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 present a novel auditing scheme, capable of utilizing feedback from dynamically-selected panels of multiple, possibly inconsistent, auditors regarding fairness violations. In particular, we show how our proposed auditing scheme allows for *algorithmically* exploring the resulting accuracy-fairness frontier, with no need for additional feedback from auditors. We then present 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 & Lugosi, 2009; György et al., 2007), allowing us to leverage algorithms for contextual combinatorial semi-bandits to establish *multi-criteria* no regret guarantees in our setting, *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. 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.

[1]Hebrew University and University of Pennsylvania [2]University of Pennsylvania. Correspondence to: Yahav Bechavod <yahav@seas.upenn.edu>, Aaron Roth <aaroth@cis.upenn.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

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. Overview of Results

Our contributions lie on a conceptual as well as a technical level. We first present a novel auditing scheme based on dynamically selected panels of multiple, possibly inconsistent, auditors, reporting fairness violations (Section 2.1). We formulate our learning problem as an optimization problem of a joint objective using a Lagrangian formulation (Section 2.2). We then present our online learning framework with one-sided label feedback and additional fairness feedback from panels (Section 2.3). Our main technical contributions are given in Sections 3, 4. We first present an efficient reduction to the contextual combinatorial semi-bandit setting, allowing us to upper bound the Lagrangian regret in our setting (Section 3). We then establish an equivalence between auditing by panels and auditing by "representative" auditors, which is an important technical step in our analysis, and then show how the Lagrangian regret guarantee can be utilized to provide *multi-criteria* no regret guarantees, that hold *simultaneously* for accuracy and fairness (Section 4). Finally, we present an oracle-efficient algorithm for our setting and analyze the resulting rates for accuracy and fairness (Section 5).

1.2. Related Work

Our work is related to two strands of literature: learning with one-sided feedback, and individual fairness in machine learning. 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. (2006b) 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. Cesa-Bianchi et al. (2006a) focuses on linear models, and propose an active learning approach based on the predictions made by the deployed predictor at each round, in the face of one-sided feedback. Sculley (2007) focused on practical challenges in learning with one-sided feedback in the context of spam filtering, and suggested the utilization of methods from the active learning literature to reduce the label complexity encountered in practice. Jiang et al. (2021) focuses on learning with generalized linear models in an online one-sided feedback setting, and propose a data-driven adaptive approach based on variance estimation techniques. De-Arteaga et al. (2018) and Lakkaraju et al. (2017) propose techniques for imputing missing labels using feedback from human experts. Zeng et al. (2017) and Lakkaraju & Rudin (2017) propose statistical techniques for assigning missing labels.

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. Ensign et al. (2018) and Elzayn et al. (2019) focus on the tasks of predictive policing and related resource allocation problems, and give algorithms for these tasks under a censored feedback model. Kleinberg et al. (2017) explores techniques to mitigate the problem of one-sided feedback in the context of judicial bail decisions. Gupta & Kamble (2019) study a time-dependent variant of individual fairness they term "individual fairness in hindsight". Yurochkin et al. (2020) consider a variant of individual fairness which asks for invariance of the learned predictors with respect to "sensitive" variables. Mukherjee et al. (2020) investigate ways to learn the metric from data. Lahoti et al. (2019) focus on the task of learning individually fair representations.

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 & Yona (2018) give a statistical treatment of individual fairness in a batch setting with examples drawn i.i.d. from some distribution, while assuming full access to a similarity metric, 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.

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.

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 $h : \mathcal{X} \to \mathcal{Y}$, and 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 by Panels

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

$$\pi(x) - \pi(x') > d(x, x') + \alpha.^{2}$$

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, $\forall \pi \in \Delta\mathcal{H}, \bar{x} \in \mathcal{X}^k$,

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

where $\bar{x} = (\bar{x}^1, \ldots, \bar{x}^k)$, $d^j : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is auditor j^α's (implicit) distance function, and $v \in \mathcal{X}$ is some "default" context. When clear from context, we will abuse notation, and simply 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 case the auditor does not consider there to be any fairness violations, we define its output to be a "default" value, $(v, v) \in \mathcal{X}^2$, 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

[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.

[2] Note, importantly, that the definition of a fairness violation not only encodes the existence, but also the "direction" of the reported fairness violation (which of the two individuals of the reported pair received the higher prediction). As we will see in Section 3, this will be important in our construction.

that there is one at one extreme, to requiring unanimity amongst the auditors at the other extreme.

We next define a panel to return a pair of individuals on which the required majority of panelists agree that a fairness violation has occurred, if one or more such pairs exists.[3]

Definition 2.4 (Panel)**.** Let $\alpha \geq 0, 0 \leq \gamma \leq 1, m \in \mathbb{N} \backslash \{0\}$. 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. 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. In particular, our framework is also capable of handling collections of disagreeing auditors, as cases where the required majority of objections is not reached within a panel with respect to any of the pairs naturally result in none of the pairs being reported.

With our proposed auditing scheme, we wish to suggest a platform that would be helpful in capturing and reflecting society's perceptions towards what should be considered "fair" in a specific context. However, this must be done *carefully*, as multiple studies have indicated the existence of implicit biases affecting human judgements (see, e.g., Levinson et al. (2016); Rachlinski et al. (2009)). Mitigating the effects of these biases, while allowing for the elicitation of critical information regarding fairness judgements is a key motivation for our proposed auditing scheme. To that end, one might envision fairness-related judgements arriving from a panel of *multiple* auditors of diverse backgrounds, trainings, life experiences, etc.[4]

Remark 2.5 (On Exploring the Accuracy-Fairness Tradeoff Frontier). 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, however, that different values of γ for the panel *do not* change the fundamental 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.

