows unfairness when he sees it". Assume specific parametric form of metric, auditor must report all violations on a given round.

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, $\ell(\hat{y}, y) = \mathbb{1}(\hat{y} \neq y)$.
- 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 $\mathcal{X}$.

Definition (α-fairness violation)

Let $\alpha \geq 0$ and let $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$. We say that a policy $\pi \in \Delta\mathcal{H}$ has an α-fairness violation (or simply "α-violation") on $(x, x') \in \mathcal{X}^2$ with respect to d if

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

where $\pi(x) = \Pr_{h\sim\pi}[h(x) = 1]$.

Auditor

- An auditor reports **one** α-violation if one or more exists.

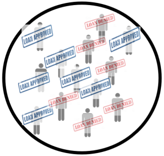
Definition (Auditor)

Let $\alpha \geq 0$. We define a fairness auditor $j^{\alpha} \in \mathcal{J}$ by, $\forall \pi \in \Delta\mathcal{H}, \bar{x} \in \mathcal{X}^k$,

$$j^{\alpha}(\pi, \bar{x}) := \begin{cases} (\bar{x}^s, \bar{x}^l) \in V^j & \text{if } V^j := \{(\bar{x}^s, \bar{x}^l) : s \neq l \in [k], \\ & \pi(\bar{x}^s) - \pi(\bar{x}^l) > d^j(x, x') + \alpha\} \neq \emptyset \ , \\ (v, v) & \text{otherwise} \end{cases}$$

where $\bar{x} = (\bar{x}^1, \dots, \bar{x}^k)$, $d^j : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is auditor j^{α}'s (implicit) distance function, and $v \in \mathcal{X}$ is some "default" context.

Auditor



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

(Features, Predictions)

Auditor$_\alpha$

Individuals 5 and 17 are being treated unfairly

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

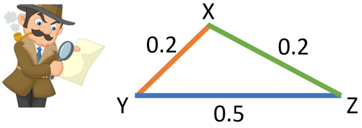
Or

I don't see any unfair treatments here.

Fairness Feedback

Metric-Free Individual Fairness

Q: Auditors' preferences may be inconsistent. What if the specified feedback from the auditor does not obey metric form?



- In our formulation, 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.
- Furthermore, we place **no parametric assumptions** on d.

How Should We Audit for Unfairness?

So far: single auditor, no metric assumption

However: unlikely that stakeholders would rely on a single auditor regarding fairness related judgements, especially in high-stakes domains:

- Human auditors may have implicit biases based on many factors: background, socio-economic level, education level, etc.
- A static auditing scheme may risk leaving too much power in the hands of the same (few) individuals over time.
- Practically speaking, may be infeasible for the same auditor to examine more than a certain amount of cases in a specific period of time.

Our Approach: Dynamic Auditing by Panels

We propose an auditing scheme based on dynamically-selected panels of multiple auditors.

Example:

- Ethicists familiar with the history of redlining
- Financial experts
- ...

Handling Inconsistent Judgements

Q: In case judgments of different auditors are inconsistent with each other, how should we handle disagreements?

Definition ((α, γ)-fairness violation)

Let $\alpha \geq 0$, $0 \leq \gamma \leq 1$, $m \in \mathbb{N} \setminus \{0\}$. We say that a policy $\pi \in \Delta\mathcal{H}$ has an (α, γ)-fairness violation on $(x, x') \in \mathcal{X}^2$ with respect to $d^1, \ldots, d^m : \mathcal{X}^2 \to [0, 1]$ if

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left[\pi(x) - \pi(x') - d^i(x, x') > \alpha\right] \geq \gamma.$$

Auditing by Panels

Definition (Panel)

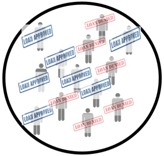
Let $\alpha \geq 0$, $0 \leq \gamma \leq 1$, $m \in \mathbb{N} \setminus \{0\}$. We define a fairness panel $\bar{j}^{\alpha,\gamma}$ by, $\forall \pi \in \Delta\mathcal{H}, \bar{x} \in \mathcal{X}^k$,

$$\bar{j}^{\alpha,\gamma}_{j^1,\ldots,j^m}(\pi,\bar{x}) = \begin{cases} (\bar{x}^s,\bar{x}^l) \in V^{\bar{j}} & \text{if } V^{\bar{j}} := \{(\bar{x}^s,\bar{x}^l) : s \neq l \in [k] \wedge \exists i_1,\ldots,i_{\lceil \gamma m \rceil} \in [m] \\ & \forall s \in [\lceil \gamma m \rceil], (\bar{x}^s,\bar{x}^l) \in V^{j^{i_s}}\} \neq \emptyset \\ (v,v) & \text{otherwise} \end{cases}$$

where $\bar{x} := (\bar{x}^1,\ldots,\bar{x}^k)$, $d^j : \mathcal{X} \times \mathcal{X} \to [0,1]$ is auditor j's (implicit) distance function, and $v \in \mathcal{X}$ is some "default" context. $\bar{j}$.

- Can vary γ and **algorithmically** explore the trade-off.

