Fairness of Exposure in Stochastic Bandits

Lequn Wang [1]   Yiwei Bai [1]   Wen Sun [1]   Thorsten Joachims [1]

Abstract

Contextual bandit algorithms have become widely used for recommendation in online systems (e.g. marketplaces, music streaming, news), where they now wield substantial influence on which items get exposed to the users. This raises questions of fairness to the items — and to the sellers, artists, and writers that benefit from this exposure. We argue that the conventional bandit formulation can lead to an undesirable and unfair winner-takes-all allocation of exposure. To remedy this problem, we propose a new bandit objective that guarantees merit-based fairness of exposure to the items while optimizing utility to the users. We formulate fairness regret and reward regret in this setting, and present algorithms for both stochastic multi-armed bandits and stochastic linear bandits. We prove that the algorithms achieve sub-linear fairness regret and reward regret. Beyond the theoretical analysis, we also provide empirical evidence that these algorithms can fairly allocate exposure to different arms effectively.

1. Introduction

Bandit algorithms (Thompson, 1933; Robbins, 1952; Bubeck & Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore & Szepesvári, 2020) provide an attractive model of learning for online platforms, and they are now widely used to optimize retail, media streaming, and news-feed. Each round of bandit learning corresponds to an interaction with a user, where the algorithm selects an arm (e.g. product, song, article), observes the user's response (e.g. purchase, stream, read), and then updates its policy. Over time, the bandit algorithm thus learns to maximize the user responses, which are often well aligned with the objective of the online platform (e.g. profit maximization, engagement maximization).

While maximizing user responses may arguably be in the interest of the platform and its users at least in the short term, there is now a growing understanding that it can also be problematic in multiple respects. In this paper, we focus on the fact that this objective ignores the interests of the items (i.e. arms), which also derive utility from the interactions. In particular, sellers, artists and writers have a strong interest in the exposure their items receive, as it affects their chance to get purchased, streamed or read. It is well understood that algorithms that maximize user responses can be unfair in how they allocate exposure to the items (Singh & Joachims, 2018). For example, two items with very similar merit (e.g. click-through rate) may receive substantially different amounts of exposure — which is not only objectionable in itself, but can also degrade the long-term objectives of the platform (e.g. seller retention (Mehrotra et al., 2018), anti-discrimination (Noble, 2018), anti-polarization (Epstein & Robertson, 2015)).

To illustrate the problem, consider a conventional (non-personalized) stochastic multi-armed bandit algorithm that is used to promote new music albums on the front-page of a website. The bandit algorithm will quickly learn which album draws the largest click-through rate and keep displaying this album, even if other albums are almost equally good. This promotes a winner-takes-all dynamic that creates superstars (Mehrotra et al., 2018), and may drive many deserving artists out of business. Analogously, a (personalized) contextual bandit for news-feed recommendation can polarize a user by quickly learning which type of articles the user is most likely to read, and then exclusively recommend such articles instead of a portfolio that is more reflective of the user's true interest distribution.

To overcome these problems of the conventional bandit objective, we propose a new formulation of the bandit problem that implements the principle of Merit-based Fairness of Exposure (Singh & Joachims, 2018; Biega et al., 2018). For brevity, we call this the FairX bandit problem. It incorporates the additional fairness requirement that each item/arm receives a share of exposure that is proportional to its merit. We define the merit of an arm as an increasing function of its mean reward, and the exposure as the probability of being selected by the bandit policy at each round. Based on these quantities, we then formulate the reward regret and the fairness regret so that minimizing these two regrets corresponds

[1]Department of Computer Science, Cornell University, Ithaca, NY, USA. Correspondence to: Lequn Wang <lw633@cornell.edu>, Wen Sun <ws455@cornell.edu>, Thorsten Joachims <tj@cs.cornell.edu>.

to maximizing responses while minimizing unfairness to the items.

For the FairX bandit problem, we present a fair upper confidence bound (UCB) algorithm and a fair Thompson sampling (TS) algorithm in the stochastic multi-armed bandits (MAB) setting, as well as a fair linear UCB algorithm and a fair linear TS algorithm in the stochastic linear bandits setting. We prove that all algorithms achieve fairness regret and reward regret with sub-linear dependence on the number of rounds, while the TS-based algorithms have computational advantages. The fairness regret of these algorithms also depends on the minimum merit of the arms and a bounded Lipschitz constant of the merit function, and we provide fairness regret lower bounds based on these quantities. Beyond the theoretical analysis, we also conduct an empirical evaluation that compares these algorithms with conventional bandit algorithms and more naive baselines, finding that the fairness-aware algorithms can fairly allocate exposure to different arms effectively while maximizing user responses.