Example 2.6. As a running example along this work, we will consider a loan approval setting, in which a learner (for instance, a government-backed financial institution) wishes to learn and deploy lending policies that are, simultaneously, *highly performing* and *individually fair*. In this example, $x \in \mathcal{X}$ would represent features of a loan applicant, such as repayment history, level of education, address, etc., and $y \in \mathcal{Y} = \{0, 1\}$ represents whether the applicant is creditworthy or not. $\pi \in \Delta\mathcal{H}$ is a lending policy, mapping incoming applicants to a score in $[0, 1]$, based on which loan decisions are made. In this context, a panel might consist of ethicists and financial experts with familiarity with the lending industry and the history of *redlining* (Rothstein, 2018). Following Definition 2.4, given a group of k loan applicants and a deployed lending policy π, the panel would inspect the assigned scores, and report back in case it finds a pair of applicants which at least a γ fraction of the panel members believe are similarly situated, but were treated very differently by the policy.

2.2. Lagrangian Loss Formulation

We define the three types of losses we will use in our setting.

Definition 2.7 (Misclassification loss)**.** We define the misclassification loss as, for all $\pi \in \Delta\mathcal{H}, \bar{x} \in \mathcal{X}^k, \bar{y} \in \{0, 1\}^k$:

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

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

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

Definition 2.8 (Unfairness loss)**.** Let $\alpha \geq 0, 0 \leq \gamma \leq 1$. We define the unfairness loss as, for all $\pi \in \Delta\mathcal{H}, \bar{x} \in \mathcal{X}^k$, $\bar{j} = \bar{j}^{\alpha,\gamma}_{j^1,\ldots,j^m} : \mathcal{X}^k \to \mathcal{X}^2$,

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

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

[3] We are agnostic as to how the panel identifies this pair — perhaps through an interactive, deliberative process. In particular, panelists might internally need to discuss more than one pair that they feel to be in violation. In particular, for $\gamma > \frac{1}{m}$, the auditors might in the worst case need to each list all pairs that they find to be in violation to establish that there is a pair of decisions that is judged to be in violation by all of them.

[4] Additionally, our model advocates for dynamically-selected panels instead of fixed ones, as having a fixed group of auditors may risk leaving too much power in the hands of the same few individuals over a long period of time. Finally, practically speaking, it may also be infeasible for human auditors to review more than a certain number of decisions in a certain period of time, a limitation which may be circumvented by periodically changing the panel.

Finally, the Lagrangian loss will be useful in our analysis.

Definition 2.9 (Lagrangian loss)**.** Let $C > 0$, $\rho = (\rho^1, \rho^2) \in \mathcal{X}^2$. We define the (C, ρ)-Lagrangian loss as, for all $\pi \in \Delta\mathcal{H}$, $\bar{x} \in \mathcal{X}^k$, $\bar{y} \in \{0,1\}^k$,

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

We will later instantiate the Lagrangian loss with (ρ^1, ρ^2) being the panel's reported pair (as in Definition 2.4). 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. The interaction proceeds in a sequential fashion, where on each round, the learner first deploys a policy $\pi^t \in \Delta\mathcal{H}$, then the environment selects k individuals $\bar{x}^t \in \mathcal{X}^k$, and their labels $\bar{y}^t \in \mathcal{Y}^k$, possibly in an adversarial fashion. The learner is only shown $\bar{x}^t$. The environment then selects a panel of auditors $(j^{t,1}, \ldots, j^{t,m}) \in \mathcal{J}^m$, possibly in an adversarial fashion. The learner predicts $\bar{x}^t$ according to π^t. The panel then reports whether a fairness violation was found, according to the predictions made by π^t. Finally, the learner observes the true label only for positively-predicted individuals in $\bar{x}^t$, and suffers two types of loss: a misclassification loss (note that this loss may *not* be observable to the learner due to the one-sided feedback) and an unfairness loss. Our setting is summarized in Algorithm 1.

One-sided feedback Our one-sided feedback structure (classically known as "apple tasting") is fundamentally 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 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. Going back to our running example - if we misclasify a creditworthy individual as not creditworthy, and deny them a loan, we never get to *observe* our error (because we do not observe the counter-factual world in which we gave them a loan that they had the opportunity to repay). Nevertheless, this is a classification error, and one that we account for in our objective.

Remark 2.10. There are several additional key motivations behind the choice of the online, adversarial, setting we explore. First, we note that when considering a batch setting, the collected data may already be "skewed", to only include individuals who were actually accepted by a past policy. This, in turn, risks replicating biases of historical policies. Additionally, when studying an online setting, note that in many problem domains relevant to our work, arriving individuals may not necessarily adhere to standard statistical assumptions, due to, for example: (i) strategic effects (individuals performing feature manipulations in anticipation of a specific policy, or choosing whether to even apply based on their perceived chances of receiving a positive outcome), (ii) distribution shifts over time (e.g. the ability to repay a loan may be affected by changes to the economy or recent events), (iii) adaptivity to previous decisions (e.g. if an individual receives a loan, that may affect the ability to repay future loans by this individual or her vicinity in the future).

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

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

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

end for

To measure performance, we will ask for algorithms that 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 2.11 (Error regret)**.** We define the error regret of an algorithm $\mathcal{A}$ against a comparator class $U \subseteq \Delta\mathcal{H}$ to be

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

Definition 2.12 (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

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

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

Definition 2.13 (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

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

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-bandit, which we will see next.

3. Reduction to Contextual Combinatorial Semi-Bandit

In this section, we present our main technical result: a reduction from individually fair online learning with one-sided feedback (Algorithm 1) to the setting of (adversarial) contextual combinatorial semi-bandit.

3.1. Contextual Combinatorial Semi-Bandit

We begin by formally describing the contextual combinatorial semi-bandit setting.[5] The setting can be viewed as an extension of the classical k-armed contextual bandit problem, to a case where k instances arrive every round, each to be labelled as either 0 or 1. However, at each round, the action set over these labellings is restricted only to a subset $A^t \subseteq \{0,1\}^k$, where each action corresponds the vector containing the predictions of a hypothesis $h \in \mathcal{H}$ on the arriving contexts. Finally, the learner suffers loss that is a linear function of all specific losses for each of the k instances, and is restricted to only observe the coordinates of the loss vector on instances predicted as 1. The setting is summarized in Algorithm 2.

