Individually Fair Learning with One-Sided Feedback

Yahav Bechavod [1] **Aaron Roth** [2]

Abstract

We consider an online learning problem with one-sided feedback, in which the learner is able to observe the true label only for positively predicted instances. On each round, k instances arrive and receive classification outcomes according to a randomized policy deployed by the learner, whose goal is to maximize accuracy while deploying *individually fair* policies. We first extend the framework of (Bechavod et al., 2020), which relies on the existence of a human fairness auditor for detecting fairness violations, to instead incorporate feedback from dynamically-selected panels of multiple, possibly inconsistent, auditors. We then construct an efficient reduction from our problem of online learning with one-sided feedback and a panel reporting fairness violations to the contextual combinatorial semi-bandit problem ((Cesa-Bianchi and Lugosi, 2009; György et al., 2007)). Finally, we show how to leverage the guarantees of two algorithms in the contextual combinatorial semi-bandit setting: Exp2 (Bubeck et al., 2012) and the oracle-efficient Context-Semi-Bandit-FTPL (Syrgkanis et al., 2016), to provide multi-criteria no regret guarantees simultaneously for accuracy and fairness. Our results eliminate two potential sources of bias from prior work: the "hidden outcomes" that are not available to an algorithm operating in the full information setting, and human biases that might be present in any single human auditor, but can be mitigated by selecting a well chosen panel.

[1]School of Computer Science and Engineering, The Hebrew University. [2]Department of Computer and Information Sciences, University of Pennsylvania.. Correspondence to: Yahav Bechavod <yahav.bechavod@cs.huji.ac.il>, Aaron Roth <aaroth@cis.upenn.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

When making many high stakes decisions about people, we receive only one-sided feedback—often we are only able to observe the outcome for people for whom we make a favorable decision. For example, we only observe the repayment history for applicants we approve for a loan—not for those we deny. We only observe the success or lack thereof for employees we hire, not for those that we pass on. We only observe the college GPA for those applicants that we admit to college, not to those we reject—and so on. In all of these domains, fairness is a major concern in addition to accuracy. Nevertheless, the majority of the literature on fairness in machine learning does not account for this "one-sided" feedback structure, operating either in a batch setting, a full information online setting, or in a more standard bandit learning setting. But when we make sequential decisions with one-sided feedback, it is crucial to explicitely account for the form of the feedback structure to avoid feedback loops that may amplify and disguise historical bias.

The bulk of the literature in algorithmic fairness also asks for fairness on a *group* or aggregate level. A standard template for this approach is to select some statistical measure of error (like false positive rate, false negative rates, or raw error rates), a partition of the data into groups (often along the lines of some "protected attribute"), and then to ask that the statistical measure of error is approximately equalized across the groups. Because these guarantees bind only over averages over many people, they promise little to individuals, as initially pointed out by Dwork et al.'s "catalogue of evils" (Dwork et al., 2012).

In an attempt to provide meaningful guarantees on an individual level, Dwork et al. (2012) introduced the notion of Individual fairness, which informally asks that "similar individuals be treated similarly". In their conception, this is a Lipschitz constraint imposed on a randomized classifier, and who is "similar" is defined by a task-specific similarity metric. Pinning down such a metric is the major challenge with using the framework of individual fairness. Gillen et al. (2018) proposed that feedback could be elicited in an online learning setting from a human auditor who "knows unfairness when she sees it" (and implicitly makes judgements according to a similarity metric), but cannot enunciate a metric — she can only identify specific violations of the fairness

constraint. Recently, Bechavod et al. (2020) gave an algorithm for operating in this setting—with full information—that was competitive with the optimal fair model, while being able to learn not to violate the notion of individual fairness underlying the feedback of a single auditor.

Our work extends that of Gillen et al. (2018); Bechavod et al. (2020) in two key ways. First, we remove the assumption of a single, consistent auditor: we assume we are given an adaptively chosen *panel* of human auditors who may have different conceptions of individual fairness and may be making inconsistent judgements (we aim to be consistent with plurality judgements of such a panel). Second, we dispense with the need to operate in a full information setting, and give oracle efficient algorithms that require only one-sided feedback. We give simultaneous no-regret guarantees for both classification error and fairness violation, with respect to models that are individually fair in hindsight (i.e. given the realization of the panels of fairness auditors who define our conception of fairness). Together these improvements eliminate two potential sources of bias from prior work: the "hidden outcomes" that are not available to an algorithm operating in the full information setting, and human biases that might be present in any single human auditor, but can be mitigated by selecting a well chosen panel.

1.1. Roadmap of our Contributions

We define an online learning framework with one-sided label feedback and additional feedback from dynamically-chosen panels of auditors regarding fairness violations (we present our formal model in Section 2). We show that auditing by panels is in fact equivalent to auditing by specific, instance-dependent, single auditors (Appendix B), which is a useful technical step in our analysis. We then cast our learning problem as an optimization problem of a joint objective using a Lagrangian formulation (Section 2.2). We construct an efficient reduction to the contextual combinatorial semi-bandit setting (Cesa-Bianchi and Lugosi, 2009; György et al., 2007) (Section 3). We then show how to leverage the regret guarantees of two algorithms for the contextual combinatorial semi-bandit setting: Exp2 (Bubeck et al., 2012) and the oracle-efficient Context-Semi-Bandit-FTPL (Syrgkanis et al., 2016), to produce regret guarantees simultaneously for each of accuracy and fairness (Section 4), where our adaptation of Context-Semi-Bandit-FTPL appears in Appendix D.

2. Preliminaries

We start by specifying the notation we will use for our setting. We denote a feature space by $\mathcal{X}$ and a label space by $\mathcal{Y}$. Throughout this work, we focus on the case where $\mathcal{Y} = \{0, 1\}$. We denote by $\mathcal{H}$ a hypothesis class of binary predictors $h : \mathcal{X} \to \mathcal{Y}$. We assume that $\mathcal{H}$ contains a constant hypothesis. For the purpose of achieving better accuracy-fairness trade-offs, we allow the deployment of randomized policies over the base class $\mathcal{H}$, which we denote by $\Delta\mathcal{H}$. As we will see later, in the context of individual fairness, it will be crucial to be able to compete with the best predictor in $\Delta\mathcal{H}$, rather than simply in $\mathcal{H}$. We model auditors as observing k-tuples of examples (the people who are present at some round of the decision making process), as well as our randomized prediction rule, and will form objections by identifying a pair of examples for which they believe our treatment was "unfair" if any such pair exists. For an integer $k \geq 2$, we denote by $\mathcal{J} : \Delta\mathcal{H} \times \mathcal{X}^k \to \mathcal{X}^2$ the domain of possible auditors. Next, we formalize the notion of fairness we will aim to satisfy.

2.1. Individual Fairness and Auditing

Here we define the notion of individual fairness and auditing that we use, following Dwork et al. (2012); Gillen et al. (2018); Bechavod et al. (2020), and extending it to the notion of a panel of auditors.

Definition 2.1 (α-fairness violation)**.** *Let $\alpha \geq 0$ and let $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$.*[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 $\pi(x) = \Pr_{h\sim\pi}[h(x) = 1]$.

A fairness auditor, parameterized by a distance function d, given a policy π and a set of k individuals, will report any single pair of the k individuals on which π represents an α-violation if one exists.

Definition 2.2 (Auditor)**.** *Let $\alpha \geq 0$. We define a fairness auditor $j^\alpha \in \mathcal{J}$ by, $\forall \pi \in \Delta\mathcal{H}, \bar{x} \in \mathcal{X}^k$, $j^\alpha(\pi, \bar{x}) :=$*

$$\begin{cases} (\bar{x}^s, \bar{x}^l) \in V^j & \text{if } V^j := \{(\bar{x}^s, \bar{x}^l) : s \neq l \in [k], \\ & \pi(\bar{x}^s) - \pi(\bar{x}^l) > d^j(x, x') + \alpha\} \neq \emptyset \\ (\bar{x}^1, \bar{x}^1) & \text{otherwise} \end{cases},$$

where $\bar{x} = (\bar{x}^1, \ldots, \bar{x}^k)$, and $d^j : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is auditor j^α's (implicit) distance function. When clear from context, we will abuse notation, and use j to denote such an auditor.