2. Related Work

The bandit problem was first introduced by Thompson (Thompson, 1933) to efficiently conduct medical trials. Since then, it has been extensively studied in different variants, and we refer to these books (Bubeck & Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore & Szepesvári, 2020) for a comprehensive survey. We focus on the classic stochastic MAB setting where each arm has a fixed but unknown reward distribution, as well as the stochastic linear bandits problem where each arm is represented as a d-dimensional vector and its expected reward is a linear function of its vector representation. In both stochastic MAB and stochastic linear bandits, some of the algorithms we designed leverage the idea of optimism in the face of uncertainty behind the UCB algorithm (Lai & Robbins, 1985), while other algorithms leverage the idea of posterior sampling behind the TS (Thompson, 1933) algorithm. The theoretical results of the proposed fair UCB and fair linear UCB algorithm borrow some ideas from several prior finite time analysis works on the conventional UCB and linear UCB algorithm (Auer, 2002; Dani et al., 2008; Abbasi-Yadkori et al., 2011). We adopt the Bayesian regret framework (Russo & Van Roy, 2014) for our theoretical analysis of the fair TS and the fair linear TS algorithm.

Algorithmic fairness has been extensively studied in binary classification (Hardt et al., 2016; Chouldechova, 2017; Kleinberg et al., 2017; Agarwal et al., 2018). These works propose statistical criteria to test algorithmic fairness that often operationalize definitions of fairness from political philosophy and sociology. Several prior works (Blum et al., 2018; Blum & Lykouris, 2019; Bechavod et al., 2019) study how to achieve these fairness criteria in online learning.

These algorithms achieve fairness to the incoming users. We, in contrast, achieve fairness to the arms.

Joseph et al. (Joseph et al., 2016b;a; 2018) study fairness in bandits that ensure a better arm is always selected with no less probability than a worse arm. Different from our definition of fairness, their optimal policy is still the one that deterministically selects the arm with the largest expected reward while giving zero exposure to all the other arms. Another type of fairness definition in bandits is to ensure a minimum and/or maximum amount of exposure to each arm or group of arms (Heidari & Krause, 2018; Wen et al., 2019; Schumann et al., 2019; Li et al., 2019; Celis et al., 2018; Claure et al., 2020; Patil et al., 2020; Chen et al., 2020). However, they do not take the merit of the items into consideration. Gillen et al. (Gillen et al., 2018) propose to optimize individual fairness defined in (Dwork et al., 2012) in the adversarial linear bandits setting, where the difference between the probabilities that any two arms are selected is bounded by the distance between their context vectors. They require additional feedback of fairness-constraint violations. We work in the stochastic bandits setting and we do not require any additional feedback beyond the reward. We also ensure that similar items obtain similar exposure, but we focus on similarity of merit, which corresponds to closeness in mean reward conditioned on context.

The most relevant work may arguably be (Liu et al., 2017), which considers fairness in stochastic MAB problems where the reward distribution is Bernoulli. They aim to achieve calibrated fairness where each arm is selected with the probability equal to that of its reward being the largest, while satisfying a smoothness constraint where arms with similar merit should receive similar exposure. They propose a TS-based algorithm that achieves fairness regret with a $T^{2/3}$ dependence on the time horizon T. Our formulation is more general in a sense that we consider arbitrary reward distributions and merit functions, with their formulation as a special case. What is more, our proposed algorithms achieve fairness regret with a $\sqrt{T}$ dependence on T. In addition, we further study the more general setting of stochastic linear bandits.

Our definition of fairness has connections to the fair division problem (Steihaus, 1948; Brams & Taylor, 1996; Procaccia, 2013), where the goal is to allocate a resource to different agents in a fair way. In our problem, we aim to allocate the users' attention among the items in a fair way. Our definition of fairness ensures proportionality, one of the key desiderata in the fair division literature to ensure each agent receives its fair share of the resource. Recently, merit-based fairness of exposure has been studied for ranking problems in the statistical batch learning framework (Singh & Joachims, 2018; 2019). We build upon this work, and extend merit-based fairness of exposure to the online-learning setting.

3. Stochastic Multi-Armed Bandits in the FairX Setting

We begin by introducing the FairX setting for stochastic MAB, including our new formulation of fairness and reward regret. We then develop two algorithms, called FairX-UCB and FairX-TS, and bound their fairness and reward regret. In the subsequent section, we will extend this approach to stochastic linear bandits.