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

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

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

3.2. Reduction

Our reduction consists of two major components: encoding the fairness constraints, and translating one-sided feedback to semi-bandit feedback. We start by extending the sample set at round t, to encode C copies of each of the individuals in the pair reported by the panel, where we append label of 0 to the first individual, and label of 1 to the second, in order to "translate" unfairness into error. We next translate one-sided feedback to semi-bandit feedback, by constructing the first half of the loss vector ℓ^t, to return a loss of 0 or 1 for positively-predicted instances (according to the, observable true label), and the second half to return a loss of $1/2$ for negatively-predicted instances (regardless of the true label, which is unobservable, since the prediction is negative). We note that this transformation of the standard $0-1$ loss is regret-preserving, and the resulting losses are also linear and non-negative (which will be important in our analysis, as we will see in Section 5). Finally, the first half of the action vector is constructed by simply invoking the selected hypothesis by the learner, h^t, on each of the contexts in the augmented $\bar{\tilde{x}}^t$, while the second half reflects the opposites of the predictions made in the first half.

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

1. $\forall a \in \{\rho^{t,1}, \rho^{t,2}, 0, 1, 1/2\}$:

[5] The combinatorial (full) bandit problem formulation is due to Cesa-Bianchi & 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).

$$\bar{a} := \overbrace{(a, \ldots, a)}^{\text{C times}}, \quad \bar{\bar{a}} := \overbrace{(a, \ldots, a)}^{\text{k+2C times}}.$$

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

The reduction is summarized in Algorithm 3.

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, within a multiplicative factor of 2 times the dimension of the output of the reduction.

For the next theorem, we assume the existence of an algorithm $\mathcal{A}$ for the contextual combinatorial semi-bandit setting (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 $\text{Regret}(\mathcal{A}, T, \mathcal{H}) \leq R^{\mathcal{A},T,\mathcal{H}}$. We next show how the regret guarantee of algorithm $\mathcal{A}$ can be used to upper bound the Lagrangian regret in our setting.

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 not only against the best classifier in $\mathcal{H}$, but rather against the set of all possibly randomized policies $\Delta\mathcal{H}$,

4. Multi-Criteria No Regret Guarantees

In this section, we will show how the guarantees established in Section 3 can be leveraged to establish multi-criteria no regret guarantees, simultaneously for accuracy and fairness.

We begin by establishing an equivalence between auditing by panels and auditing by instance-specific "representative" auditors, which will be useful in our analysis.

Lemma 4.1. *Let $(x, x') \in \mathcal{X}^2$, $(j^1, ..., j^m) \in \mathcal{J}^m$. Then, there exists an index $s = s_{x,x'}(j^1, \ldots, j^m) \in [m]$ such that 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 $\bar{j}^{\alpha,\gamma}_{j^1,\ldots,j^m}$.*
2. *π has an α-violation on (x, x') w.r.t. auditor j^s.*

The crucial aspect of this lemma is that the index $s_{x,x'}(j^1, \ldots, j^m)$ of the "pivotal" auditor is defined independently of π. We refer the reader to Appendix C.1 for a complete discussion, and in particular, to Figure 1 in Appendix C.1, for an illustration of Lemma 4.1.

Next, we will see how the guarantees established in Section 3, along with the reduction to "representative" auditors (Lemma 4.1), allow for providing simultaneous guarantees for each of accuracy and fairness. We begin by defining the comparator set as all policies in $\mathcal{H}$ that are, for every round $t \in [T]$, (α, γ)-fair on the arriving individuals of the round $\bar{x}^t$, with respect to the realized panel in that round $\bar{j}^t$. Note that this set is only defined in *hindsight*.

Definition 4.2 ((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 the rounds in the run of the algorithm as

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

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

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

High-level proof idea It is sufficient to prove that for every $\pi^* \in Q_{\alpha-\epsilon,\gamma}$, if we set π^* as the comparator (instead of taking the minima for each the error regret and the Lagrangian regret), the inequality holds. We will hence fix such $\pi^* \in Q_{\alpha-\epsilon,\gamma}$, and using Definition 2.9, see that

the sums of the $Error$ terms in both sides of the inequality cancel out. We will then divide the analysis to two cases: rounds on which no (α, γ)-violation was detected, and rounds where such a violation was detected. For the first type, the equality holds, since by definition of the panel, $\rho^{t,1} = \rho^{t,2}$. For the second type, the left hand side of the equality is simply $C\epsilon$. As for the right hand side, we first use Lemma 4.1 to move from panels to carefully defined "representative" single auditors, then argue that the right hand side of the equality is at least $C\epsilon$, since π^* is guaranteed to have no $(\alpha - \epsilon)$-violations with respect to any of the "representative" auditors of the panels in the interaction.

Remark 4.4. Note that while Lemma 4.3 is crucial in our analysis, it does not directly give the desired guarantees for accuracy and fairness. The reason is that $Regret^{err}(\mathcal{A}, T, Q_{\alpha-\epsilon,\gamma})$ can be *negative*. Hence, even a sub-linear bound on the Lagrangian regret in the right hand side of the inequality cannot be immediately translated to *simultaneous* sub-linear guarantees for accuracy and fairness. For that, we will need an extra step to carefully interpolate between the two forms of loss in the Lagrangian, which can be found in the proof of Theorem 5.3 in Appendix D.2.