Note that if there exist multiple pairs in $\bar{x}$ on which an α-violation exist, we only require the auditor to report one. In

[1] d represents a function specifying the auditor's judgement of the "similarity" between individuals in a specific context. We do not require that d be a metric: only that it be non-negative and symmetric. It is important that we make as few assumptions as possible when modeling human auditors, as in general, we cannot expect this form of feedback to take specific parametric form, or even be a metric.

the case in which the auditor does not consider there to be any fairness violations, we define its output to be a "default" value, $(\bar{x}^1, \bar{x}^1)$, to indicate that no violation was detected.

Thus far our formulation of fairness violations and auditors follows the formulation in Bechavod et al. (2020). In the following, we generalize the notion of fairness violations to panels of multiple fairness auditors which extends beyond the framework of Bechavod et al. (2020).

Definition 2.3 ((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 (or simply "(α,γ)-violation") on $(x,x') \in \mathcal{X}^2$ with respect to $d^1,\ldots,d^m : \mathcal{X}^2 \to [0,1]$ if*

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

Definition 2.3 specifies that a policy π has an (α,γ)-fairness violation on a pair of examples when a γ fraction of the auditors consider π to have an α-fairness violation on that pair. By varying γ, we can interpolate between considering there to be a violation when any *single* auditor determines that there is one at one extreme, to requiring unanimity amongst the auditors at the other extreme.

Definition 2.4 (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, $\forall \pi \in \Delta\mathcal{H}, \bar{x} \in \mathcal{X}^k$, $\bar{j}^{\alpha,\gamma}_{j^1,\ldots,j^m}(\pi,\bar{x}) :=$*

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

where $\bar{x} := (\bar{x}^1,\ldots,\bar{x}^k)$, and $d^j : \mathcal{X}\times\mathcal{X} \to [0,1]$ is auditor j's (implicit) distance function. When clear from context, we will abuse notation and simply denote such a panel by $\bar{j}$.

Again panels need only report a *single* (α,γ)-violation even if many exist. The rationale behind extending the auditing scheme to panels is that human auditors have their own implicit biases, and so there may be no single human auditor that a collection of stakeholders would agree to entrust with fairness judgements. It is much easier to agree on a representative panel of authorities. As already noted, the γ parameter allows us to adjust the degree to which we require consensus amongst panel members: we can interpolate all the way between requiring full unanimity on all judgements of unfairness (when $\gamma = 1$) to giving any single panel member effective "veto power" (when $\gamma \leq 1/m$).

Note that different values of γ for the panel do not change the auditing task for individual auditors: in all cases, each auditor is only asked to report α-violations according to their own judgement. Thus, using the same feedback from a panel of auditors, we can algorithmically vary γ to explore an entire frontier of fairness/accuracy tradeoffs. We refer the reader to Appendix B, where an equivalence between consensus-based auditing schemes and auditing by instance-specific single auditors is proven and further discussed.

2.2. Lagrangian Loss Formulation

Next, we define the two types of loss we will use in our setting.

Definition 2.5 (Misclassification loss)**.** *We define the misclassification loss as, for all $\pi \in \Delta\mathcal{H}$, $\bar{x} \in \mathcal{X}^k$, $\bar{y} \in \{0,1\}^k$,*

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

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 $\forall i \in [k] : \ell^{0-1}(h,(\bar{x}^i,\bar{y}^i)) = \mathbb{1}[h(\bar{x}^i) \neq \bar{y}^i]$.

Next, we define the unfairness loss, to reflect the existence of one or more fairness violations according to a panel's judgement.

Definition 2.6 (Unfairness loss)**.** *Let $\alpha \geq 0$, $0 \leq \gamma \leq 1$. We define the unfairness loss as, for all $\pi \in \Delta\mathcal{H}$, $\bar{x} \in \mathcal{X}^k$, $\bar{y} \in \{0,1\}^k$, $\bar{j} : \mathcal{X}^k \to \mathcal{X}^2$, Unfair$^{\alpha,\gamma}(\pi,\bar{x},\bar{y},\bar{j}) :=$*

$$\begin{cases} 1 & \pi \text{ has an } (\alpha,\gamma) - \text{violation on a pair } (x,x') \in \bar{x} \\ & \text{w.r.t. panel } \bar{j} \\ 0 & \text{otherwise} \end{cases}.$$

Finally, we define the Lagrangian loss.

Definition 2.7 (Lagrangian loss)**.** *Let $C > 0$, $\rho = (\rho^1,\rho^2) \in \mathcal{X}^2$. We define the (C,ρ)-Lagrangian loss as, for all $\pi \in \Delta\mathcal{H}$, $\bar{x} \in \mathcal{X}^k$, $\bar{y} \in \{0,1\}^k$,*

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

We are now ready to formally define our learning environment, which we do next.

2.3. Individually Fair Online Learning with One-Sided Feedback

In this section, we formally define our learning environment with one-sided feedback. Our setting is formally defined in Algorithm 1.

One-sided feedback Our one-sided feedback structure (classically known as "apple tasting") is different from the standard bandit setting. In the bandit setting, the feedback visible to the learner is the loss for the selected action in each round. In our setting, feedback may or may not be

Algorithm 1 Individually Fair Online Learning 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 $Unfair(\pi^t, \bar{x}^t, \bar{y}^t, \bar{j}^t)$;

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

end for

observable for a selected action: if we classify an individual as positive, we observe feedback for our action—and for the counterfactual action we could have taken (classifying them as negative). On the other hand, if we classify an individual as negative, we do not observe (but still suffer) our classification error.

To measure performance, we will ask for algorithms who are competitive with the best possible (fixed) policy in hindsight. This is captured using the notion of regret, which we define next for relevant loss functions.

Definition 2.8 (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 2.9 (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*

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

Finally, we define the Lagrangian regret, which will be useful in our analysis.

Definition 2.10 (Lagrangian regret)**.** *Let $C > 0$, and $(\rho^t)_{t=1}^T$ be a sequence s.t. $\forall t \in [T] : \rho^t \in \mathcal{X}^2$. We define the Lagrangian regret of an algorithm $\mathcal{A}$ against a comparator class $U \subseteq \Delta\mathcal{H}$ to be*

$$Regret^{L,C,\rho^1,\ldots,\rho^T}(\mathcal{A}, T, U) = \sum_{t=1}^{T} L_{C,\rho^t}(\pi^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in U} \sum_{t=1}^{T} L_{C,\rho^t}(\pi^*, \bar{x}^t, \bar{y}^t).$$

In order to construct an algorithm that achieves no regret simultaneously for accuracy and fairness, our approach will be to reduce the setting of individually fair learning with one-sided feedback (Algorithm 1) to the setting of contextual combinatorial semi-bandits, which we will see next.

3. Reduction to Contextual Combinatorial Semi-Bandit

In this section, we present our main result: a reduction from individually fair online learning with one-sided feedback (Algorithm 1) to the setting of (adversarial) contextual combinatorial semi-bandits, to be specified next. The proofs for this section appear in Appendix C. We begin by formally describing the contextual combinatorial semi-bandit setting.[2] The setting is formally defined in Algorithm 2.

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 for

We next define regret in the context of contextual combinatorial semi-bandit (Algorithm 2).

[2] The combinatorial (full) bandit problem formulation is due to Cesa-Bianchi and Lugosi (2009). We consider a contextual variant of the problem. Our setting operates within a relaxation of the feedback structure, known as "semi-bandit" (György et al., 2007).

Definition 3.1 (Regret). *In the setting of Algorithm 2, we define the regret of an algorithm $\mathcal{A}$ against a comparator class $U \subseteq \Delta\mathcal{H}$ to be*

$$Regret(\mathcal{A}, T, U) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{a^t \sim \pi^t} \left\langle a^t, \ell^t \right\rangle - \min_{\pi^* \in U} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{a^* \sim \pi^*} \left\langle a^t, \ell^t \right\rangle.$$

Reduction Our reduction is summarized in Algorithm 3.