Auditing by Panels



$\{(x_i, \pi(x_i))\}_i^k$

(Features, Predictions)

m auditors

Panel

$j^{\alpha,\gamma}(x_1, \dots, x_k, \pi)$

Violation threshold

Required consensus

There exist two individuals (e.g. 5 and 17) for which at least γ fraction of the auditors see a violation: $|\pi(x_5) - \pi(x_{17})| > d^i(x_5, x_{17}) + \alpha$

Or

No unfair treatments.

Fairness Feedback

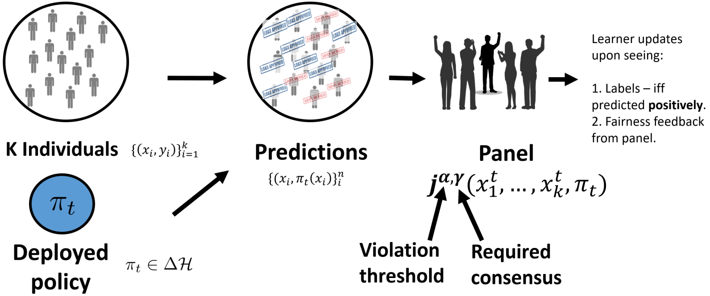
Outline

- Fairness Framework: Metric-Free Individual Fairness via Panels
- Individually Fair Online Batch Classification
- Reduction to Contextual Combinatorial Semi-Bandit
- Multi-Criteria No Regret Guarantees for Accuracy, Fairness
- Oracle-Efficient Algorithm

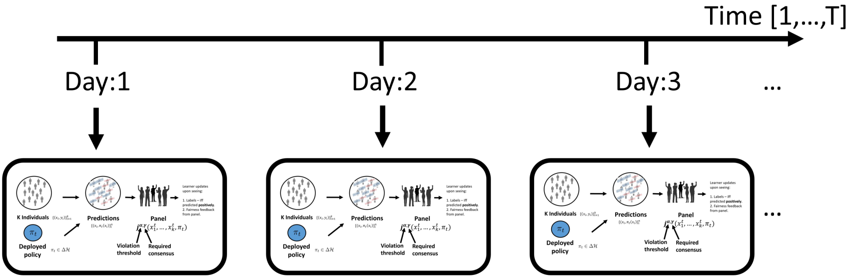
Our Setting

- Online classification
- Arriving individuals:
  - Possibly adversarial
  - Possibly multiple arrivals each round
  - Label information for positive predictions only
- Auditing panels:
  - Dynamically selected

Individually fair online batch classification: single round



Our Setting


Time [1,...,T]
Day:1
Day:2
Day:3
...
K Individuals
Predictions
Panel
Deployed policy
Violation threshold
Required consensus
Learner updates upon seeing:
1. Labels – iff predicted positively.
2. Fairness feedback from panel.
...

Individually fair online batch classification with one-sided feedback

Algorithm 1: Individually fair online batch classification with one-sided feedback

Input: Number of rounds T, hypothesis class $\mathcal{H}$;
Learner initializes $\pi^1 \in \Delta\mathcal{H}$;
for $t = 1, \ldots, T$ **do**

Environment selects individuals $\bar{x}^t \in \mathcal{X}^k$, and labels $\bar{y}^t \in \mathcal{Y}^k$, learner only observes $\bar{x}^t$;

Environment selects panel of auditors $(j^{t,1}, \ldots, j^{t,m}) \in \mathcal{J}^m$;

Learner draws $h^t \sim \pi^t$, predicts $\hat{y}^{t,i} = h^t(\bar{x}^{t,i})$ for each $i \in [k]$, observes $\bar{y}^{t,i}$ iff $\hat{y}^{t,i} = 1$;

Panel reports its feedback $\rho^t = \bar{j}^{t,\alpha,\gamma}_{j^1,\ldots,j^m}(\pi^t, \bar{x}^t)$;

Learner suffers misclassification loss $Error(h^t, \bar{x}^t, \bar{y}^t)$ (not necessarily observed by learner);

Learner suffers unfairness loss $Unfair(\pi^t, \bar{x}^t, \bar{j}^t)$;

Learner updates $\pi^{t+1} \in \Delta\mathcal{H}$;

end

Online Fair Batch Classification

Definition (Misclassification loss)

We define the misclassification loss as, for all $\pi \in \Delta\mathcal{H}$, $\bar{x} \in \mathcal{X}^k$, $\bar{y} \in \{0,1\}^k$ as:

$$Error(\pi, \bar{x}, \bar{y}) := \mathop{\mathbb{E}}_{h\sim\pi}[\ell^{0-1}(h, \bar{x}, \bar{y})].$$

Where for all $h \in \mathcal{H}$, $\ell^{0-1}(h, \bar{x}, \bar{y}) := \sum_{i=1}^{k} \ell^{0-1}(h, (\bar{x}^i, \bar{y}^i))$, and $\forall i \in [k] : \ell^{0-1}(h, (\bar{x}^i, \bar{y}^i)) = \mathbb{1}[h(\bar{x}^i) \neq \bar{y}^i]$.