5. Oracle-Efficient Algorithm

In Sections 3 and 4, we have shown how it is possible to translate the guarantees of an algorithm $\mathcal{A}$ for the contextual combinatorial semi-bandit problem (as described in Algorithm 2) to establish simultaneous guarantees for accuracy and fairness in our setting of individually fair online learning with one-sided feedback (Algorithm 1). But our work is not done. Note, in particular, that in order to invoke our reduction (Algorithm 3), one must first query the panel for fairness violations regarding the deployed policy. But what if this policy is not kept *explicitly* by the algorithm $\mathcal{A}$? Generally speaking, the most straightforward way for an algorithm to deploy a randomized policy is to explicitly maintain a vector of weights representing the probability of each classifier in the base class $\mathcal{H}$. This is, for example, what the multiplicative weights algorithm does. This is very costly; it has running time that scales linearly with $|\mathcal{H}|$, which is typically at least exponential in the data dimension. On the other hand, *oracle-efficient* algorithms generally refrain from explicitly maintaining the deployed policy $\pi \in \Delta\mathcal{H}$, and instead implement a way to sample from this (implicitly defined) probability distribution via an efficient reduction to a batch learning problem.[6]

Hence, our efforts next are towards establishing our desired accuracy and fairness guarantees *efficiently*. To this end, we present an oracle-efficient algorithm for our setting - a resampling-based variant of the Context-Semi-Bandit-FTPL algorithm by Syrgkanis et al. (2016). In line with the above discussion, Context-Semi-Bandit-FTPL does not explicitly maintain the deployed policy at any point in its run, but rather, on each round, samples a *realization* of this policy and calculates the loss with respect to this realization. This is useful, as given the specific implementation of the algorithm, one can efficiently sample a realization $h \in \mathcal{H}$ according to $\pi \in \Delta\mathcal{H}$ even though π is not maintained explicitly. However, in the case of a our joint Lagrangian loss, evaluating the loss on single realizations could be problematic, as it may lead to overestimating the unfairness loss (being a strictly sub-additive function). We elaborate on this point in Lemma D.5 in Appendix D.

Hence, we next construct a variant where the process of sampling the hypothesis on each round is repeated R times, to form an accurate enough estimate of the implicit distribution.[7] Our construction, which we term Context-Semi-Bandit-FTPL-With-Resampling, is formally defined in Algorithm 4 and Algorithm 5 in Appendix D.2, and yields the guarantees we present next. For the following theorem, it is assumed that Context-Semi-Bandit-FTPL-With-Resampling has access to a (pre-computed) separator set S of size s for the class $\mathcal{H}$, and an (offline) optimization oracle for $\mathcal{H}$. The optimization oracle assumption can be viewed equivalently as assuming access to a weighted ERM oracle for $\mathcal{H}$. We next describe the separator set assumption.

Definition 5.1. 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 $x \in S$ such that $h(x) \neq h'(x)$.

Remark 5.2. Classes for which small separator sets are known include conjunctions, disjunctions, parities, decision lists, discretized linear classifiers. Please see more elaborate discussions in Syrgkanis et al. (2016) and Neel et al. (2019).

The guarantees of Theorem 5.3 can be interpreted as follows: accuracy-wise, the resulting algorithm is competitive with the performance of the most accurate policy that is fair (i.e. in $Q_{\alpha-\epsilon,\gamma}$). Fairness-wise, the number of rounds in which there exist (one or more) fairness violations, is sub-linear.

Theorem 5.3. *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 guarantees, for any $\epsilon \in [0, \alpha]$,*

[6] In Appendix D.1 we also present a simpler, *inefficient*, algorithm for the contextual combinatorial semi-bandit problem, due to Bubeck et al. (2012), and analyze its resulting rates in our setting.

[7] Resampling is required, as it is observed in general (see, e.g. the discussion in Neu & Bartók (2013)), the specific weights the implicit distribution maintained by Context-Semi-Bandit-FTPL places on each of $h \in \mathcal{H}$ cannot be expressed in closed-form.

simultaneously:

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

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

Remark 5.4. In particular, one could imagine that a potential issue with such a resampling-based approach is that since we lack access to the implicit, deployed, policy π^t, we instead define the Lagrangian loss incorporating reported fairness violations based on the *realized* estimate $\hat{\pi}^t$ of π^t. In adversarial online learning, however, revealing the random bits of the learner to the adversary is generally not permitted. To remedy this issue, In their algorithm for the full information, single auditor, setting, Bechavod et al. (2020) suggested restricting the power of the adversary. We show how this issue can be circumvented without making such (or other) assumptions in Appendix D.2.

6. Conclusion and Future Directions

In this work, we wished to study the ability to provide individual fairness guarantees, while attempting to minimize surrounding assumptions as much as possible — regarding the feedback structure, the data generation process, the availability of a fairness metric, the level of disagreement among the auditors, and the specific form of the auditors' preferences. Our work suggests a number of future directions. First, it would be desirable to determine whether efficiently achieving faster rates is possible. We note, however, that our proposed algorithm is the first to establish simultaneous guarantees for accuracy and fairness in the setting we study. Second, Context-Semi-Bandit-FTPL-WR is oracle-efficient, but is limited only to classes for which small separator sets are known. We inherit this limitation from the contextual bandit literature — it holds even without the additionally encoded fairness constraints. Third, our resampling-based variant 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 faster, more efficient algorithms which can provide multi-criteria accuracy and fairness guarantees of the sort we give here using one-sided feedback with auditors? 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?

7. Acknowledgements

YB is supported in part by the Israeli Council for Higher Education Postdoctoral Fellowship and the Apple Scholars in AI/ML PhD Fellowship. AR is supported in part by NSF grant FAI-2147212 and the Simons Collaboration on the Theory of Algorithmic Fairness.

References

Antos, A., Bartók, G., Pál, D., and Szepesvári, C. Toward a classification of finite partial-monitoring games. *Theor. Comput. Sci.*, 473:77–99, 2013. doi: 10.1016/j.tcs.2012.10.008. URL https://doi.org/10.1016/j.tcs.2012.10.008.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. Gambling in a rigged casino: The adversarial multi-arm bandit problem. In *36th Annual Symposium on Foundations of Computer Science, Milwaukee, Wisconsin, USA, 23-25 October 1995*, pp. 322–331. IEEE Computer Society, 1995. doi: 10.1109/SFCS.1995.492488. URL https://doi.org/10.1109/SFCS.1995.492488.