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

$$(i)\ \forall a \in \{\rho^{t,1}, \rho^{t,2}, 0, 1, 1/2\}:$$
$$\bar{a} := \overbrace{(a, \ldots, a)}^{\text{C times}}, \quad \bar{\bar{a}} := \overbrace{(a, \ldots, a)}^{\text{k+2C times}}.$$
$$(ii)\ h(\bar{\bar{x}}^t) := (h(\bar{\bar{x}}^{t,1}), \ldots, h(\bar{\bar{x}}^{t,2k+4C})).$$

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);

We next prove that the reduction described in Algorithm 3 can be used to upper bound an algorithm's Lagrangian regret in the individually fair online learning with one-sided feedback setting. For the following theorem, we will assume the existence of an algorithm $\mathcal{A}$ for the contextual combinatorial semi-bandit setting 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 3.2. *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 3 (when invoked every round on $\bar{x}^t$, $\bar{y}^t$, h^t, ρ^t, and C), yields the following guarantee, for any $V \subseteq \Delta\mathcal{H}$,*

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

Note that the guarantee of Theorem 3.2 holds when competing with the set of all possibly randomized policies $\Delta\mathcal{H}$ over the base class, instead of only with respect to the best classifier in $\mathcal{H}$. As we will see next, the bound on the Lagrangian loss regret in Theorem 3.2 will be useful in simultaneously upper bounding each of error regret and unfairness regret.

Definition 3.3 ((α,γ)-fairness). *Let $\alpha \geq 0$, $0 \leq \gamma \leq 1$, $m \in \mathbb{N} \setminus \{0\}$. We denote the set of (α,γ)-fair policies with respect to all of the panels in the run of the algorithm as*

$$Q_{\alpha,\gamma} := \{\pi \in \Delta\mathcal{H} : \forall (x,x') \in \mathcal{X}^2, \forall t \in [T] : \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left[\pi(x) - \pi(x') - d^{t,i}(x,x') > \alpha\right] < \gamma\},$$

where $d^{t,i}$ is auditor $j^{t,i}$'s underlying distance function.

Next, we show how the Lagrangian regret guarantee established in Theorem 3.2 can be utilized to provide simultaneous guarantees for accuracy and fairness, when compared with the most accurate policy in $Q_{\alpha-\epsilon,\gamma}$. Note, in particular, that by setting $Q_{\alpha-\epsilon,\gamma}$ as the comparator set, we will be able to upper bound the number of rounds in which an (α,γ)-violation has occurred.

Lemma 3.4. *For any $\epsilon \in [0, \alpha]$,*

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

4. Multi-Criteria No Regret Guarantees

In this section, we present two algorithms for the contextual combinatorial semi-bandit setting (Algorithm 2), Exp2 (Bubeck et al., 2012) and the oracle-efficient Context-Semi-Bandit-FTPL (Syrgkanis et al., 2016), and show how they can be leveraged to produce simultaneous accuracy and fairness guarantees in the setting of individually fair online learning with one-sided feedback (Algorithm 1). The full constructions described next, as well as the proofs for this section, appear in Appendix D. In the following, we use the notation $\|\ell^t\|_* = \max_{a \in A^t} |\langle \ell^t, a \rangle|$, and use $\tilde{O}$ to hide logarithmic factors.

4.1. Exp2

Theorem 4.1 (via Bubeck et al. (2012)). *The expected regret of Exp2 in the contextual combinatorial semi-bandit setting, against any adaptively and adversarially chosen sequence of contexts and linear losses such that $\|\ell^t\|_* \leq 1$, is at most:*

$$Regret(T) \leq O\left(\sqrt{kT \log |\mathcal{H}|}\right).$$

Next, we show how, when leveraging our reduction as described in Section 3, Exp2 can be utilized to provide multi-criteria guarantees, for accuracy and fairness.

Theorem 4.2. *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 3 (when invoked on each round using $\bar{x}^t$, $\bar{y}^t$, h^t, ρ^t, and $C = T^{\frac{1}{5}}$), yields the following regret 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:

$$\sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\pi^t, \bar{x}^t, \bar{y}^t, \bar{j}^t) \leq O\left(\frac{1}{\epsilon} k^{\frac{3}{2}} T^{\frac{4}{5}} \log|\mathcal{H}|^{\frac{1}{2}}\right).$$

While presenting statistically optimal performance in terms of its dependence on the number of rounds and the cardinality of the hypothesis class, Exp2 is in general computationally inefficient, with runtime and space requirements that are linear in $|\mathcal{H}|$, which is prohibitive for large hypothesis classes. We hence next propose an oracle-efficient algorithm, based on a combinatorial semi-bandit variant of the classical Follow-The-Perturbed-Leader (FTPL) algorithm (Kalai and Vempala, 2005; Hannan, 1957).

4.2. Context-Semi-Bandit-FTPL

Context-Semi-Bandit-FTPL assumes access to two key components: an offline optimization oracle for the base class $\mathcal{H}$, and a small separator set for $\mathcal{H}$. The optimization oracle assumption is simply equivalent to access to a weighted ERM oracle for $\mathcal{H}$ (we elaborate on the adaptation in Appendix D). We next describe the small separator set assumption.

Definition 4.3 (Separator set)**.** *We say $S \subseteq \mathcal{X}$ is a separator set for a class $\mathcal{H} : \mathcal{X} \to \{0, 1\}$, if for any two distinct hypotheses $h, h' \in \mathcal{H}$, there exists $x \in S$ s.t. $h(x) \neq h'(x)$. We denote $s := |S|$ the size of the separator set.*

Remark 4.4. *Classes for which small separator sets are known include conjunctions, disjunctions, parities, decision lists, discretized linear classifiers. Please see an elaborate discussion in Syrgkanis et al. (2016); Neel et al. (2019).*

Theorem 4.5 (via Syrgkanis et al. (2016))**.** *The expected regret of Context-Semi-Bandit-FTPL in the contextual combinatorial semi-bandit setting, against any adaptively and adversarially chosen sequence of contexts and linear, non-negative losses such that $\|\ell^t\|_* \leq 1$, is at most:*

$$Regret(T) \leq O\left(k^{\frac{7}{4}} s^{\frac{3}{4}} T^{\frac{2}{3}} \log|\mathcal{H}|^{\frac{1}{2}}\right).$$

We note that Context-Semi-Bandit-FTPL does not, at any point, maintain its deployed distribution over the class $\mathcal{H}$ explicitly. Instead, on each round, it "samples" a hypothesis according to such (implicit) distribution — where the process of perturbing then solving described above can equivalently be seen as sampling a single hypothesis from such underlying distribution over $\mathcal{H}$.

Resampling-based adaptation For our purposes, however, we will have to adapt the implementation of Context-Semi-Bandit-FTPL so that the process of sampling the hypothesis at each round is repeated, and we are able to form an empirical estimate of the implicit distribution. This is required for two reasons: first, as we wish to compete with the best fair *policy* in $\Delta\mathcal{H}$, rather than only with the best fair classifier in $\mathcal{H}$ (We elaborate on this in Appendix D.2, and in Lemma D.4). Second, as it is observed in general (see, e.g. the discussion in Neu and Bartók (2013)), the specific weights this implicit distribution places on each of $h \in \mathcal{H}$ cannot be expressed in closed-form. In Appendix D, We construct an adaptation, Context-Semi-Bandit-FTPL-With-Resampling, which is based on resampling the hypothesis R times and deploying the empirical estimate $\hat{\pi}^t$ of the (implicit) distribution π^t. This adaptation is summarized in Appendix D, and yields the following guarantees.

Theorem 4.6. *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 3 (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 regret guarantees, for any $\epsilon \in [0, \alpha]$, simultaneously:*

Accuracy: $$Regret^{err}(\textit{CSB-FTPL-WR}, T, Q_{\alpha-\epsilon,\gamma}) \leq \tilde{O}\left(k^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{41}{45}} \log|\mathcal{H}|^{\frac{1}{2}}\right).$$

Fairness: $$\sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\hat{\pi}^t, \bar{x}^t, \bar{y}^t, \bar{j}^t) \leq \tilde{O}\left(\frac{1}{\epsilon} k^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{41}{45}} \log|\mathcal{H}|^{\frac{1}{2}}\right).$$

5. Limitations and Future Directions

We next discuss limitations and future directions. First, the Exp2 algorithm has runtime and space requirements that are linear in $|\mathcal{H}|$, which is prohibitive for large hypothesis classes. Context-Semi-Bandit-FTPL is oracle-efficient, but is limited only to classes for which small separator sets are known. We inherit these limitations from the contextual bandit literature — they hold even without the additionally encoded fairness constraints. Second, our adaptation of Context-Semi-Bandit-FTPL requires T additional oracle calls at each iteration, to estimate the implicit distribution by the learner. Taken together, these limitations suggest the

following important open question: are there efficient algorithms which can provide accuracy and fairness guarantees of the sort we give here using one-sided feedback with auditors, which are not restricted by the limitations above? This question is interesting also in less adversarial settings than we consider here. For example, do things become easier if the panel is selected i.i.d. from a distribution every round, rather than being chosen by an adversary?