Definition (Unfairness loss)

Let $\alpha \geq 0$, $0 \leq \gamma \leq 1$. We define the unfairness loss as, for all $\pi \in \Delta\mathcal{H}$, $\bar{x} \in \mathcal{X}^k$, $\bar{j} = \bar{j}^{\alpha,\gamma}_{j^1,\dots,j^m} : \mathcal{X}^k \to \mathcal{X}^2$,

$$Unfair^{\alpha,\gamma}(\pi, \bar{x}, \bar{j}) := \begin{cases} 1 & \bar{j}(\pi, \bar{x}) = (\bar{x}^s, \bar{x}^l) \wedge s \neq l \\ 0 & \text{otherwise} \end{cases},$$

where $\bar{x} := (\bar{x}^1, \dots, \bar{x}^k)$.

Lagrangian Loss

Definition (Lagrangian loss)

Let $C > 0$, $\rho = (\rho^1, \rho^2) \in \mathcal{X}^2$. We define the (C, ρ)-Lagrangian loss as, for all $\pi \in \Delta\mathcal{H}$, $\bar{x} \in \mathcal{X}^k$, $\bar{y} \in \{0, 1\}^k$,

$$L_{C,\rho}(\pi, \bar{x}, \bar{y}) := Error(\pi, \bar{x}, \bar{y}) + C \cdot \left[\pi(\rho^1) - \pi(\rho^2)\right].$$

Linear in $\Delta\mathcal{H}$.

Regret

Definition (Error regret)

We define the error regret of an algorithm $\mathcal{A}$ against a comparator class $U \subseteq \Delta\mathcal{H}$ to be

$$Regret^{err}(\mathcal{A}, T, U) = \sum_{t=1}^{T} Error(\pi^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in U} \sum_{t=1}^{T} Error(\pi^*, \bar{x}^t, \bar{y}^t).$$

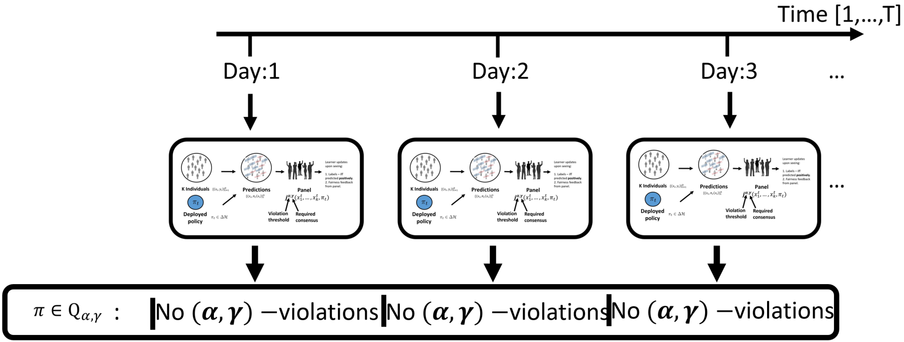
Definition (Unfairness regret)

Let $\alpha \geq 0$, $0 \leq \gamma \leq 1$. We define the unfairness regret of an algorithm $\mathcal{A}$ against a comparator class $U \subseteq \Delta\mathcal{H}$ to be

$$Regret^{unfair,\alpha,\gamma}(\mathcal{A}, T, U) = \sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\pi^t, \bar{x}^t, \bar{j}^t) - \min_{\pi^* \in U} \sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\pi^*, \bar{x}^t, \bar{j}^t).$$

Measuring Performance

“Competing” against most accurate policy that does not violate individual fairness.



Measuring Performance

We wish to compare performance with the highest-performing policy that is individually fair.

Definition ((α, γ)-fair policies)

Let $\alpha \geq 0$, $0 \leq \gamma \leq 1$, $m \in \mathbb{N} \setminus \{0\}$. We denote the set of all (α, γ)-fair policies with respect to all of the rounds in the run of the algorithm as

$$Q_{\alpha,\gamma} := \left\{ \pi \in \Delta\mathcal{H} : \forall t \in [T],\ \bar{j}^{t,\alpha,\gamma}_{j^{t,1},\dots,j^{t,m}}(\pi, \bar{x}^t) = (v, v) \right\}.$$

- Class is only defined **in hindsight** - realization is over both arriving individuals and panel members.

Simultaneous No-Regret Guarantees

We want, **simultaneously**:

1. **Accuracy:**
$$Regret^{err}(\mathcal{A}, T, Q_{\alpha,\gamma}) = o(T).$$
2. **Fairness:**
$$Regret^{unfair,\alpha,\gamma}(\mathcal{A}, T, Q_{\alpha,\gamma}) = o(T).$$

We know:
Gillen, Jung, Kearns, Roth (2018) - If auditor's judgements are according to a metric, of **particular parametric form** (Mahalanobis), and reports **all violations** - this is possible (with fast, logarithmic rate for the fairness regret).

Q: Can we still achieve simultaneous sub-linear rates under:

- no parametric or metric assumptions?
- auditor not reporting **all** violations?