Bechavod, Y., Ligett, K., Roth, A., Waggoner, B., and Wu, Z. S. 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. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8972–8982, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/084afd913ab1e6ea58b8ca73f6cb41a6-Abstract.html.

Bechavod, Y., Jung, C., and Wu, Z. S. Metric-free individual fairness in online learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/80b618ebcac7aa97a6dac2ba65cb7e36-Abstract.html.

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

2012. URL http://proceedings.mlr.press/v23/bubeck12a/bubeck12a.pdf.

Cao, X. and Liu, K. J. R. Online convex optimization with time-varying constraints and bandit feedback. *IEEE Transactions on Automatic Control*, 64(7):2665–2680, 2019. doi: 10.1109/TAC.2018.2884653.

Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL http://www.cs.mcgill.ca/%7Ecolt2009/papers/024.pdf#page=1.

Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., and Warmuth, M. K. How to use expert advice. *J. ACM*, 44(3):427–485, May 1997. ISSN 0004-5411. doi: 10.1145/258128.258179. URL https://doi.org/10.1145/258128.258179.

Cesa-Bianchi, N., Gentile, C., and Zaniboni, L. Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research*, 7(44):1205–1230, 2006a. URL http://jmlr.org/papers/v7/cesa-bianchi06b.html.

Cesa-Bianchi, N., Lugosi, G., and Stoltz, G. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006b. doi: 10.1287/moor.1060.0206. URL https://doi.org/10.1287/moor.1060.0206.

Coston, A., Rambachan, A., and Chouldechova, A. Characterizing fairness over the set of good models under selective labels. In Meila, M. and Zhang, T. (eds.), *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*, pp. 2144–2155. PMLR, 2021. URL http://proceedings.mlr.press/v139/coston21a.html.

Dani, V., Hayes, T. P., and Kakade, S. M. The price of bandit information for online optimization. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T. (eds.), *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pp. 345–352. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/hash/bf62768ca46b6c3b5bea9515d1a1fc45-Abstract.html.

De-Arteaga, M., Dubrawski, A., and Chouldechova, A. Learning under selective labels in the presence of expert consistency. *CoRR*, abs/1807.00905, 2018. URL http://arxiv.org/abs/1807.00905.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In Goldwasser, S. (ed.), *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pp. 214–226. ACM, 2012. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Elzayn, H., Jabbari, S., Jung, C., Kearns, M., Neel, S., Roth, A., and Schutzman, Z. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 170–179, 2019.

Ensign, D., Sorelle, F., Scott, N., Carlos, S., and Suresh, V. Decision making with limited feedback. In Janoos, F., Mohri, M., and Sridharan, K. (eds.), *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pp. 359–367. PMLR, 07–09 Apr 2018. URL https://proceedings.mlr.press/v83/ensign18a.html.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504. URL https://doi.org/10.1006/jcss.1997.1504.

Gillen, S., Jung, C., Kearns, M. J., and Roth, A. 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.*, pp. 2605–2614, 2018. URL http://papers.nips.cc/paper/7526-online-learning-with-an-unknown-fairness-metric.

Gupta, S. and Kamble, V. Individual fairness in hindsight. In *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, pp. 805–806, 2019. doi: 10.1145/3328526.3329605. URL https://doi.org/10.1145/3328526.3329605.

György, A., Linder, T., Lugosi, G., and Ottucsák, G. The on-line shortest path problem under partial monitoring. *J. Mach. Learn. Res.*, 8:2369–2403, 2007. URL http://dl.acm.org/citation.cfm?id=1314575.

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

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Helmbold, D. P., Littlestone, N., and Long, P. M. Apple tasting. *Inf. Comput.*, 161(2):85–139, 2000. doi: 10.1006/inco.2000.2870. URL https://doi.org/10.1006/inco.2000.2870.

Ilvento, C. Metric learning for individual fairness. In Roth, A. (ed.), *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020, Harvard University, Cambridge, MA, USA (virtual conference)*, volume 156 of *LIPIcs*, pp. 2:1–2:11. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi: 10.4230/LIPIcs.FORC.2020.2. URL https://doi.org/10.4230/LIPIcs.FORC.2020.2.

Jenatton, R., Huang, J., and Archambeau, C. Adaptive algorithms for online convex optimization with long-term constraints. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 402–411, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/jenatton16.html.

Jiang, H., Jiang, Q., and Pacchiano, A. Learning the truth from only one side of the story. In Banerjee, A. and Fukumizu, K. (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2413–2421. PMLR, 2021. URL http://proceedings.mlr.press/v130/jiang21b.html.

Joseph, M., Kearns, M. J., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 325–333, 2016. URL http://papers.nips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.

Joseph, M., Kearns, M. J., Morgenstern, J., Neel, S., and Roth, A. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pp. 158–163, 2018. doi: 10.1145/3278721.3278764. URL https://doi.org/10.1145/3278721.3278764.

Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L., and Wu, Z. S. An algorithmic framework for fairness elicitation. In Ligett, K. and Gupta, S. (eds.), *2nd Symposium on Foundations of Responsible Computing, FORC 2021, June 9-11, 2021, Virtual Conference*, volume 192 of *LIPIcs*, pp. 2:1–2:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi: 10.4230/LIPIcs.FORC.2021.2. URL https://doi.org/10.4230/LIPIcs.FORC.2021.2.

Kalai, A. T. and Vempala, S. S. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.*, 71(3):291–307, 2005. doi: 10.1016/j.jcss.2004.10.016. URL https://doi.org/10.1016/j.jcss.2004.10.016.

Kim, M. P., Reingold, O., and Rothblum, G. N. 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*, pp. 4847–4857, 2018. URL http://papers.nips.cc/paper/7733-fairness-through-computationally-bounded-awareness.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 08 2017. ISSN 0033-5533. doi: 10.1093/qje/qjx032. URL https://doi.org/10.1093/qje/qjx032.