3.1. FairX Setting for Stochastic MAB

A stochastic MAB instance can be represented as a collection of reward distributions $v = (P_a : a \in [K])$, where P_a is the reward distribution of arm a with mean $\mu_a^\star = \mathbb{E}_{r \sim P_a}[r]$. The learner interacts with the environment sequentially over T rounds. In each round $t \in [T]$, the learner has to choose a policy π_t over the K arms based on the interaction history before round t. The learner then samples an arm $a_t \sim \pi_t$. In response to the selected arm a_t, the environment samples a reward $r_{t,a_t} \sim P_{a_t} \in \mathbb{R}$ from the reward distribution P_{a_t} and reveals the reward r_{t,a_t} to the learner. The history $\mathcal{H}_t = (\pi_1, a_1, r_{1,a_1}, \ldots, \pi_{t-1}, a_{t-1}, r_{t-1,a_{t-1}})$ consists of all the deployed policies, chosen arms, and their associated rewards. Conventionally, the goal of learning is to maximize the cumulative expected reward $\sum_{t=1}^{T} \mathbb{E}_{a_t \sim \pi_t} \mu_{a_t}^\star$. Thus conventional bandit algorithms converge to a policy that deterministically selects the arm with the largest expected reward.

As many have pointed out in other contexts (Singh & Joachims, 2018; Mehrotra et al., 2018; Biega et al., 2018; Beutel et al., 2019; Geyik et al., 2019; Abdollahpouri et al., 2020), such winner-takes-all allocations can be considered unfair to the items in many applications and can lead to undesirable long-term dynamics. Bringing this insight to the task of bandit learning, we propose to incorporate merit-based fairness-of-exposure constraints (Singh & Joachims, 2018) into the bandits objective. Specifically, we aim to learn a policy $\pi^\star$ which ensures that each arm receives an amount of exposure proportional to its merit, where merit is quantified through an application-dependent merit function $f(\cdot) > 0$ that maps the expected reward of an arm to a positive merit value.

$$\frac{\pi^\star(a)}{f(\mu_a^\star)} = \frac{\pi^\star(a')}{f(\mu_{a'}^\star)} \quad \forall a, a' \in [K].$$

The merit function f is an input to the bandit algorithm, and it provides a design choice that permits tailoring the fairness criterion to different applications. The following theorem shows that there is a unique policy that satisfies the above fairness constraints.

Theorem 3.1.1 (Optimal Fair Policy). *For any mean reward parameter $\mu^\star$ and any choice of merit function $f(\cdot) > 0$,*

there exist a unique policy $\pi^\star$ of the form

$$\pi^\star(a) = \frac{f(\mu_a^\star)}{\sum_{a'} f(\mu_{a'}^\star)} \quad \forall a \in [K],$$

that fulfills the merit-based fairness of exposure constraints.

We refer to $\pi^\star$ as the optimal fair policy. All the proofs of the theorems are in Appendix A.

When the bandit converges to this optimal fair policy $\pi^\star$, the expected reward also converges to the expected reward of the optimal fair policy. We thus define the *reward regret* RR_T at round T as the gap between the expected reward of the deployed policy and the expected reward of the optimal fair policy $\pi^\star$

$$\text{RR}_T = \sum_{t=1}^{T} \sum_a \pi^\star(a)\mu_a^\star - \sum_{t=1}^{T} \sum_a \pi_t(a)\mu_a^\star. \quad (1)$$

While this reward regret quantifies how quickly the reward is optimized, we also need to quantify how effectively the algorithm learns to enforce fairness. We thus define the following *fairness regret* FR_T, which measures the cumulative ℓ^1 distance between the deployed policy and the optimal fair policy at round T

$$\text{FR}_T = \sum_{t=1}^{T} \sum_a |\pi^\star(a) - \pi_t(a)|. \quad (2)$$

The fairness regret and the reward regret depend on both the randomly sampled rewards, as well as the arms randomly sampled from the policy. They are thus random variables and we aim to minimize the regrets with high probability.

To prepare for the theoretical analysis, we introduce the following two conditions on the merit function f to suitably characterize a FairX bandit problem.

Condition 3.1.2 (Minimum Merit). *The merit of each arm is positive, i.e. $\min_\mu f(\mu) \geq \gamma$ for some positive constant $\gamma > 0$.*

Condition 3.1.3 (Lipschitz Continuity). *The merit function f is L-Lipschitz continuous, i.e. $\forall \mu_1, \mu_2, |f(\mu_1) - f(\mu_2)| \leq L|\mu_1 - \mu_2|$ for some positive constant $L > 0$.*

The following two theorems show that neither of the two conditions can be dropped, if we want to obtain bandit algorithms with fairness regret that is sub-linear in the number of rounds T.