Solution Strategy

1. Construct a reduction from our setting to the contextual combinatorial semi-bandit problem.
2. Show that, under certain conditions, the Lagrangian loss may be used to upper bound both error and unfairness losses.
3. Propose an oracle efficient algorithm by adapting Context-Semi-Bandit-FTPL (Syrgkanis et al. 2016), which would allow invoking our reduction.

Outline

- Fairness Framework: Metric-Free Individual Fairness via Panels
- Individually Fair Online Batch Classification
- Reduction to Contextual Combinatorial Semi-Bandit
- Multi-Criteria No Regret Guarantees for Accuracy, Fairness
- Oracle-Efficient Algorithm

Contextual Combinatorial Semi-Bandit

Algorithm 2: Contextual Combinatorial Semi-Bandit

Parameters: Class of predictors $\mathcal{H}$, number of rounds T;
Learner deploys $\pi^1 \in \Delta\mathcal{H}$;
for $t = 1, \ldots, T$ **do**
- Environment selects loss vector $\ell^t \in [0,1]^k$ (without revealing it to learner);
- Environment selects contexts $\bar{x}^t \in \mathcal{X}^k$, and reveals them to the learner;
- Learner draws action $a^t \in A^t \subseteq \{0,1\}^k$ according to π^t (where $A^t = \{a^t_h = (h(\bar{x}^{t,1}), \ldots, h(\bar{x}^{t,k})) : \forall h \in \mathcal{H}\})$;
- Learner suffers linear loss $\langle a^t, \ell^t \rangle$;
- Learner observes $\ell^{t,i}$ iff $a^{t,1} = 1$;
- Learner deploys π^{t+1};

end

Reduction

In describing the reduction, we use the following notations (For integers $k \geq 2$, $C \geq 1$):

$$(i)\ \forall a \in \{\rho^{t,1}, \rho^{t,2}, 0, 1, 1/2\} : \quad \bar{a} := \overbrace{(a, \ldots, a)}^{\text{C times}}, \quad \bar{\bar{a}} := \overbrace{(a, \ldots, a)}^{\text{k+2C times}}.$$

$$(ii)\ h(\bar{\bar{x}}^t) := (h(\bar{\bar{x}}^{t,1}), \ldots, h(\bar{\bar{x}}^{t,2k+4C})).$$

Algorithm 3: Reduction to Contextual Combinatorial Semi-Bandit

---

Input: Contexts $\bar{x}^t \in \mathcal{X}^k$, labels $\bar{y}^t \in \{0,1\}^k$, hypothesis h^t, pair $\rho^t \in \mathcal{X}^2$, parameter $C \in \mathbb{N}$;

Define: $\bar{\bar{x}}^t = (\bar{x}^t, \bar{\rho}^{t,1}, \bar{\rho}^{t,2}) \in \mathcal{X}^{k+2C}, \bar{\bar{y}}^t = (\bar{y}^t, \bar{0}, \bar{1}) \in \{0,1\}^{k+2C}$;

Construct loss vector: $\ell^t = (\bar{\bar{1}} - \bar{\bar{y}}^t, \bar{\bar{1/2}}) \in [0,1]^{2k+4C}$;

Construct action vector: $a^t = (h^t(\bar{\bar{x}}^t), \bar{\bar{1}} - h^t(\bar{\bar{x}}^t)) \in \{0,1\}^{2k+4C}$;

Output: (ℓ^t, a^t);

Reduction

In describing the reduction, we use the following notations (For integers $k \geq 2$, $C \geq 1$):

$$(i)\ \forall a \in \{\rho^{t,1}, \rho^{t,2}, 0, 1, 1/2\} : \quad \bar{a} := \overbrace{(a, \ldots, a)}^{\text{C times}}, \quad \bar{\bar{a}} := \overbrace{(a, \ldots, a)}^{\text{k+2C times}}.$$

$$(ii)\ h(\bar{\bar{x}}^t) := (h(\bar{\bar{x}}^{t,1}), \ldots, h(\bar{\bar{x}}^{t,2k+4C})).$$

Algorithm 4: Reduction to Contextual Combinatorial Semi-Bandit

---

Input: Contexts $\bar{x}^t \in \mathcal{X}^k$, labels $\bar{y}^t \in \{0,1\}^k$, hypothesis h^t, pair $\rho^t \in \mathcal{X}^2$, parameter $C \in \mathbb{N}$;

Define: $\bar{\bar{x}}^t = (\bar{x}^t, \bar{\rho}^{t,1}, \bar{\rho}^{t,2}) \in \mathcal{X}^{k+2C}, \bar{\bar{y}}^t = (\bar{y}^t, \bar{0}, \bar{1}) \in \{0,1\}^{k+2C}$;