Lahoti, P., Gummadi, K. P., and Weikum, G. Operationalizing individual fairness with pairwise fair representations. *Proc. VLDB Endow.*, 13(4):506–518, 2019. doi: 10.14778/3372716.3372723. URL http://www.vldb.org/pvldb/vol13/p506-lahoti.pdf.

Lakkaraju, H. and Rudin, C. Learning cost-effective and interpretable treatment regimes. In Singh, A. and Zhu, X. J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 166–175. PMLR, 2017. URL http://proceedings.mlr.press/v54/lakkaraju17a.html.

Lakkaraju, H., Kleinberg, J. M., Leskovec, J., Ludwig, J., and Mullainathan, S. The selective labels problem: Evaluating algorithmic predictions in the presence

of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pp. 275–284. ACM, 2017. doi: 10.1145/3097983.3098066. URL https://doi.org/10.1145/3097983.3098066.

Levinson, J. D., Bennett, M. W., and Hioki, K. Judging implicit bias: A national empirical study of judicial stereotypes. *Law & Courts eJournal*, 2016.

Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, 1994. doi: 10.1006/inco.1994.1009. URL https://doi.org/10.1006/inco.1994.1009.

Mahdavi, M., Jin, R., and Yang, T. Trading regret for efficiency: Online convex optimization with long term constraints. *Journal of Machine Learning Research*, 13(81):2503–2528, 2012. URL http://jmlr.org/papers/v13/mahdavi12a.html.

Mukherjee, D., Yurochkin, M., Banerjee, M., and Sun, Y. Two simple ways to learn individual fairness metrics from data. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7097–7107. PMLR, 2020. URL http://proceedings.mlr.press/v119/mukherjee20a.html.

Neel, S., Roth, A., and Wu, Z. S. How to use heuristics for differential privacy. In Zuckerman, D. (ed.), *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pp. 72–93. IEEE Computer Society, 2019. doi: 10.1109/FOCS.2019.00014. URL https://doi.org/10.1109/FOCS.2019.00014.

Neu, G. and Bartók, G. An efficient algorithm for learning with semi-bandit feedback. In Jain, S., Munos, R., Stephan, F., and Zeugmann, T. (eds.), *Algorithmic Learning Theory - 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings*, volume 8139 of *Lecture Notes in Computer Science*, pp. 234–248. Springer, 2013. doi: 10.1007/978-3-642-40935-6_17. URL https://doi.org/10.1007/978-3-642-40935-6_17.

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., and Guthrie, C. Does unconscious racial bias affect trial judges. *Notre Dame Law Review*, 84:1195, 2009.

Rothblum, G. N. and Yona, G. Probably approximately metric-fair learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 5666–5674, 2018. URL http://proceedings.mlr.press/v80/yona18a.html.

Rothstein, R. *The color of law: A forgotten history of how our government segregated America*. Liveright Publishing Corporation, New York, NY, 2018.

Sculley, D. Practical learning from one-sided feedback. In Berkhin, P., Caruana, R., and Wu, X. (eds.), *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pp. 609–618. ACM, 2007. doi: 10.1145/1281192.1281258. URL https://doi.org/10.1145/1281192.1281258.

Syrgkanis, V., Krishnamurthy, A., and Schapire, R. E. 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*, pp. 2159–2168, 2016. URL http://proceedings.mlr.press/v48/syrgkanis16.html.

Vovk, V. G. Aggregating strategies. In Fulk, M. A. and Case, J. (eds.), *Proceedings of the Third Annual Workshop on Computational Learning Theory, COLT 1990, University of Rochester, Rochester, NY, USA, August 6-8, 1990*, pp. 371–386. Morgan Kaufmann, 1990. URL http://dl.acm.org/citation.cfm?id=92672.

Yu, H. and Neely, M. J. A low complexity algorithm with o(âˆšt) regret and o(1) constraint violations for online convex optimization with long term constraints. *Journal of Machine Learning Research*, 21(1):1–24, 2020. URL http://jmlr.org/papers/v21/16-494.html.

Yu, H., Neely, M., and Wei, X. Online convex optimization with stochastic constraints. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/da0d1111d2dc5d489242e60ebcbaf988-Paper.pdf.

Yurochkin, M., Bower, A., and Sun, Y. Training individually fair ML models with sensitive subspace robustness. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1gdkxHFDH.

Zeng, J., Ustun, B., and Rudin, C. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society Series A*, 180(3):689–722, June

2017. URL https://ideas.repec.org/a/bla/jorssa/v180y2017i3p689-722.html.

A. Extended Related Work

In the context of individual fairness, Joseph et al. (2016; 2018) study a notion of individual fairness in a more traditional contextual bandit setting, in which k individuals arrive at each round, and some subset of them are "selected", which yields observable reward to the learner. Their notion of individual fairness mandates that selection probability should be monotone in the (true) label of an individual (and in particular individuals with the same true label should be selected with the same probability). True labels cannot in general be ascertained, and as a result they only give positive results under strong realizability assumptions.

An additional line of work which is relevant in the context of online learning with individual fairness constraints is the one regarding online learning with long-term constraints (Mahdavi et al., 2012; Jenatton et al., 2016; Yu et al., 2017; Cao & Liu, 2019; Yu & Neely, 2020). We refer the reader to a detailed discussion of similarities and differences in Bechavod et al. (2020).

B. Omitted Details from Section 3

In order to prove Theorem 3.2, we first state and prove two lemmas, which express the Lagrangian regret in the setting of Individually Fair Online Learning with One-Sided Feedback (Algorithm 1) in terms of the regret in the contextual combinatorial semi-bandit setting (Algorithm 2). In what follows, we denote $k' = k + 2C$.