References

Antos, A., Bartók, G., Pál, D., and Szepesvári, C. (2013). Toward a classification of finite partial-monitoring games. *Theor. Comput. Sci.*, 473:77–99.

Bechavod, Y., Jung, C., and Wu, Z. S. (2020). Metric-free individual fairness in online learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Bechavod, Y., Ligett, K., Roth, A., Waggoner, B., and Wu, Z. S. (2019). Equal opportunity in online classification with partial feedback. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8972–8982.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. (2012). Towards minimax policies for online linear optimization with bandit feedback. In Mannor, S., Srebro, N., and Williamson, R. C., editors, *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, volume 23 of *JMLR Proceedings*, pages 41.1–41.14. JMLR.org.

Cesa-Bianchi, N. and Lugosi, G. (2009). Combinatorial bandits. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*.

Cesa-Bianchi, N., Lugosi, G., and Stoltz, G. (2006). Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580.

Coston, A., Rambachan, A., and Chouldechova, A. (2021). Characterizing fairness over the set of good models under selective labels. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2144–2155. PMLR.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012). Fairness through awareness. In Goldwasser, S., editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM.

Gillen, S., Jung, C., Kearns, M. J., and Roth, A. (2018). Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2605–2614.

Gupta, S. and Kamble, V. (2019). Individual fairness in hindsight. In *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, pages 805–806.

György, A., Linder, T., Lugosi, G., and Ottucsák, G. (2007). The on-line shortest path problem under partial monitoring. *J. Mach. Learn. Res.*, 8:2369–2403.

Hannan, J. (1957). Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3(2):97–139.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Helmbold, D. P., Littlestone, N., and Long, P. M. (2000). Apple tasting. *Inf. Comput.*, 161(2):85–139.

Ilvento, C. (2020). Metric learning for individual fairness. In Roth, A., editor, *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020, Harvard University, Cambridge, MA, USA (virtual conference)*, volume 156 of *LIPIcs*, pages 2:1–2:11. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L., and Wu, Z. S. (2021). An algorithmic framework for fairness elicitation. In Ligett, K. and Gupta, S., editors, *2nd Symposium on Foundations of Responsible Computing, FORC 2021, June 9-11, 2021, Virtual Conference*, volume 192 of *LIPIcs*, pages 2:1–2:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Kalai, A. T. and Vempala, S. S. (2005). Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.*, 71(3):291–307.

Kim, M. P., Reingold, O., and Rothblum, G. N. (2018). Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 4847–4857.

Neel, S., Roth, A., and Wu, Z. S. (2019). How to use heuristics for differential privacy. In Zuckerman, D., editor, *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 72–93. IEEE Computer Society.

Neu, G. and Bartók, G. (2013). An efficient algorithm for learning with semi-bandit feedback. In Jain, S., Munos, R., Stephan, F., and Zeugmann, T., editors, *Algorithmic Learning Theory - 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings*, volume 8139 of *Lecture Notes in Computer Science*, pages 234–248. Springer.

Rothblum, G. N. and Yona, G. (2018). Probably approximately metric-fair learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5666–5674.

Syrgkanis, V., Krishnamurthy, A., and Schapire, R. E. (2016). Efficient algorithms for adversarial contextual learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2159–2168.

A. Additional Related Work

Our work is related to two strands of literature: learning with one-sided feedback, and individual fairness in machine learning. Despite the prevalence of the problem across a wide variety of domains, there has been relatively little work in the one-sided feedback model for binary classification that we consider. The problem of learning from positive-prediction-only feedback first appeared in Helmbold et al. (2000), under the name of "apple tasting". Subsequently, Cesa-Bianchi et al. (2006) studied a generalization of the one-sided feedback setting, in which the feedback at each round is a function of the combined choice of two players. Follow-up work by Antos et al. (2013) showed that it is possible to reduce the online one-sided feedback setting to the better studied contextual bandit problem. In the context of algorithmic fairness, Bechavod et al. (2019) considers a stochastic online setting with one-sided feedback, in which the aim is to learn a binary classifier while enforcing the statistical fairness condition of "equal opportunity" (Hardt et al., 2016). Coston et al. (2021) operate in a batch setting with potentially missing labels due to one-sided feedback in historical decisions, and attempt to impute missing labels using statistical techniques.

Dwork et al. (2012) introduced the notion of individual fairness. In their formulation, a similarity metric is explicitly given, and they ask that predictors satisfy a Lipschitz condition (with respect to this metric) that roughly translates into the condition that "similar individuals should have similar distributions over outcomes". Rothblum and Yona (2018) give a statistical treatment of individual fairness in a batch setting with examples drawn i.i.d. from some distribution, and prove PAC-style generalization bounds for both accuracy and individual fairness violations. Ilvento (2020) suggests learning the similarity metric from human arbiters, using a hybrid model of comparison queries and numerical distance queries. Kim et al. (2018) study a group-based relaxation of individual fairness, while relying on access to an auditor returning unbiased estimates of distances between pairs of individuals. Jung et al. (2021) consider a batch setting, with a fixed set of "stakeholders" which provide fairness feedback regarding pairs of individuals in a somewhat different model of fairness, and give oracle-efficient algorithms and generalization bounds. Gupta and Kamble (2019) study a time-dependent variant of individual fairness they term "individual fairness in hindsight".

The papers most related to ours are Gillen et al. (2018) and Bechavod et al. (2020). Gillen et al. (2018) introduces the idea of online learning with human auditor feedback as an approach to individual fairness, but give algorithms that are limited to a single auditor that makes decisions with respect to a restrictive parametric form of fairness metrics in the full information setting. Bechavod et al. (2020) generalize this to a much more permisive definition of a human auditor, but still operate in the full information setting and are limited to single human auditors. See Appendix A for additional related work.

B. From Panels to Single Auditors

In this section, we prove a reduction from auditing by panels to auditing by single auditors. In particular, we prove that the decisions of any panel are equivalent to the decisions of a single auditor from the panel. We note that when considering the space of auditors, it is not possible in general to fully order or compare the level of strictness of different auditors, as some may be stricter than others on different regions of the space of pairs from $\mathcal{X}^2$, and this order may be reversed when considering different regions. For illustration, consider the following example: let $\mathcal{X} = \{x^1, x^2, x^3\}, \mathcal{J} = \{j^1, j^2\}$ and assume that $d^{j^1}(x^1, x^2) > d^{j^2}(x^1, x^2)$, and $d^{j^1}(x^2, x^3) < d^{j^2}(x^2, x^3)$. In the context of this example, asking who is stricter or who is more lenient among the auditors, in an absolute sense, is undefined.

However, as we restrict the attention to a single pair (x, x'), such a task becomes feasible. Namely, in spite of the fact that we do not have access to auditors' underlying distance measures (we only observe feedback regarding violations), we know that there is an implicit ordering among the auditors' level of strictness with respect to that specific pair. The idea is to then utilize this (implicit) ordering to argue that a panel's judgements with respect to this pair are in fact equivalent to the judgements of a specific single auditor from the panel, which can be viewed as a "representative auditor", having the "swing-vote" among the panel. We formalize the argument in Lemma B.1.