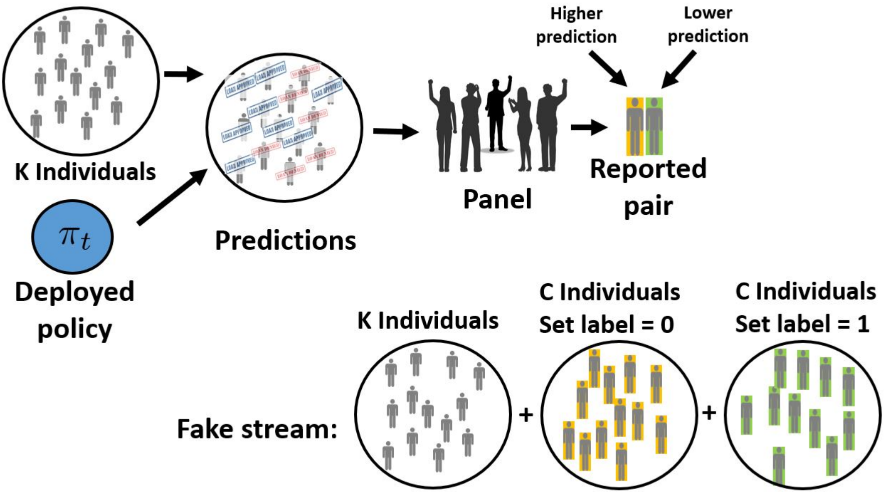
Construct loss vector: $\ell^t = (\bar{\bar{1}} - \bar{\bar{y}}^t, \bar{\bar{1/2}}) \in [0,1]^{2k+4C}$;

Construct action vector: $a^t = (h^t(\bar{\bar{x}}^t), \bar{\bar{1}} - h^t(\bar{\bar{x}}^t)) \in \{0,1\}^{2k+4C}$;

Output: (ℓ^t, a^t);

Reduction

1. Encoding unfairness loss in terms of misclassification loss, by generating a "fake" stream of samples.



Reduction

In describing the reduction, we use the following notations (For integers $k \geq 2$, $C \geq 1$):

$$(i)\ \forall a \in \{\rho^{t,1}, \rho^{t,2}, 0, 1, 1/2\}: \quad \bar{a} := \overbrace{(a, \ldots, a)}^{\text{C times}}, \quad \bar{\bar{a}} := \overbrace{(a, \ldots, a)}^{\text{k+2C times}}.$$

$$(ii)\ h(\bar{\bar{x}}^t) := (h(\bar{\bar{x}}^{t,1}), \ldots, h(\bar{\bar{x}}^{t,2k+4C})).$$

Algorithm 5: Reduction to Contextual Combinatorial Semi-Bandit

Input: Contexts $\bar{x}^t \in \mathcal{X}^k$, labels $\bar{y}^t \in \{0,1\}^k$, hypothesis h^t, pair $\rho^t \in \mathcal{X}^2$, parameter $C \in \mathbb{N}$;

Define: $\bar{\bar{x}}^t = (\bar{x}^t, \bar{\rho}^{t,1}, \bar{\rho}^{t,2}) \in \mathcal{X}^{k+2C}, \bar{\bar{y}}^t = (\bar{y}^t, \bar{0}, \bar{1}) \in \{0,1\}^{k+2C}$;

Construct loss vector: $\ell^t = (\bar{\bar{1}} - \bar{\bar{y}}^t, \bar{\bar{1/2}}) \in [0,1]^{2k+4C}$;

Construct action vector: $a^t = (h^t(\bar{\bar{x}}^t), \bar{\bar{1}} - h^t(\bar{\bar{x}}^t)) \in \{0,1\}^{2k+4C}$;

Output: (ℓ^t, a^t);

2. Handling one-sided feedback: misclassification loss manipulation:

$$\ell = \begin{array}{c c} & \begin{array}{cc} \textit{Good} & \textit{Bad} \end{array} \\ \begin{array}{c} \textit{Accept} \\ \textit{Reject} \end{array} & \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array} \rightarrow \tilde{\ell} = \begin{array}{c c} & \begin{array}{cc} \textit{Good} & \textit{Bad} \end{array} \\ \begin{array}{c} \textit{Accept} \\ \textit{Reject} \end{array} & \begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix} \end{array}$$

Manipulation is **regret-preserving**:

$$\forall h \in \mathcal{H} : \tilde{\ell}(h,(x,y)) = \ell(h,(x,y)) + \mathbb{1}[y=0]$$
$$\implies \forall h, h' \in \mathcal{H} : \tilde{\ell}(h,(x,y)) - \tilde{\ell}(h',(x,y)) = \ell(h,(x,y)) - \ell(h',(x,y))$$

Allows for moving from one-sided to bandit setting.

Upper Bounding Lagrangian Regret

For the following theorem, we will assume the existence of an algorithm $\mathcal{A}$ for the contextual combinatorial semi-bandit setting (as summarized in Algorithm 2) whose expected regret (compared to only fixed hypotheses in $\mathcal{H}$), against any adaptively and adversarially chosen sequence of loss functions ℓ^t and contexts $\bar{x}^t$, is bounded by $Regret(\mathcal{A}, T, \mathcal{H}) \leq R^{\mathcal{A},T,\mathcal{H}}$.