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

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

Proof. Observe that

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

Which proves the lemma. □

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

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

Proof. Observe that

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

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

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. We can see that

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

Which concludes the proof. □

C. Omitted Details from Section 4

C.1. From Panels to "Representative" Auditors

Here, we give a reduction from auditing by panels to auditing by instance-specific single auditors. In particular, we prove that the feedback given by any panel can be viewed as equivalent to the decisions of single, "representative", auditors from the panel, where the identity of the relevant auditor is determined only as a function of the specific pair (x, x') in question.

We observe that when it comes to a single pair (x, x'), we can order auditors by their "strictness" on this pair, as measured by $d(x, x')$. However it is not possible in general to order or compare the level of "strictness" of different auditors beyond a single pair, as some may be stricter than others on some pairs, but have the opposite relation on others. 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

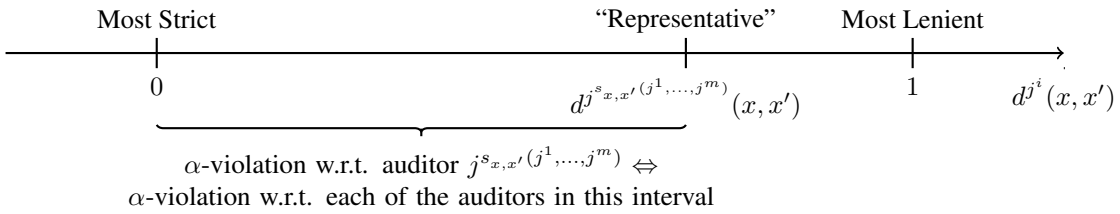

Figure 1. An illustration of an ordering of a panel of auditors $(j^1, \ldots, j^m)$ according to their (implicit) distances on (x, x'). $j^{s_{x,x'}(j^1,\ldots,j^m)}$ denotes the auditor who is in the $\lceil \gamma m \rceil$ position in this ordering, which can also be viewed as having the "swing vote"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

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". We formalize the argument in Lemma 4.1.

Lemma 4.1. *Let $(x, x') \in \mathcal{X}^2$, $(j^1, ..., j^m) \in \mathcal{J}^m$. Then, there exists an index $s = s_{x,x'}(j^1, \ldots, j^m) \in [m]$ such that 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 $\bar{j}^{\alpha,\gamma}_{j^1,\ldots,j^m}$.*
2. *π has an α-violation on (x, x') w.r.t. auditor j^s.*

The crucial aspect of this lemma is that the index $s_{x,x'}(j^1, \ldots, j^m)$ of the "pivotal" auditor is defined independently of π. We refer the reader to Figure 1 for an illustration of Lemma 4.1.

Proof of Lemma 4.1. Fix (x, x'). Then, we can define an ordering of $(j^1, ..., j^m)$ according to their (underlying) distances on (x, x'),

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

Then, set

$$s := s_{x,x'}(j^1, ..., j^m) = i_{\lceil \gamma m \rceil}. \tag{2}$$

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 on (x, x') with respect to panel $\bar{j}$, we know that

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

which is equivalent to indicating an α-violation on (x, x') with respect to auditor j^s. This concludes the proof. □

C.2. Multi-Criteria No Regret Guarantees

Proof of Lemma 4.3. To prove the lemma, it is sufficient to prove that for every $\pi^* \in Q_{\alpha-\epsilon,\gamma}$,

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

Which, using Definition 2.9, is equivalent to proving that

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

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 4.1 and Definition 4.2, we know that

$$\pi^t(\rho^{t,1}) - \pi^t(\rho^{t,2}) > d^{s_{\rho^{t,1},\rho^{t,2}}(j^{t,1},\ldots,j^{t,m})}(\rho^{t,1}, \rho^{t,2}) + \alpha \tag{4}$$

And

$$-(\pi^*(\rho^{t,1}) - \pi^*(\rho^{t,2})) \geq \epsilon - \alpha - d^{s_{\rho^{t,1},\rho^{t,2}}(j^{t,1},\ldots,j^{t,m})}(\rho^{t,1}, \rho^{t,2}) \tag{5}$$

Hence, combining Equation 4 and Equation 5, we get

$$\pi^t(\rho^{t,1}) - \pi^t(\rho^{t,2}) - (\pi^*(\rho^{t,1}) - \pi^*(\rho^{t,2}))$$
$$\geq d^{s_{\rho^{t,1},\rho^{t,2}}(j^{t,1},\ldots,j^{t,m})}(\rho^{t,1}, \rho^{t,2}) + \alpha + \epsilon - \alpha - d^{s_{\rho^{t,1},\rho^{t,2}}(j^{t,1},\ldots,j^{t,m})}(\rho^{t,1}, \rho^{t,2})$$
$$\geq \epsilon.$$

The lemma hence follows. □

D. Omitted Detailes from Section 5

In this section, we present two algorithms for the contextual combinatorial semi-bandit setting (Algorithm 2), and show how they can be leveraged to establish accuracy and fairness guarantees in the setting of individually fair online learning with one-sided feedback (Algorithm 1). 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.

D.1. Exp2

We begin by presenting the Exp2 algorithm (Bubeck et al., 2012; Dani et al., 2007; Cesa-Bianchi & Lugosi, 2009) and showing how it can be adapted to our setting.[8] Exp2 is an adaptation of the classical exponential weights algorithm (Littlestone & Warmuth, 1994; Auer et al., 1995; Vovk, 1990; Freund & Schapire, 1997; Cesa-Bianchi et al., 1997), which, 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 on $|\mathcal{H}|$.

[8] The contextual combinatorial semi-bandit setting considered in this paper subsumes the standard contextual k-armed bandit setting. To see this, consider the case where $A^t = A = \{a^{t,i} = (\mathbb{1}[i = 1], \ldots, \mathbb{1}[i = k]) : i \in [k]\}$. Naively applying the classical EXP4 algorithm for contextual bandits in the combinatorial semi-bandit setting would result in a regret bound of $O(\sqrt{|\mathcal{H}|T})$, whose square root dependence on $|\mathcal{H}|$ we prefer to avoid.

