Individually Fair Learning with One-Sided Feedback

Yahav Bechavod, Aaron Roth
yahav.bechavod@cs.huji.ac.il, aaroth@cis.upenn.edu

**Individual Fairness**

“Similar individuals should be treated similarly.”

Meaningful guarantee at the individual level.

**Problem:** Metric often unavailable.

**Auditor-based Approach**

“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.”

**Issue #1:** single auditors are prone to biases.

- Decision-makers less likely to entrust a single auditor with fairness-related judgements in high-stakes scenarios.
- How to reconcile cases disagreed upon by different auditors?

**Auditing by Panels**

- Fairness violation – only when a consensus is reached within a panel.
- Possible to alter the required fraction to algorithmically explore the fairness-accuracy frontier.

**One-Sided Feedback**

**Issue #2:** real-life feedback is often one-sided.

- “Hidden outcomes” of rejected individuals.
- Uncareful treatment may result in feedback loops.

Example: loan approvals

- Deployed policy

\[ (x, y) \]

- \( y \) observable

- \( y \) unobservable

**Our Setting**

Online Learning with One-Sided Feedback + Feedback from Dynamically-Chosen Panels

Time \([1,...,T]\)

Learner updates upon seeing:

1. Labels – iff predicted positively.
2. Fairness feedback from panel.

**Results**

Result #1: Reduction from online learning with one-sided feedback and feedback from dynamically-chosen panels to Contextual Combinatorial Semi-Bandit.

Result #2: Multi-Criteria No-Regret Guarantees

Using regret bound of any algorithm for Contextual Combinatorial Semi-Bandit, upper bounding, simultaneously:

1. **Accuracy:** sub-linear regret vs. best fair policy.
2. **Fairness:** sub-linear number of rounds on which fairness violations exist.

**Accuracy + Fairness Guarantees**

**Thm. 1 (simplified.):** Using Exp2 algorithm,

**Accuracy:**

\[
\text{Regret} (\text{Exp}2, T, Q_{\alpha-\epsilon}) \leq O\left(\frac{3}{2} T \frac{3}{4} T \frac{3}{4} \log |H| \frac{1}{2}\right)
\]

**Fairness:**

\[
\sum_{t=1}^{T} \text{Unfair}^{\alpha, \gamma} (\pi_t, \bar{x}^t, \bar{y}^t) \leq O\left(\frac{1}{\epsilon} \frac{1}{2} \frac{3}{2} T \log |H| \frac{1}{2}\right)
\]

**Thm. 2 (simplified.):** Using (adapted) Context-Semi-Bandit-FTPL,

**Accuracy:**

\[
\text{Regret} (\text{CSB} - \text{FTPL} - WR, T, Q_{\alpha-\epsilon}) \leq \tilde{O}\left(\frac{11}{2} T \frac{3}{4} T \frac{3}{4} T \frac{3}{4} \log |H| \frac{1}{2}\right)
\]

**Fairness:**

\[
\sum_{t=1}^{T} \text{Unfair}^{\alpha, \gamma} (\pi_t, \bar{x}^t, \bar{y}^t) \leq \tilde{O}\left(\frac{1}{\epsilon} \frac{1}{2} \frac{11}{2} T \frac{3}{4} T \log |H| \frac{1}{2}\right)
\]