Theorem 3.1.4 (Lower Bound on Fairness Regret is Linear without Minimum-Merit Condition). *For time horizon $T > 0$, there exists a 1-Lipschitz continuous merit function f where $\min_\mu f(\mu) = 1/\sqrt{T}$, such that for any bandit algorithm, there must exist a MAB instance such that the expected fairness regret is at least $\mathbb{E}[FR_T] \geq 0.015T$.*

Theorem 3.1.5 (Lower Bound on Fairness Regret is Linear without Bounded Lipschitz-Continuity Condition). *For time horizon $T > 0$, there exists a $\sqrt{T}$-Lipschitz continuous merit function f with minimum merit 1, such that for any bandit algorithm, there must exist a MAB instance such that the expected fairness regret is at least $\mathbb{E}[FR_T] \geq 0.015T$.*

3.2. FairX-UCB Algorithm

Algorithm 1 FairX-UCB Algorithm

1: **input:** K, T, f, w_0
2: **for** $t = 1$ to T **do**
3:     $\forall a \; N_{t,a} = \sum_{\tau=1}^{t-1} \mathbf{1}\{a_\tau = a\}$
4:     $\forall a \; \hat{\mu}_{t,a} = \sum_{\tau=1}^{t-1} \mathbf{1}\{a_\tau = a\} r_{\tau,a_\tau}/N_{t,a}$
5:     $\forall a \; w_{t,a} = w_0/\sqrt{N_{t,a}}$
6:     $\mathrm{CR}_t = (\mu : \forall a \; \mu_a \in [\hat{\mu}_{t,a} - w_{t,a}, \hat{\mu}_{t,a} + w_{t,a}])$
7:     $\mu_t = \arg\max_{\mu \in \mathrm{CR}_t} \sum_a \frac{f(\mu_a)}{\sum_{a'} f(\mu_{a'})} \mu_a$
8:     Construct policy $\pi_t(a) = \frac{f(\mu_{t,a})}{\sum_{a'} f(\mu_{t,a'})}$
9:     Sample arm $a_t \sim \pi_t$
10:     Observe reward r_{t,a_t}
11: **end for**

The first algorithm we introduce is called FairX-UCB and it is detailed in Algorithm 1. It utilizes the idea of optimism in the face of uncertainty. At each round t, the algorithm constructs a confidence region CR_t which contains the true parameter $\mu^\star$ with high probability. Then the algorithm optimistically selects a parameter $\mu_t \in \mathbb{R}^K$ within the confidence region CR_t that maximizes the estimated expected reward subject to the constraint that we construct a fair policy as if the selected parameter is the true parameter. Compared to the conventional UCB algorithm which selects the arm with the largest UCB deterministicly in each round, the proposed FairX-UCB algorithm selects arms stochastically to ensure fairness. Finally, we apply the constructed policy π_t, observe the feedback, and update the confidence region.

The following two theorems characterize the fairness and reward regret upper bounds of the FairX-UCB algorithm.

Theorem 3.2.1 (FairX-UCB Fairness Regret). *Under Condition 3.1.2 and 3.1.3, suppose $\forall t, a : r_{t,a} \in [-1, 1]$, when $T > K$, for any $\delta \in (0, 1)$, set $w_0 = \sqrt{2 \ln(4TK/\delta)}$, the fairness regret of the FairX-UCB algorithm is $FR_T = \widetilde{O}\left(L\sqrt{KT}/\gamma\right)$ with probability at least $1 - \delta$.*

Theorem 3.2.2 (FairX-UCB Reward Regret). *Suppose $\forall t, a : r_{t,a} \in [-1, 1]$, when $T > K$, for any $\delta \in (0, 1)$, set $w_0 = \sqrt{2 \ln(4TK/\delta)}$, the reward regret of the FairX-UCB algorithm is $RR_T = \widetilde{O}\left(\sqrt{KT}\right)$ with probability at least $1 - \delta$.*