Lemma B.1. *Let $(x, x') \in \mathcal{X}^2$, $(j^1, ..., j^m) \in \mathcal{J}^m$. Then, the following are equivalent, for all $\pi \in \Delta\mathcal{H}$:*

1. *π has an (α, γ)-violation on (x, x') with respect to panel $\bar{j}^{\alpha,\gamma}_{j^1,...,j^m}$.*
2. *π has an α-violation on (x, x') with respect to auditor j^s, where $s = s_{x,x'}(j^1, \ldots, j^m)$ is an index in $[m]$.*

Proof of Lemma B.1. Fix (x, x'). Then, we can define an ordering of $(j^1, ..., j^m)$ according to their (underlying) distances

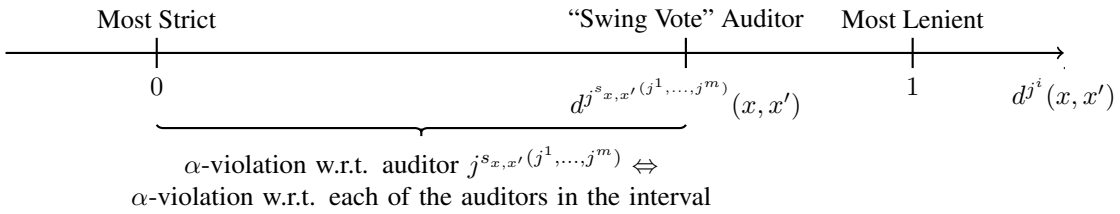

Figure 1. An illustration of an ordering of a panel of auditors $(j^1, \ldots, j^m)$ according to their (implicit) distances on (x, x'). $j^{s_{x,x'}(j^1,\ldots,j^m)}$ denotes the auditor who is in the $\lceil \gamma m \rceil$ position in this ordering, which can also be viewed as having the "swing vote" with respect to deciding an (α, γ)-violation in this instance.

on (x, x'),

$$d^{j^{i_1}}(x,x') \leq \cdots \leq d^{j^{i_m}}(x,x'). \tag{1}$$

Then, set

$$s := s_{x,x'}(j^1, ..., j^m) = \operatorname*{argmin}_{s'} \left\{ s' \in [m] : \frac{s'}{m} \geq \gamma \right\}. \tag{2}$$

Note that s in eq. (2) is well-defined, since $\gamma \leq 1$.

We also note that, using the ordering defined in eq. (1), for any $r \in [m]$,

$$\pi(x) - \pi(x') > d^{j^{i_r}}(x,x') + \alpha \implies \forall r' \leq r : \pi(x) - \pi(x') > d^{j^{i_{r'}}}(x,x') + \alpha. \tag{3}$$

Hence, when considering a random variable indicating an (α, γ)-violation, we know that

$$\begin{aligned}
&\mathbb{1}\left[\left[\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[\pi(x) - \pi(x') - d^{j^i}(x,x') > \alpha\right]\right] \geq \gamma\right] \\
&= \mathbb{1}\left[\left[\frac{1}{m}\sum_{i=1}^{s} \mathbb{1}\left[\pi(x) - \pi(x') - d^{j^s}(x,x') > \alpha\right]\right] \geq \gamma\right] && \text{(by eq. (2) and eq. (3))} \\
&= \mathbb{1}\left[\left[\mathbb{1}\left[\pi(x) - \pi(x') - d^{j^s}(x,x') > \alpha\right]\frac{s}{m}\right] \geq \gamma\right] && \text{(by eq. (1))} \\
&= \mathbb{1}\left[\pi(x) - \pi(x') - d^{j^s}(x,x') > \alpha\right] && \text{(by eq. (2)).}
\end{aligned}$$

Which concludes the proof. □

C. Omitted Details from Section 3

In what follows, we denote $k' = k + 2C$.

Lemma C.1. *For all $\pi, \pi' \in \Delta\mathcal{H}$, $\bar{x}^t \in \mathcal{X}^k$, $\bar{y}^t \in \{0,1\}^k$,*

$$L_{C,\rho^t}(\pi, \bar{x}^t, \bar{y}^t) - L_{C,\rho^t}(\pi', \bar{x}^t, \bar{y}^t) = \sum_{i=1}^{k'} Error(\pi, \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) - \sum_{i=1}^{k'} Error(\pi', \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}).$$

Proof. Observe that

$$
\begin{aligned}
&L_{C,\rho^t}(\pi, \bar{x}^t, \bar{y}^t) - L_{C,\rho^t}(\pi', \bar{x}^t, \bar{y}^t) \\
&= Error(\pi, \bar{x}^t, \bar{y}^t) + C \cdot \left[\pi(\rho^{t,1}) - \pi(\rho^{t,2})\right] - Error(\pi', \bar{x}^t, \bar{y}^t) - C \cdot \left[\pi'(\rho^{t,1}) - \pi'(\rho^{t,2})\right] \\
&= \sum_{i=1}^{k} Error(\pi, \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) - Error(\pi', \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) + \sum_{i=k+1}^{k+C} \pi(\rho^{t,1}) - \pi'(\rho^{t,1}) + \sum_{i=k+C+1}^{k+2C} 1 - \pi(\rho^{t,2}) - 1 + \pi'(\rho^{t,2}) \\
&= \sum_{i=1}^{k} Error(\pi, \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) - Error(\pi', \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) + \sum_{i=k+1}^{k+C} Error(\pi, \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) - Error(\pi', \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) \\
&+ \sum_{i=k+C+1}^{k+2C} Error(\pi, \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) - Error(\pi', \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) \\
&= \sum_{i=1}^{k'} Error(\pi, \bar{\bar{x}}^{t}, \bar{\bar{y}}^{t}) - \sum_{i=1}^{k'} Error(\pi', \bar{\bar{x}}^{t}, \bar{\bar{y}}^{t}).
\end{aligned}
$$

Which proves the lemma. □

Lemma C.2. *For all $\pi, \pi' \in \Delta\mathcal{H}$, $\bar{\bar{x}}^t \in \mathcal{X}^{k'}$,$\bar{\bar{y}}^t \in \mathcal{Y}^{k'}$,*

$$
\sum_{i=1}^{k'} Error(\pi, \bar{\bar{x}}^{t}, \bar{\bar{y}}^{t}) - \sum_{i=1}^{k'} Error(\pi', \bar{\bar{x}}^{t}, \bar{\bar{y}}^{t}) = 2\left[\mathop{\mathbb{E}}_{h\sim\pi}\left[\langle a^h, \ell^t\rangle\right] - \mathop{\mathbb{E}}_{h'\sim\pi'}\left[\langle a^{h'}, \ell^t\rangle\right]\right].
$$

Proof. Observe that

$$
\begin{aligned}
&\sum_{i=1}^{k'} Error(\pi, \bar{\bar{x}}^{t}, \bar{\bar{y}}^{t}) - \sum_{i=1}^{k'} Error(\pi', \bar{\bar{x}}^{t}, \bar{\bar{y}}^{t}) \\
&= \left[\sum_{i=1}^{k'} Error(\pi, \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) + \mathbb{1}\left[\bar{\bar{y}}^{t,i} = 0\right]\right] - \left[\sum_{i=1}^{k'} Error(\pi', \bar{\bar{x}}^{t,i}, \bar{\bar{y}}^{t,i}) + \mathbb{1}\left[\bar{\bar{y}}^{t,i} = 0\right]\right] \\
&= 2\Bigg[\left\langle \left(\pi(\bar{\bar{x}}^{t,1}), \ldots, \pi(\bar{\bar{x}}^{t,k'}), 1-\pi(\bar{\bar{x}}^{t,1}), \ldots, 1-\pi(\bar{\bar{x}}^{t,k'})\right), \left(1-\bar{\bar{y}}^{t,1}, \ldots, 1-\bar{\bar{y}}^{t,k'}, 1/2, \ldots, 1/2\right)\right\rangle \\
&- \left\langle \left(\pi'(\bar{\bar{x}}^{t,1}), \ldots, \pi'(\bar{\bar{x}}^{t,k'}), 1-\pi'(\bar{\bar{x}}^{t,1}), \ldots, 1-\pi'(\bar{\bar{x}}^{t,k})\right), \left(1-\bar{\bar{y}}^{t,1}, \ldots, 1-\bar{\bar{y}}^{t,k'}, 1/2, \ldots, 1/2\right)\right\rangle\Bigg] \\
&= 2\left[\mathop{\mathbb{E}}_{h\sim\pi}\left[\langle a^h, \ell^t\rangle\right] - \mathop{\mathbb{E}}_{h'\sim\pi'}\left[\langle a^{h'}, \ell^t\rangle\right]\right].
\end{aligned}
$$

Where the last transitions stems from the linearity of Error$(\cdot, \bar{\bar{x}}^t, \bar{\bar{y}}^t)$. This concludes the proof. □