Theorem D.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:*

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

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

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

Proof of Theorem D.2. Combining Theorems 3.2, D.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 4.3, we get

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

And,

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

□

The guarantees of Theorem D.2 can be interpreted as follows: accuracy-wise, the resulting algorithm is competitive with the performance of the most accurate policy that is fair (i.e. in $Q_{\alpha-\epsilon,\gamma}$). Fairness-wise, the number of rounds in which there exist (one or more) fairness violations, is sub-linear.

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 & Vempala, 2005; Hannan, 1957).

D.2. Context-Semi-Bandit-FTPL

We next present an oracle-efficient algorithm, Context-Semi-Bandit-FTPL (Syrgkanis et al., 2016), and construct an variant of it for the setting of individually fair online learning with one-sided feedback. Broadly speaking, FTPL-style algorithms' approach is to solve "perturbed" optimization problems. Namely, at each round, the set of data samples observed so far is augmented, using carefully drawn additional noisy samples. Then, the resulting "perturbed" optimization problem over the augmented sample set is solved. In doing so, the procedure carefully combines the objectives of stability and error minimization, in order to provide no regret guarantees.

In order to construct an *efficient* implementation of this approach in the setting of contextual combinatorial semi-bandit, 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 can be viewed equivalently as assuming access to a weighted ERM oracle for $\mathcal{H}$. We next describe the small separator set assumption.

Definition 5.1. 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 $x \in S$ such that $h(x) \neq h'(x)$.

Remark 5.2. Classes for which small separator sets are known include conjunctions, disjunctions, parities, decision lists, discretized linear classifiers. Please see more elaborate discussions in Syrgkanis et al. (2016) and Neel et al. (2019).

For the following theorem, it is assumed that Context-Semi-Bandit-FTPL has access to a (pre-computed) separator set S of size s for the class $\mathcal{H}$, and access to an (offline) optimization oracle for $\mathcal{H}$.

Theorem D.3 (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:*

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

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 variant 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 accurate enough 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 point in Lemma D.5). Second, as it is observed in general (see, e.g. the discussion in Neu & Bartók (2013)), the specific weights this implicit distribution places on each of $h \in \mathcal{H}$ cannot be expressed in closed-form.

We therefore next construct an adaptation we term 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) underlying distribution π^t. This adaptation is summarized in Algorithm 4 and Algorithm 5 below, and yields the following guarantee.

Theorem 5.3. *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 guarantees, for any $\epsilon \in [0, \alpha]$, simultaneously:*

1. Accuracy:

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

2. Fairness:

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

We next describe the adaptation of Context-Semi-Bandit-FTPL (Syrgkanis et al., 2016) to our setting. Context-Semi-Bandit-FTPL relies on access to an optimization oracle for the corresponding (offline) problem. We elaborate on the exact implementation of this oracle in our setting next.

Definition D.4 (Optimization oracle)**.** Context-Semi-Bandit-FTPL assumes access to an 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 $f^t(a) = \langle a, \ell\rangle$, where ℓ is a non-negative vector. In our construction, this is equivalent to

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

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 the beginning of 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., Neu & 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 rely on having sampling access to actions from the distribution in obtaining expected no regret guarantees.

For our purposes, however, such a method of assessing the loss on realized (single) hypotheses $h^t \sim \pi^t$ 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.5. *There exist $\alpha, \gamma, m, k > 0$, $\mathcal{H} : \mathcal{X} \to \{0, 1\}$, $\bar{x} \in \mathcal{X}^k$, $\bar{j} : \mathcal{X}^k \to \mathcal{X}^2$, and $\pi \in \Delta\mathcal{H}$ for which, simultaneously,*

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

Proof of Lemma D.5. We set $\alpha = 0.2, \gamma = 1$ and $k = 2$. We define the context space to be $\mathcal{X} = \{x, x'\}$, 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 $\bar{j}^{\alpha,\gamma}$, that hence consists of a single auditor, to reflect the judgements of j^α, where $d^j(x, x') = 0.1$. Finally, we define $\pi \in \Delta\mathcal{H}$ to return h with probability 0.5, and h' with probability 0.5. We denote $\bar{x} = (x, x')$.

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,

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

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

On the other hand,

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

Hence,

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

Which proves the lemma. □

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, 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 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.

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.

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

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 $\hat{\pi}^t$ 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.6. *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 with $\bar{x}^t$, $\bar{y}^t$, $\hat{h}^t$, $\hat{\rho}^t$, and C), yields, with probability $1-\delta$, 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.6, 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.7. *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 $(\rho^t)_{t=1}^T$ such that $\forall t \in [T], \rho^t \in (\bar{x}^t \times \bar{x}^t) \cup \{(v, v)\}$,*

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

Proof. Using Chernoff bound, we can bound the difference in predictions between the underlying and the estimated distributions over base classifiers, for each of the contexts 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 $\bar{x}^t$, $\bar{y}^t$, $\hat{h}^t$, $\hat{\rho}^t$, and C), with probability $1-\delta$,

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

Summing over rounds, with probability $1-\delta$, for any arbitrary sequence of reported pairs $(\rho^t)_{t=1}^T$, such that $\forall t \in [T], \rho^t \in (\bar{x}^t \times \bar{x}^t) \cup \{(v,v)\}$:

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

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.6. Using Theorems D.3 and 3.2 along with the fact that $\|\ell^t\|_* \leq 2k+4C$, for any sequence $(\rho^t)_{t=1}^T$ such that $\forall t \in [T], \rho^t \in (\bar{x}^t \times \bar{x}^t) \cup \{(v,v)\}$,

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

Using Lemma D.7 and the triangle inequality, we conclude that, with probability $1-\delta$,

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

□

We are now ready to prove Theorem 5.3

Proof of Theorem 5.3. Using Theorem D.6 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 4.3, we get, with probability $1-\delta$,

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

And,

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

□