$\widetilde{O}$ ignores logarithmic factors in O. Note that the well-

---
Algorithm 2 FairX-TS Algorithm

1: **input:** $f, \mathcal{V}_1$
2: **for** $t = 1$ to ∞ **do**
3:     Sample parameter from posterior $\mu_t \sim \mathcal{V}_t$
4:     Construct policy $\pi_t(a) = \frac{f(\mu_{t,a})}{\sum_{a'} f(\mu_{t,a'})}$
5:     Sample arm $a_t \sim \pi_t$
6:     Observe reward r_{t,a_t}
7:     Update posterior $\mathcal{V}_{t+1} = \mathrm{Update}(\mathcal{V}_1, \mathcal{H}_{t+1})$
8: **end for**
---

known $\Omega\left(\sqrt{KT}\right)$ reward regret lower bound (Auer et al., 2002) developed for the conventional bandit problem also holds for the FairX setting because the conventional stochastic MAB problem that only minimizes the reward regret is a special case of the FairX setting where we set the merit function f to be an infinitely steep increasing function. Since the reward regret upper bound of FairX-UCB we proved does not depend on Conditions 3.1.2 and 3.1.3 about the merit function f, our reward regret upper bound of the FairX-UCB algorithm is tight up to logarithmic factors.

The fairness regret has the same dependence on the number of arms K and the number of rounds T as the reward regret. It further depends on the minimum merit constant γ and the Lipschitz continuity constant L, which we treat as absolute constants due to Theorem 3.1.4 and Theorem 3.1.5.

Compared to Fair_SD_TS algorithms proposed in (Liu et al., 2017), our proposed FairX-UCB algorithm focuses on fairness and reward regret across rounds instead of achieving a smooth fairness constraint for each round. This allows FairX-UCB to achieve improved fairness and reward regret ($\sqrt{KT}$ compared to $(KT)^{2/3}$). In addition, FairX-UCB works for general reward distributions and merit functions while SD_TS only works for Bernoulli reward distribution and identity merit function.

One challenge in implementing Algorithm 1 lies in Step 7, since finding the most optimistic parameter is a non-convex constrained optimization problem. We solve this optimization problem approximately with projected gradient descent in our empirical evaluation. In the next subsection, we will introduce the FairX-TS algorithm that avoids this optimization problem.

3.3. FairX-TS Algorithm

Another approach to designing stochastic bandit algorithms that has proven successful both empirically and theoretically is Thompson Sampling (TS). We find that this approach can also be applied to the FairX setting. In particular, our FairX-TS as shown in Algorithm 2 uses posterior sampling similar to a conventional TS bandit. The algorithm puts a prior distribution $\mathcal{V}_1$ on the expected reward of each arm

$\mu^\star$. For each round t, the algorithm samples a parameter μ_t from the posterior $\mathcal{V}_t$, and constructs a fair policy π_t from the sampled parameter to deploy. Finally, the algorithm observes the feedback and updates the posterior distribution of the true parameter.

Following (Russo & Van Roy, 2014), we analyze the Bayesian reward and fairness regret of the algorithm. The Bayesian regret framework assumes that the true parameter $\mu^\star$ is sampled from the prior, and the Bayesian regret is the expected regret taken over the prior distribution

$$\text{BayesRR}_T = \mathbb{E}_{\mu^\star} \left[\mathbb{E}[\text{RR}_T | \mu^\star] \right] \qquad (3)$$

$$\text{BayesFR}_T = \mathbb{E}_{\mu^\star} \left[\mathbb{E}[\text{FR}_T | \mu^\star] \right]. \qquad (4)$$

In the following two theorems we provide bounds on both the Bayesian reward regret and the Bayesian fairness regret of the FairX-TS algorithm.

Theorem 3.3.1 (FairX-TS Fairness Regret). *Under Condition 3.1.2 and 3.1.3, suppose the mean reward $\mu_a^\star$ of each arm a is independently sampled from standard normal distribution $\mathcal{N}(0, 1)$, and $\forall t, a\ r_{t,a} \sim \mathcal{N}(\mu_a^\star, 1)$, the Bayesian fairness regret of the FairX-TS algorithm at any round T is BayesFR$_T = \widetilde{O}\left(L\sqrt{KT}/\gamma \right)$.*

Theorem 3.3.2 (FairX-TS Reward Regret). *Suppose the mean reward $\mu_a^\star$ of each arm a is independently sampled from standard normal distribution $\mathcal{N}(0, 1)$, and $\forall t, a\ r_{t,a} \sim \mathcal{N}(\mu_a^\star, 1)$, the Bayesian fairness regret of the FairX-TS algorithm at any round T is BayesRR$_T = \widetilde{O}\left(\sqrt{KT} \right)$.*

Note that these regret bounds are on the same order as the fairness and reward regret of the FairX-UCB algorithm. However, FairX-TS relies on sampling from the posterior and thus avoids the non-convex optimization problem that makes the use of FairX-UCB more challenging.