Proof of Theorem 3.2. We can see that

$$\begin{aligned}
&\sum_{t=1}^{T} L_{C,\rho^t}(\pi^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in V} \sum_{t=1}^{T} L_{C,\rho^t}(\pi^*, \bar{x}^t, \bar{y}^t) \\
&\leq \sum_{t=1}^{T} L_{C,\rho^t}(\pi^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in \Delta\mathcal{H}} \sum_{t=1}^{T} L_{C,\rho^t}(\pi^*, \bar{x}^t, \bar{y}^t) && (V \subseteq \Delta\mathcal{H}) \\
&= \sum_{t=1}^{T} L_{C,\rho^t}(\pi^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in \mathcal{H}} \sum_{t=1}^{T} L_{C,\rho^t}(\pi^*, \bar{x}^t, \bar{y}^t) && (\text{Linearity of } L_{C,\rho^t}(\cdot, \bar{x}^t, \bar{y}^t) \\
&= 2\left[\sum_{t=1}^{T} \mathop{\mathbb{E}}_{h^t \sim \pi^t}\left[\langle a^{h^t}, \ell^t\rangle\right] - \min_{\pi^* \in \mathcal{H}} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{h^* \sim \pi^*}\left[\langle a^{h^*}, \ell^t\rangle\right]\right]. && (\text{Lemma C.1+Lemma C.2}) \\
&= 2(2k + 4C)R^{\mathcal{A},T,\mathcal{H}} && (\forall t \in [T] : \ell^t \in [0, 2k + 4C]).
\end{aligned}$$

Which concludes the proof. □

Proof of Lemma 3.4. To prove the lemma, it is sufficient to prove that for every $\pi^* \in Q_{\alpha-\epsilon,\gamma}$,

$$\begin{aligned}
&C\epsilon \sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\pi^t, \bar{x}^t, \bar{y}^t, \bar{j}^t) + \sum_{t=1}^{T} Error(\pi^t, \bar{x}^t, \bar{y}^t) - \sum_{t=1}^{T} Error(\pi^*, \bar{x}^t, \bar{y}^t) \\
&\leq \sum_{t=1}^{T} L_{C,\rho^t}(\pi^t, \bar{x}^t, \bar{y}^t) - \sum_{t=1}^{T} L_{C,\rho^t}(\pi^*, \bar{x}^t, \bar{y}^t).
\end{aligned}$$

Which, using Definition 2.7, is equivalent to proving that

$$C\epsilon \sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\pi^t, \bar{x}^t, \bar{y}^t, \bar{j}^t) \leq \sum_{t=1}^{T} C \cdot \left[\pi^t(\rho^{t,1}) - \pi^t(\rho^{t,2})\right] - \sum_{t=1}^{T} C \cdot \left[\pi^*(\rho^{t,1}) - \pi^*(\rho^{t,2})\right].$$

We consider two cases:

1. For rounds t where the panel $\bar{j}^t$ did not detect any (α, γ)-fairness violations, the left hand side of the inequality is 0, and so is the right hand side, since $\rho^{t,1} = \rho^{t,2}$.

2. For rounds t where the panel $\bar{j}^t$ detected an (α, γ)-violation, the left hand side is equal to $C\epsilon$, and the right hand side is at least $C\epsilon$, since, using Lemma B.1, we know

$$\begin{aligned}
\pi^t(\rho^{t,1}) - \pi^t(\rho^{t,2}) &> d^{s_{\rho^{t,1},\rho^{t,2}}(j^{t,1},\ldots,j^{t,m})}(\rho^{t,1}, \rho^{t,2}) + \alpha, \\
-(\pi^*(\rho^{t,1}) - \pi^*(\rho^{t,2})) &\geq \epsilon - \alpha - \min_{r \in [T]} d^{s_{\rho^{t,1},\rho^{t,2}}(j^{r,1},\ldots,j^{r,m})}(\rho^{t,1}, \rho^{t,2}).
\end{aligned}$$

   And hence,

$$\begin{aligned}
&\pi^t(\rho^{t,1}) - \pi^t(\rho^{t,2}) - (\pi^*(\rho^{t,1}) - \pi^*(\rho^{t,2})) \\
&\geq d^{s_{\rho^{t,1},\rho^{t,2}}(j^{t,1},\ldots,j^{t,m})}(\rho^{t,1}, \rho^{t,2}) + \alpha + \epsilon - \alpha - \min_{r \in [T]} d^{s_{\rho^{t,1},\rho^{t,2}}(j^{r,1},\ldots,j^{r,m})}(\rho^{t,1}, \rho^{t,2}) \\
&\geq \epsilon.
\end{aligned}$$

The lemma hence follows.

□

D. Omitted Details from Section 4

D.1. Accuracy and Fairness Regret Rates of Exp2

Proof of Theorem 4.2. Combining Theorems 3.2, 4.1, we know that

$$\sum_{t=1}^{T} L_{C,\rho^t}(\pi^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in Q_{\alpha-\epsilon,\gamma}} \sum_{t=1}^{T} L_{C,\rho^t}(\pi^*, \bar{x}^t, \bar{y}^t) \leq O\left((2k+4C)^{\frac{3}{2}}\sqrt{T\log|\mathcal{H}|}\right).$$

Setting $C = T^{\frac{1}{5}}$, and using Lemma 3.4, we get

$$\begin{aligned} Regret^{err}(\text{Exp2}, T, Q_{\alpha-\epsilon,\gamma}) &\leq O\left((2k+4C)^{\frac{3}{2}}\sqrt{T\log|\mathcal{H}|}\right) - C\epsilon\sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\pi^t, \bar{x}^t, \bar{y}^t, \bar{j}^t) \\ &\leq O\left((2k+4C)^{\frac{3}{2}}\sqrt{T\log|\mathcal{H}|}\right) \\ &\leq O\left(k^{\frac{3}{2}}T^{\frac{4}{5}}\log|\mathcal{H}|^{\frac{1}{2}}\right). \end{aligned}$$

And,

$$\begin{aligned} \sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\pi^t, \bar{x}^t, \bar{y}^t, \bar{j}^t) &\leq \frac{1}{C\epsilon}\left[O\left((2k+4C)^{\frac{3}{2}}\sqrt{T\log|\mathcal{H}|}\right) - Regret^{err}(\text{Exp2}, T, Q_{\alpha-\epsilon,\gamma})\right] \\ &\leq \frac{1}{C\epsilon}\left[O\left((2k+4C)^{\frac{3}{2}}\sqrt{T\log|\mathcal{H}|}\right) + kT\right] \\ &\leq O\left(\frac{1}{\epsilon}k^{\frac{3}{2}}T^{\frac{4}{5}}\log|\mathcal{H}|^{\frac{1}{2}}\right). \end{aligned}$$

□

D.2. Adaptation of Context-Semi-Bandit-FTPL

We next describe an adaptation of Context-Semi-Bandit-FTPL (Syrgkanis et al., 2016) to our setting. Context-Semi-Bandit-FTPL relies on the class $\mathcal{H}$ having a small separator set.

Definition D.1 (Separator set)**.** *We say $S \subseteq \mathcal{X}$ is a separator set for a class $\mathcal{H} : \mathcal{X} \to \{0,1\}$, if for any two distinct hypotheses $h, h' \in \mathcal{H}$, there exists $x \in S$ s.t. $h(x) \neq h'(x)$. We denote $s := |S|$ the size of the separator set.*

Remark D.2. *Classes for which small separator sets are known include conjunctions, disjunctions, parities, decision lists, discretized linear classifiers. Please see an elaborate discussion in Syrgkanis et al. (2016); Neel et al. (2019).*

Context-Semi-Bandit-FTPL further relies on access to an (offline) optimization oracle for the corresponding problem (equivalent to a weighted ERM oracle), which we define next.