Theorem (Upper Bounding Lagrangian Regret)

In the setting of individually fair online learning with one-sided feedback (Algorithm 1), running $\mathcal{A}$ while using the sequence $(a^t, \ell^t)_{t=1}^T$ generated by the reduction in Algorithm 5 (when invoked every round on $\bar{x}^t$, $\bar{y}^t$, h^t, ρ^t, and C), yields the following guarantee, for any $V \subseteq \Delta\mathcal{H}$,

$$\sum_{t=1}^{T} L_{C,\rho^t}(\pi^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in V} \sum_{t=1}^{T} L_{C,\rho^t}(\pi^*, \bar{x}^t, \bar{y}^t) \leq (2k + 4C)R^{\mathcal{A},T,\mathcal{H}}.$$

Simultaneous No-Regret Guarantees

Reminder: we want, **simultaneously**:

1. **Accuracy:**
$$Regret^{err}(\mathcal{A}, T, Q_{\alpha,\gamma}) = o(T).$$
2. **Fairness:**
$$Regret^{unfair,\alpha,\gamma}(\mathcal{A}, T, Q_{\alpha,\gamma}) = o(T).$$

Upper Bounding Misclassification, Unfairness

Theorem (Upper Bounding Misclassification, Unfairness Simultaneously)

For any $\epsilon \in [0, \alpha]$,

$$C\epsilon \sum_{t=1}^{T} \mathit{Unfair}^{\alpha,\gamma}(\pi^t, \bar{x}^t, \bar{j}^t) + \mathit{Regret}^{err}(\mathcal{A}, T, Q_{\alpha-\epsilon,\gamma})$$
$$\leq \sum_{t=1}^{T} L_{C,\rho^t}(\pi^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in Q_{\alpha-\epsilon,\gamma}} \sum_{t=1}^{T} L_{C,\rho^t}(\pi^*, \bar{x}^t, \bar{y}^t).$$

And remember that the right hand side is upper bounded by $(2k + 4C)R^{\mathcal{A},T,\mathcal{H}}$.

Careful...

Theorem (Upper Bounding Misclassification, Unfairness Simultaneously)

For any $\epsilon \in [0, \alpha]$,

$$C\epsilon \sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\pi^t, \bar{x}^t, \bar{j}^t) + Regret^{err}(\mathcal{A}, T, Q_{\alpha-\epsilon,\gamma})$$
$$\leq \sum_{t=1}^{T} L_{C,\rho^t}(\pi^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in Q_{\alpha-\epsilon,\gamma}} \sum_{t=1}^{T} L_{C,\rho^t}(\pi^*, \bar{x}^t, \bar{y}^t).$$

$Regret^{err}(\mathcal{A}, T, Q_{\alpha-\epsilon,\gamma})$ can be **negative**!

$\implies$ Even if Lagrangian regret is sublinear, number of fairness violations can still be **linear**.

$\implies$ We will need to carefully interpolate between the two objectives.

Outline

- Fairness Framework: Metric-Free Individual Fairness via Panels
- Individually Fair Online Batch Classification
- Reduction to Contextual Combinatorial Semi-Bandit
- Multi-Criteria No Regret Guarantees for Accuracy, Fairness
- Oracle-Efficient Algorithm

So Far

- (Any) no regret algorithm for contextual combinatorial semi-bandit $\implies$ simultaneous no regret for each of accuracy, fairness.
- Important: our reduction requires that the panel **sees the predictions** (not the realization!) of the deployed policy on incoming individuals:
  - Fine with exponential weights style algorithms.
  - FTPL style algorithms **do not** explicitly maintain the distribution deployed over base predictors every round.

Multi-Criteria No-Regret Guarantees: Exp2 ("Expanded Exp")

Exp2 (Bubeck et al. 2012) is an adaptation of the classical exponential weights algorithm for linear bandits.

- in order to cope with the semi-bandit nature of the online setting, leverages the linear structure of the loss functions in order to share information regarding the observed feedback between all experts (hypotheses in $\mathcal{H}$).
- Such information sharing is then utilized in decreasing the variance in the formed loss estimators, resulting in a regret rate that depends only logarithmically (instead of linearly) on $|\mathcal{H}|$.

Multi-Criteria No-Regret Guarantees: Exp2 ("Expanded Exp")

Theorem

In the setting of individually fair online learning with one-sided feedback (Algorithm 1), running Exp2 for contextual combinatorial semi-bandits (Algorithm 2) while using the sequence $(a^t, \ell^t)_{t=1}^{T}$ generated by the reduction in Algorithm 5 (when invoked each round using $\bar{x}^t$, $\bar{y}^t$, h^t, ρ^t, and $C = T^{\frac{1}{5}}$), yields the following guarantees, for any $\epsilon \in [0, \alpha]$, simultaneously:

1. **Accuracy:** $Regret^{err}(Exp2, T, Q_{\alpha-\epsilon,\gamma}) \leq O\left(k^{\frac{3}{2}} T^{\frac{4}{5}} \log |\mathcal{H}|^{\frac{1}{2}}\right)$.
2. **Fairness:** $\sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\pi^t, \bar{x}^t, \bar{j}^t) \leq O\left(\frac{1}{\epsilon} k^{\frac{3}{2}} T^{\frac{4}{5}} \log |\mathcal{H}|^{\frac{1}{2}}\right)$.