4. Stochastic Linear Bandits in the FairX Setting

In this section, we extend the two algorithms introduced in the MAB setting to the more general stochastic linear bandits setting where the learner is provided with contextual information for making decisions. We discuss how the two algorithms can be adapted to this setting to achieve both sub-linear fairness and reward regret.

4.1. FairX Setting for Stochastic Linear Bandits

In stochastic linear bandits, each arm a at round t comes with a context vector $x_{t,a} \in \mathbb{R}^d$. A stochastic linear bandits instance $v = (P_x : x \in \mathbb{R}^d)$ is a collection of reward distributions for each context vector. The key assumption of stochastic linear bandits is that there exists a true parameter $\mu^\star$ such that, regardless of the interaction history $\mathcal{H}_t$, the

mean reward of arm a at round t is the product between the context vector and the true parameter $\mathbb{E}_{r \sim P_{x_{t,a}}}[r | \mathcal{H}_t] = \mu^\star \cdot x_{t,a}$ for all t, a. The noise sequence

$$\eta_t = r_{t,a_t} - \mu^\star \cdot x_{t,a_t}$$

is thus a martingale difference sequence, since

$$\mathbb{E}[\eta_t | \mathcal{H}_t] = \mathbb{E}_{a \sim \pi_t}[\mathbb{E}_{r \sim P_{x_{t,a}}}[r | \mathcal{H}_t] - \mu^\star \cdot x_{t,a}] = 0.$$

At each round t, the learner is given a set of context vectors $\mathcal{D}_t \subset \mathbb{R}^d$ representing the arms, and it has to choose a policy π_t over these K arms based on the interaction history $\mathcal{H}_t = (\mathcal{D}_1, \pi_1, a_1, r_{1,a_1}, \ldots, \mathcal{D}_{t-1}, \pi_{t-1}, a_{t-1}, r_{t-1,a_{t-1}})$. We focus on problems where the number of available arms is finite $\forall t : |\mathcal{D}_t| = K$, but where K could be large.

Again, we want to ensure that the policy provides each arm with an amount of exposure proportional to its merit

$$\frac{\pi_t^\star(a)}{f(\mu^\star \cdot x_{t,a})} = \frac{\pi_t^\star(a')}{f(\mu^\star \cdot x_{t,a'})} \quad \forall t, x_{t,a}, x_{t,a'} \in \mathcal{D}_t,$$

where f is the merit function that maps the mean reward of the arm to a positive merit value. Since the set of arms changes over time, the optimal fair policy $\pi_t^\star$ at round t is time-dependent

$$\pi_t^\star(a) = \frac{f(\mu^\star \cdot x_{t,a})}{\sum_{a'} f(\mu^\star \cdot x_{t,a'})} \quad \forall t, a.$$

Analogous to the MAB setting, we define the reward regret as the expected reward difference between the optimal fair policy and the deployed policy

$$\text{RR}_T = \sum_{t=1}^{T} \sum_a \pi_t^\star(a) \mu^\star \cdot x_{t,a} - \sum_{t=1}^{T} \sum_a \pi_t(a) \mu^\star \cdot x_{t,a}, \quad (5)$$

and fairness regret as the cumulative ℓ^1 distance between the optimal fair policy and the deployed policy

$$\text{FR}_T = \sum_{t=1}^{T} \sum_a |\pi_t^\star(a) - \pi_t(a)|. \qquad (6)$$

The lower bounds on the fairness regret derived in Theorem 3.1.4 and Theorem 3.1.5 in the MAB setting also apply to the stochastic linear bandit setting, since we can easily convert a MAB instance into a stochastic linear bandits instance by constructing K K-dimensional basis vectors, each representing one arm. Thus we again employ Condition 3.1.2 and 3.1.3 to design algorithms that have fairness regret with sub-linear dependence on the horizon T.