Definition D.3 (Optimization oracle)**.** *Context-Semi-Bandit-FTPL assumes access to oracle of the form*

$$M((\bar{\bar{x}}^t)_{t=1}^N, (\hat{\ell}^t)_{t=1}^N) = \operatorname*{argmin}_{h\in\mathcal{H}} L(h, (\bar{\bar{x}}^t, \hat{\ell}^t)),$$

where $\hat{\ell}^t$ denotes the loss estimates held by Context-Semi-Bandit-FTPL for round t, and L denotes the cumulative loss, over

linear loss functions of the form $f^t(a) = \langle a, \ell \rangle$. In our construction, this is equivalent to

$$\begin{aligned}
&\operatorname*{argmin}_{h \in \mathcal{H}} L(h^t, (\bar{\bar{x}}^t, \hat{\ell}^t)) \\
&:= \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{t=1}^{N} \left\langle a_h^t, \hat{\ell}^t \right\rangle && \textit{(Definition of L)} \\
&= \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{t=1}^{N} \sum_{i=1}^{k+2C} h(\bar{\bar{x}}^{t,i}) \cdot \hat{\ell}^{t,i} + (1 - h(\bar{\bar{x}}^{t,i})) \cdot \frac{1}{2} && \textit{(Algorithm 3)} \\
&= \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{t=1}^{N} \sum_{i=1}^{k+2C} h(\bar{\bar{x}}^{t,i}) \cdot (\hat{\ell}^{t,i} - \frac{1}{2}) && \textit{(Subtraction of constant).}
\end{aligned}$$

In broad strokes, Context-Semi-Bandit-FTPL operates by, at each round, first sampling a set of "fake" samples z^t, that is added to the history of observed contexts and losses by the beginning of round t, denoted by H^t. The algorithm then invokes the optimization oracle on the extended set $z^t \cup H^t$, and deploys $h^t \in \mathcal{H}$ that is returned by the oracle.

Equivalently, this process can be seen as the learner, at each round t, (implicitly) deploying a distribution over hypotheses from the base class $\mathcal{H}$, denoted by π^t, then sampling and deploying a single hypothesis $h^t \sim \pi^t$. As it is observed in general (see, e.g. the discussion in Neu and Bartók (2013)), the specific weights this implicit distribution places on each of $h \in \mathcal{H}$ on any given round cannot be expressed in closed-form. Instead, FTPL-based algorithms resort to sampling actions from the distribution, leveraging the linearity of the loss function in obtaining expected regret guarantees.

For our purposes, however, such a method of assessing the loss on single realized hypotheses h^t from π^t could be problematic, since we rely on the panel $\bar{j}^t$ reporting its feedback upon observing the actual distribution π^t. Querying the panel instead using realizations $h^t \sim \pi^t$ could lead to an over-estimation of the unfairness loss, as we demonstrate next.

Lemma D.4. *There exist $\alpha, \gamma, m, k > 0$, $\mathcal{H} : \mathcal{X} \to \{0, 1\}$, $(x, x') \in \mathcal{X}^2$, $\bar{j} : \mathcal{X}^k \to \mathcal{X}^2$, and $\pi \in \Delta\mathcal{H}$ for which, simultaneously,*

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

We defer the proof of Lemma D.4 to the end of this section.

We therefore adapt Context-Semi-Bandit-FTPL to our setting by adding a resampling process at each iteration of the algorithm. Our approach is similar in spirit to the resampling-based approach in (Bechavod et al., 2020) (which offer an adaptation for the full information variant of the algorithm), however, unlike their suggested scheme, which requires further restricting the power of the adversary to, at each round t, to not depend on the policy π^t deployed by the learner (instead, they only allow dependence on the history of the interaction until round $t - 2$), the adaptation we next propose would not require such a relaxation.

We next abstract out the implementation details of the original Context-Semi-Bandit-FTPL that remain unchanged (namely, the addition of "fake" samples, and solving of the resulting optimization problem at the beginning of each round, and the loss estimation process at the end of it), to focus on the adaptation.

Our adaptation will work as follows: the learner first initializes Context-Semi-Bandit-FTPL-With-Resampling with a pre-computed separator set S for $\mathcal{H}$. Then, at each round t, the learner (implicitly) deploys π^t according to Context-Semi-Bandit-FTPL-With-Resampling. The environment then selects individuals $\bar{x}^t$ and their labels $\bar{y}^t$, only revealing $\bar{x}^t$ to the learner. The environment proceeds to select a panel of auditors $(j^{t,1}, \ldots, j^{t,m})$. The learner invokes Context-Semi-Bandit-FTPL-With-Resampling and receives an estimated policy $\hat{\pi}^t$, and a realized predictor $\hat{h}^t$ sampled from $\hat{\pi}^t$. The learner then predicts the arriving individuals $\bar{x}^t$ using $\hat{h}^t$, only observing feedback on positively labelled instances. The panel then reports its feedback $\hat{\rho}^t$ on $(\hat{\pi}^t, \bar{x}^t)$. The learner invokes the reduction (Algorithm 3), using $\bar{x}^t$, $\bar{y}^t$, $\hat{h}^t$, $\hat{\rho}^t$, and C, and receives (ℓ^t, a^t). The learner updates Context-Semi-Bandit-FTPL-With-Resampling with (ℓ^t, a^t) and lets it finish the loss estimation process and deploy the policy for the next round. Finally, the learner suffers misclassification loss with respect to $\hat{h}^t$, and unfairness loss with respect to $\hat{\pi}^t$. The interaction is summarized in Algorithm 4.

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 $\bar{x}^t$;
  Learner 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 $\rho^t = \vec{j}^{t,\alpha,\gamma}_{j^1,\ldots,j^m}(\hat{\pi}^t, \bar{x}^t)$;
  (ℓ^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 $Unfair(\hat{\pi}^t, \bar{x}^t, \bar{y}^t, \vec{j}^t)$;
  Learner deploys $\pi^{t+1} \in \Delta\mathcal{H}$ according to Context-Semi-Bandit-FTPL-With-Resampling;
end for

Algorithm 5 Context-Semi-Bandit-FTPL-With-Resampling(S, ω, L)

Parameters: Class of predictors $\mathcal{H}$, number of rounds T, optimization oracle M, separator set S, parameters ω, L;
for $t = 1, \ldots, T$ **do**
  for $r = 1, \ldots, R$ **do**
    Sample predictor $h^{t,r}$ according to $\mathcal{D}^t$;
  end for
  Set and report $\hat{\pi}^t = \mathbb{U}(h^{t,1}, \ldots, h^{t,R})$, $\hat{h}^t \sim \hat{\pi}^t$;
  Receive back (ℓ^t, a^t) from reduction;
  Continue as original Context-Semi-Bandit-FTPL;
end for

As for the resampling process we add to the original Context-Semi-Bandit-FTPL: at each round we define "sampling from $\mathcal{D}^t$" to refer to the process of first sampling the additional "fake" samples to be added, and then solving the resulting optimization problem over the original and the "fake" samples, to produce a predictor $h^{t,r}$. We repeat this process R times, to produce an empirical distribution $\hat{\pi}^t$, and select a single predictor $\hat{h}^t$ from it, which are reported to the learner. Once receiving back (ℓ^t, a^t) from the learner, Context-Semi-Bandit-FTPL-With-Resampling proceeds to perform loss estimation, as well as selecting the next policy, in a similar fashion to the original version of Context-Semi-Bandit-FTPL. This adaptation is summarized in Algorithm 5.

We note that for the described adaptation, we will next prove accuracy and fairness guarantees for the sequence of estimated policies, $(\hat{\pi}^t)_{t=1}^T$, rather than for the underlying policies $(\pi^t)_{t=1}^T$. One potential issue with this approach is that the Lagrangian loss at each round is defined using the panel's reported pair ρ^t, which is assumed to be reported with respect to π^t. Here, we instead consider the Lagrangian loss using $\hat{\rho}^t$, which is based on the realized estimation $\hat{\pi}^t$. However, this issue can be circumvented with the following observation: on each round, there are k^2 options for selecting ρ^t, which are simply all pairs in $\bar{x}^t$. We will prove next, that since resampling for $\hat{\rho}^t$ is done after $\bar{x}^t$ is fixed, with high probability, the Lagrangian loss for each of π^t and $\hat{\pi}^t$ will take values that are close to each other, when defined using any possible pair $\hat{\rho}^t$ from $\bar{x}^t$. Hence, by allowing the adversary the power to specify $\hat{\rho}^t$ after $\hat{\pi}^t$ is realized, we do not lose too much. We formalize this argument next.