However, Exp2 has space and time requirements linear in T. Could be prohibitive for large classes.

Multi-Criteria Guarantees: Context-Semi-Bandit-FTPL

Context-Semi-Bandit-FTPL (Syrgkanis et al. 2016) is an oracle-efficient algorithm for combinatorial bandits. It requires access to:

- (Offline) optimization oracle.
- Pre-computed (small) separator set.

However, in our specific setting, it cannot simply be applied off the shelf.

Multi-Criteria Guarantees: Adapting Context-Semi-Bandit-FTPL

In order to not have runtime, memory complexity that scales with $|\mathcal{H}|$, Context-Semi-Bandit-FTPL **does not** explicitly maintain the deployed distribution over $\mathcal{H}$.

- Instead, it samples a single hypothesis according to this distribution every round, utilizing the linearity of the loss function.
- However, for individual fairness this is problematic, as it can lead to extreme overestimation of unfairness, if panel is queried using single hypotheses. This is since the unfairness loss is **sub-additive**.

Lemma

There exist $\alpha, \gamma, m, k > 0$, $\mathcal{H} : \mathcal{X} \rightarrow \{0,1\}$, $\bar{x} \in \mathcal{X}^k$, $\bar{j} : \mathcal{X}^k \rightarrow \mathcal{X}^2$, and $\pi \in \Delta\mathcal{H}$ for which, simultaneously,

1. $\mathop{\mathbb{E}}_{h\sim\pi}\left[unfair^{\alpha,\gamma}(h, \bar{x}, \bar{j})\right] = 1.$
2. $unfair^{\alpha,\gamma}(\pi, \bar{x}, \bar{j}) = 0.$

Adapting Context-Semi-Bandit-FTPL

- **Potential solution:** Closed-form expression for the (implicit) weights the algorithm places on each $h \in \mathcal{H}$.
- However, the weights are generally not efficiently computable in closed form (see e.g. the discussion in Neu and Bartok 2013).
- **Our solution:** Instead, we will resample the deployed hypothesis every round.
- **Problem:** In order to use adversarial online learning algorithms, the realized randomness of the learner cannot be revealed to the adversary before it picks its loss vector.
- In general: adversary can tailor the losses to the realized randomness and force linear regret.

Adapting Context-Semi-Bandit-FTPL

Algorithm 4: Utilization of Context-Semi-Bandit-FTPL

Parameters: Class of predictors $\mathcal{H}$, number of rounds T, separator set S, parameters ω, L;
Initialize Context-Semi-Bandit-FTPL-With-Resampling(S, ω, L);
Learner deploys $\pi^1 \in \Delta\mathcal{H}$ according to Context-Semi-Bandit-FTPL-With-Resampling;
for $t = 1, \ldots, T$ **do**
  Environment selects individuals $\bar{x}^t \in \mathcal{X}^k$, and labels $\bar{y}^t \in \mathcal{Y}^k$, learner only observes $\bar{x}^t$;
  Environment selects panel of auditors $(j^{t,1}, \ldots, j^{t,m}) \in \mathcal{J}^m$;
  $(\hat{\pi}^t, \hat{h}^t) =$ Context-Semi-Bandit-FTPL-With-Resampling($\bar{x}^t, \omega, L$);
  Learner predicts $\hat{y}^{t,i} = h^t(\bar{x}^{t,i})$ for each $i \in [k]$, observes $\bar{y}^{t,i}$ iff $\hat{y}^{t,i} = 1$;
  Panel reports its feedback $\rho^t = \bar{j}^{t,\alpha,\gamma}_{j^1,\ldots,j^m}(\hat{\pi}^t, \bar{x}^t)$;
  $(\ell^t, a^t) =$ Reduction($\bar{x}^t, \bar{y}^t, \hat{h}^t, \rho^t, C$);
  Update Context-Semi-Bandit-FTPL-With-Resampling with (ℓ^t, a^t);
  Learner suffers misclassification loss $Error(\hat{h}^t, \bar{x}^t, \bar{y}^t)$ (not necessarily observed by learner);
  Learner suffers unfairness loss $Unfair(\hat{\pi}^t, \bar{x}^t, \bar{j}^t)$;
  Learner deploys $\pi^{t+1} \in \Delta\mathcal{H}$ according to Context-Semi-Bandit-FTPL-With-Resampling;
end

Adapting Context-Semi-Bandit-FTPL

- **Potential solution:** Closed-form expression for the (implicit) weights the algorithm places on each $h \in \mathcal{H}$.
- The weights are generally not efficiently computable in closed form (see e.g. the discussion in Neu and Bartok 2013).
- **Our solution:** Instead, we will resample the deployed hypothesis every round.
- **Problem:** In order to use adversarial online learning algorithms, the realized randomness of the learner cannot be revealed to the adversary before it picks its loss vector.
- In general: adversary can "tailor" its losses to the realized randomness and force linear regret.
- However, since our "adversary" is **restricted** to act according to the (fixed) implicit distance functions of the auditors in the panel, it cannot really adversarially adapt to the realized estimate: with high probability, the fairness loss for the realized (estimated) policy and the underlying distribution is close.