4.2. FairX-LinUCB Algorithm

Similar to the FairX-UCB algorithm, the FairX-LinUCB algorithm constructs a confidence region CR_t of the true

---

Algorithm 3 FairX-LinUCB Algorithm

1: **input:** β_t, f, λ
2: **initialization:** $\Sigma_1 = \lambda \mathbf{I}_d, b_1 = \mathbf{0}_d$
3: **for** $t = 1$ to ∞ **do**
4:      Observe contexts $\mathcal{D}_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,K})$
5:      $\hat{\mu}_t = \Sigma_t^{-1} b_t$ {The ridge regression solution}
6:      $\mathrm{CR}_t = (\mu : \|\mu - \hat{\mu}_t\|_{\Sigma_t} \leq \sqrt{\beta_t})$
7:      $\mu_t = \arg\max_{\mu \in \mathrm{CR}_t} \sum_a \frac{f(\mu \cdot x_{t,a})}{\sum_{a'} f(\mu \cdot x_{t,a'})} \mu \cdot x_{t,a}$
8:      Construct policy $\pi_t(a) = \frac{f(\mu_t \cdot x_{t,a})}{\sum_{a'} f(\mu_t \cdot x_{t,a'})}$
9:      Sample arm $a_t \sim \pi_t$
10:     Observe reward r_{t,a_t}
11:     $\Sigma_{t+1} = \Sigma_t + x_{t,a_t} x_{t,a_t}^\top$
12:     $b_{t+1} = b_t + x_{t,a_t} r_{t,a_t}$
13: **end for**

Algorithm 4 FairX-LinTS Algorithm

1: **input:** $f, \mathcal{V}_1$
2: **for** $t = 1$ to ∞ **do**
3:      Observe contexts $\mathcal{D}_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,K})$
4:      Sample parameter from posterior $\mu_t \sim \mathcal{V}_t$
5:      Construct policy $\pi_t(a) = \frac{f(\mu_{t,a} \cdot x_{t,a})}{\sum_{a'} f(\mu_{t,a'} \cdot x_{t,a})}$
6:      Sample arm $a_t \sim \pi_t$
7:      Observe reward r_{t,a_t}
8:      Update posterior $\mathcal{V}_{t+1} = \mathrm{Update}(\mathcal{V}_1, \mathcal{H}_{t+1})$
9: **end for**

---

find approximate solutions in our empirical evaluation.

4.3. FairX-LinTS Algorithm

To avoid the difficult optimization problem of FairX-LinUCB, we again explore the use of Thompson sampling. Algorithm 4 shows our proposed FairX-LinTS. At each round t, the algorithm samples a parameter μ_t from the posterior distribution $\mathcal{V}_t$ of the true parameter $\mu^\star$ and derives a fair policy π_t from the sampled parameter. Then the algorithm deploys the policy and observes the feedback for the selected arm. Finally, the algorithm updates the posterior distribution of the true parameter given the observed data. Note that sampling from the posterior is efficient for a variety of models (e.g. normal distribution), as opposed to the non-convex optimization problem in FairX-LinUCB.

Appropriately extending our definition of Bayesian reward regret and fairness regret

$$\mathrm{BayesRR}_T = \mathbb{E}_{\mu^\star}\left[\mathbb{E}[\mathrm{RR}_T | \mu^\star]\right] \qquad (7)$$

$$\mathrm{BayesFR}_T = \mathbb{E}_{\mu^\star}\left[\mathbb{E}[\mathrm{FR}_T | \mu^\star]\right], \qquad (8)$$

we can prove the following regret bounds for the FairX-LinTS algorithm.

Theorem 4.3.1 (FairX-LinTS Fairness Regret). *Under Condition 3.1.2 and 3.1.3, suppose each dimension of the true parameter $\mu^\star$ is independently sampled from standard normal distribution $\mathcal{N}(0,1)$, $\forall t, a \|x_{t,a}\|_2 \leq 1$, η_t is sampled from standard normal distribution $\mathcal{N}(0,1)$, the Bayesian fairness regret of the FairX-LinTS algorithm is $\mathrm{BayesFR} = \widetilde{O}\left(L\sqrt{dT}/\gamma\right)$.*

Theorem 4.3.2 (FairX-LinTS Reward Regret). *Suppose each dimension of the true parameter $\mu^\star$ is independently sampled from standard normal distribution $\mathcal{N}(0,1)$, $\forall t, a \|x_{t,a}\|_2 \leq 1$, η_t is sampled from standard normal distribution $\mathcal{N}(0,1)$, the Bayesian reward regret of the FairX-LinTS algorithm is $\mathrm{BayesRR} = \widetilde{O}\left(d\sqrt{T}\right)$.*

Similar to the FairX-TS algorithm in the MAB setting, the Bayesian fairness regret of FairX-LinTS assumes a normal prior. Note that the Bayesian fairness regret of FairX-LinTS

---

parameter $\mu^\star$ at each round t. The center of the confidence region $\hat{\mu}_t$ is the solution of a ridge regression over the existing data, which can be updated incrementally. The radius of the confidence ball β_t is an input to the algorithm. The algorithm proceeds by repeatedly selecting a parameter μ_t that is optimistic about the expected reward within the confidence region, subject to the constraint that we construct a fair policy from the parameter. We prove the following upper bounds on the fairness regret and reward regret of the FairX-LinUCB algorithm.