Theorem D.5. *In the setting of (adapted) individually fair online learning with one-sided feedback (Algorithm 4), running Context-Semi-Bandit-FTPL-With-Resampling (Algorithm 5) with $L = T^{\frac{1}{3}}$, and optimally selected ω, using the sequence $(a^t, \ell^t)_{t=1}^T$ generated by the reduction in Algorithm 3 (when invoked every round on $\bar{x}^t$, $\bar{y}^t$, $\hat{h}^t$, $\hat{\rho}^t$, and C), yields the*

following guarantee, for any $U \subseteq \Delta\mathcal{H}$,

$$\sum_{t=1}^{T} L_{C,\hat{\rho}^t}(\hat{\pi}^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in U} \sum_{t=1}^{T} L_{C,\hat{\rho}^t}(\pi^*, \bar{x}^t, \bar{y}^t) \leq O\left((2k+4C)^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{2}{3}} \log|\mathcal{H}|^{\frac{1}{2}}\right) + 2(2k+4C)T\sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2R}}.$$

In order to prove Theorem D.5, we will first prove the following lemma, regarding the difference of losses between the underlying π^t and the estimated $\hat{\pi}^t$.

Lemma D.6. *With probability $1-\delta$ (over the draw of $(h^{t,1}, \ldots, h^{t,R})_{t=1}^T$), for any arbitrary sequence of reported pairs $(\rho^t)_{t=1}^T$,*

$$\sum_{t=1}^{T} \left| \mathop{\mathbb{E}}_{\hat{h}^t \sim \hat{\pi}^t} \left[\langle a^{\hat{h}^t}, \ell^t \rangle\right] - \mathop{\mathbb{E}}_{h^t \sim \pi^t} \left[\langle a^{h^t}, \ell^t \rangle\right] \right| \leq 2(2k+4C)T\sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2R}}.$$

Proof. Using Chernoff bound, we can bound the difference in predictions between the underlying and the estimated distributions, for each of the contexts context in $\bar{x}^t$, for any round t:

$$\forall t \in [T], i \in [k] : \Pr\left[\left|\hat{\pi}^t(\bar{x}^{t,i}) - \pi^t(\bar{x}^{t,i})\right| \geq \sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2R}}\right] \leq \frac{\delta}{kT}.$$

Union bounding over all rounds, and each of the contexts in a round, we get that, with probability $1-\delta$,

$$\forall t \in [T], i \in [k] : \left|\hat{\pi}^t(\bar{x}^{t,i}) - \pi^t(\bar{x}^{t,i})\right| \leq \sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2R}}.$$

Hence, when considering pairs of individuals, and using triangle inequality, we know that with probability $1-\delta$,

$$\forall t \in [T], i,j \in [k] : \left|\left[\hat{\pi}^t(\bar{x}^{t,i}) - \hat{\pi}^t(\bar{x}^{t,j})\right] - \left[\pi^t(\bar{x}^{t,i}) - \pi^t(\bar{x}^{t,j})\right]\right| \leq 2\sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2R}}.$$

Hence, by construction of the losses and actions sequence (using the reduction in Algorithm 3 with respect to $\hat{\rho}^t$), with probability $1-\delta$,

$$\forall t \in [T], \hat{\rho}^t \in \bar{x}^t : \left| \mathop{\mathbb{E}}_{\hat{h}^t \sim \hat{\pi}^t} \left[\langle a^{\hat{h}^t}, \ell^t \rangle\right] - \mathop{\mathbb{E}}_{h^t \sim \pi^t} \left[\langle a^{h^t}, \ell^t \rangle\right] \right| \leq 2(2k+4C)\sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2R}}.$$

Summing over rounds, with probability $1-\delta$, for any arbitrary sequence of reported pairs $(\rho^t)_{t=1}^T$,

$$\sum_{t=1}^{T} \left| \mathop{\mathbb{E}}_{\hat{h}^t \sim \hat{\pi}^t} \left[\langle a^{\hat{h}^t}, \ell^t \rangle\right] - \mathop{\mathbb{E}}_{h^t \sim \pi^t} \left[\langle a^{h^t}, \ell^t \rangle\right] \right| \leq 2(2k+4C)T\sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2R}}.$$

Which concludes the proof of the lemma. □

We are now ready to prove the regret bound of Context-Semi-Bandit-FTPL-With-Resampling.

Proof of Theorem D.5. Note that by the guarantees of Context-Semi-Bandit-FTPL, and since $\|\ell^t\|_* \leq 2k+4C$, for any arbitrary sequence $(\rho^t)_{t=1}^T$,

$$2\left[\sum_{t=1}^{T} \mathop{\mathbb{E}}_{h^t \sim \pi^t} \left[\langle a^{h^t}, \ell^t \rangle\right] - \min_{\pi^* \in \Delta\mathcal{H}} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{h^* \sim \pi^*} \left[\langle a^{h^*}, \ell^t \rangle\right]\right] \leq O\left((2k+4C)^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{2}{3}} \log|\mathcal{H}|^{\frac{1}{2}}\right).$$

Using Lemma D.6 and the triangle inequality, we conclude

$$\sum_{t=1}^{T} L_{C,\hat{\rho}^t}(\hat{\pi}^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in U} \sum_{t=1}^{T} L_{C,\hat{\rho}^t}(\pi^*, \bar{x}^t, \bar{y}^t) \leq O\left((2k+4C)^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{2}{3}} \log|\mathcal{H}|^{\frac{1}{2}}\right) + 2(2k+4C)T\sqrt{\frac{\log\left(\frac{2kT}{\delta}\right)}{2R}}.$$

□

Proof of Theorem 4.6. Using Theorem D.5 with $C = T^{\frac{4}{45}}$, $R = T^{\frac{38}{45}}$, we know that, with probability $1-\delta$,

$$\sum_{t=1}^{T} L_{C,\bar{\rho}^t}(\hat{\pi}^t, \bar{x}^t, \bar{y}^t) - \min_{\pi^* \in Q_{\alpha-\epsilon,\gamma}} \sum_{t=1}^{T} L_{C,\hat{\rho}^t}(\pi^*, \bar{x}^t, \bar{y}^t) \leq \tilde{O}\left(k^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{41}{45}} \log|\mathcal{H}|^{\frac{1}{2}}\right).$$

Using Lemma 3.4, we get, with probability $1-\delta$,

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

And,

$$\begin{aligned} \sum_{t=1}^{T} Unfair^{\alpha,\gamma}(\hat{\pi}^t, \bar{x}^t, \bar{y}^t, \bar{j}^t) &\leq \frac{1}{C\epsilon}\left[\tilde{O}\left(k^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{41}{45}} \log|\mathcal{H}|^{\frac{1}{2}}\right) - Regret^{err}(T)\right] \\ &\leq \frac{1}{C\epsilon}\left[\tilde{O}\left(k^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{41}{45}} \log|\mathcal{H}|^{\frac{1}{2}}\right) + kT\right] \\ &\leq \tilde{O}\left(\frac{1}{\epsilon} k^{\frac{11}{4}} s^{\frac{3}{4}} T^{\frac{41}{45}} \log|\mathcal{H}|^{\frac{1}{2}}\right). \end{aligned}$$

□

Proof of Lemma D.4. We set $\alpha = 0.2, \gamma = 1$ and $k = 2$. We define the context space to be $\mathcal{X} = \{x, x'\}$ (each with label 1), and the hypothesis class as $\mathcal{H} = \{h, h'\}$, where $h(x) = h'(x') = 1$, and $h(x') = h'(x) = 0$. We set $m = 1$, and the panel $\bar{j}$, that hence consists of a single auditor, to reflect the judgements of j, where $d^j(x, x') = 0.1$. Define $\pi \in \Delta\mathcal{H}$ as h with probability 0.5, and as h' with probability 0.5.

We denote $\bar{x} = (x, x')$, $\bar{y} = (1, 1)$.

Next, note that

$$\begin{aligned} h(x) - h(x') = 1 > 0.3 = d^j(x, x') + \alpha, \\ h'(x') - h'(x) = 1 > 0.3 = d^j(x, x') + \alpha. \end{aligned}$$

Hence,

$$\mathop{\mathbb{E}}_{h\sim\pi}\left[unfair^{\alpha,\gamma}(h, \bar{x}, \bar{y}, \bar{j})\right] = 0.5 unfair^{\alpha,\gamma}(h, \bar{x}, \bar{y}, \bar{j}) + 0.5 unfair^{\alpha,\gamma}(h', \bar{x}, \bar{y}, \bar{j}) = 1.$$

On the other hand,

$$\pi(x) - \pi(x') = \pi(x') - \pi(x) = 0 < 0.3 = d^j(x, x') + \alpha.$$

Hence,

$$unfair^{\alpha,\gamma}(\pi, \bar{x}, \bar{y}, \bar{j}) = 0.$$

Which proves the lemma. □