Oracle-Efficient Algorithm: Context-Semi-Bandit-FTPL-With-Resampling

Theorem

In the setting of individually fair online learning with one-sided feedback (Algorithm 1), running Context-Semi-Bandit-FTPL-With-Resampling for contextual combinatorial semi-bandit (Algorithm 5) as specified in Algorithm 4, with $R = T$, and using the sequence $(\ell^t, a^t)_{t=1}^{T}$ generated by the reduction in Algorithm 5 (when invoked on each round using $\bar{x}^t$, $\bar{y}^t$, $\hat{h}^t$, $\hat{\rho}^t$, and $C = T^{\frac{4}{45}}$), yields, with probability $1-\delta$, the following guarantees, for any $\epsilon \in [0, \alpha]$, simultaneously:

1. **Accuracy:** $\mathit{Regret}^{err}(\mathit{CSB\text{-}FTPL\text{-}WR}, T, Q_{\alpha-\epsilon,\gamma}) \leq \tilde{O}\left(k^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{41}{45}} \log|\mathcal{H}|^{\frac{1}{2}}\right)$.
2. **Fairness:** $\sum_{t=1}^{T} \mathit{Unfair}^{\alpha,\gamma}(\hat{\pi}^t, \bar{x}^t, \bar{j}^t) \leq \tilde{O}\left(\frac{1}{\epsilon} k^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{41}{45}} \log|\mathcal{H}|^{\frac{1}{2}}\right)$.

Overview of Results

Full Information	Inefficient	**Accuracy:** $\tilde{O}\left(kT^{\frac{3}{4}}\right)$ **Fairness:** $\tilde{O}\left(\frac{1}{\alpha}kT^{\frac{3}{4}}\right)$
	Efficient	**Accuracy:** $\tilde{O}\left(s^{\frac{3}{4}}k^{\frac{5}{4}}T^{\frac{7}{9}}\right)$ **Fairness:** $\tilde{O}\left(\frac{1}{\alpha}s^{\frac{3}{4}}k^{\frac{5}{4}}T^{\frac{7}{9}}\right)$
One-Sided	Inefficient	**Accuracy:** $\tilde{O}\left(k^{\frac{3}{2}}T^{\frac{4}{5}}\right)$ **Fairness:** $\tilde{O}\left(\frac{1}{\alpha}k^{\frac{3}{2}}T^{\frac{4}{5}}\right)$
	Efficient	**Accuracy:** $\tilde{O}\left(s^{\frac{3}{4}}k^{\frac{11}{4}}T^{\frac{41}{45}}\right)$ **Fairness:** $\tilde{O}\left(\frac{1}{\alpha}s^{\frac{3}{4}}k^{\frac{11}{4}}T^{\frac{41}{45}}\right)$

Limitations

- Exp2 prohibitive for large hypothesis classes.
- Context-Semi-Bandit-FTPL-WR:
  - Small separator sets only known for specific classes (conjunctions, disjunctions, parities, decision lists, discretized linear classifiers).
  - Our implementation requires $O(T^2)$ calls to the (offline) optimization oracle.

We "inherit" some of the limitations from the contextual bandit literature.

Rich Subgroup Fairness

- Kearns et al. 2018, Hébert-Johnson et al. 2018. Many follow up works.
- A "middleground" between group and individual fairness - equalizing across a pre-defined set of (potentially) exponentially many, possibly overlapping, groups in the population.
- Allows for significantly stronger guarantees for individuals than simple group notions.

Individual Fairness and Rich Subgroup Fairness

- Individual fairness sits on one extreme of subgroup fairness, treating each individual as a subgroup.
- However, individual fairness does not equalize some statistic over all individuals, but rather according to a very specific structure - given by an extra component, specifying who is similar.
- Individual fairness gives direct influence to people's preferences in forming the fairness definition.
- However, harder to elicit. Could trigger larger tension with accuracy if similarity preferences are not well-aligned with labels.

Takeaways

- Meaningful fairness guarantees to individuals, while minimizing surrounding assumptions, regarding:
  - The availability or form of similarity metrics
  - Data generation process
  - The observable feedback for made decisions
- Fairness auditing framework which can handle multiple auditors with (possibly) conflicting opinions
  - Possible to **algorithmically** change the required consensus for a fairness violation and explore the frontier.
- Possible to achieve simultaneous no regret for accuracy and individual fairness, under
  - No parametric (or even metric) assumptions on similarity judgements
  - Adversarial arrivals
  - One-sided label feedback

Future Directions

- Is it possible to achieve faster rates? The regret lower bound for combinatorial bandits is $\Omega(k\sqrt{T \log |\mathcal{H}|})$.
- Can we give an oracle efficient algorithm in the general case (without requiring small separators)?
- Relaxing some of the assumptions:
  - What if only contexts are adversarial, but labels are selected from a distribution given the context?
  - What if panels are selected stochastically?
  - Parametric assumptions?
- Faster algorithms?