Theorem 4.2.1 (FairX-LinUCB Fairness Regret). *Under Condition 3.1.2 and 3.1.3, suppose $\forall t, a \|x_{t,a}\|_2 \leq 1$, η_t is 1 sub-Gaussian, $\|\mu^\star\|_2 \leq 1$, set $\lambda = 1$, with proper choice of β_t, the fairness regret at any round $T > 0$ is $FR_T = \widetilde{O}\left(Ld\sqrt{T}/\gamma\right)$ with high probability.*

Theorem 4.2.2 (FairX-LinUCB Reward Regret). *Suppose $\forall t, a \|x_{t,a}\|_2 \leq 1$, η_t is 1 sub-Gaussian, $\|\mu^\star\|_2 \leq 1$, set $\lambda = 1$, with proper choice of β_t, the reward regret at any round $T > 0$ is $RR_T = \widetilde{O}\left(d\sqrt{T}\right)$ with high probability.*

Both fairness and reward regret have square root dependence on the horizon T and a linear dependence on the feature dimension d, and the fairness regret depends on the absolute constants L and γ. Note that the reward regret is not tight in terms of d and there exist algorithms (Chu et al., 2011; Lattimore & Szepesvári, 2020) that achieve reward regret $\widetilde{O}(\sqrt{dT})$. However these algorithms are based on the idea of arm elimination and thus will likely not achieve low fairness regret. Also LinUCB is a much more practical option than the ones based on arm elimination (Chu et al., 2011).

The optimization Step 7 in Algorithm 3, where we need to find an optimistic parameter μ_t that maximizes the estimated expected reward within the confidence region CR_t subject to the fairness constraint, is again a non-convex constrained optimization problem. We use projected gradient descent to

differs by order of $\sqrt{d}$ from the non-Bayesian fairness regret of the FairX-LinUCB algorithm. The Bayesian setting and the normal prior assumption enable us to explicitly bound the total variation distance between our policy and the optimal fair policy, which allows us to avoid going through the UCB-based analysis of the LinUCB algorithms as in the conventional way of proving Bayesian regret bound (Russo & Van Roy, 2014).

5. Experiments

While the theoretical analysis provides worst case guarantees for the algorithms, we now evaluate empirically how the algorithms perform on a range of tasks. We perform this evaluation both on synthetic and real-world data. The synthetic data allows us to control properties of the learning problem for internal validity, and the real-world data provides a data-point for external validity of the analysis.

5.1. Experiment Setup

For the experiments where we control the properties of the synthetic data, we derive bandit problems from the multi-label datasets *yeast* (Horton & Nakai, 1996) and *mediamill* (Snoek et al., 2006). The yeast dataset consists of $2,417$ examples. Each example has 103 features and belongs to one or multiple of the 14 classes. We randomly split the dataset into two sets, 20% as the validation set to tune hyper-parameters and 80% as the test set to test the performance of different algorithms. For space reasons, the details and the results of the mediamill dataset are in Appendix B. To simulate the bandit environment, we treat classes as arms and their labels (0 or 1) as rewards. For each round t, the bandit environment randomly samples an example from the dataset. Then the bandit algorithm selects an arm (class), and its reward (class label) is revealed to the algorithm. To construct context vectors for the arms, we generate 50-dimensional random Fourier features (Rahimi et al., 2007) from the outer product between the features of the example and the one-hot representation of the arms.

For the experiments on real-world data, we use data from the Yahoo! Today Module (Li et al., 2010), which contains user click logs from a news-article recommender system that was fielded for 10 days. Each day logged around 4.6 million events from a bandit that selected articles uniformly at random, which allows the use of the replay methodology (Li et al., 2010) for unbiased offline evaluation of new bandit algorithms. We use the data from the first day for hyperparameter selection and report the results on the data from the second day. The results using all the data are presented in Appendix B. Each article and each user is represented by a 6-dimensional feature vector respectively. Following (Li et al., 2010), we use the outer product between the user features and the article features as the context vector.

To calculate the fairness and reward regret, we determine the optimal fair policy as follows. For MAB experiments, we use the empirical mean reward of each arm as the mean parameter for each arm. For linear bandit experiments, we fit a linear least square model that maps the context vector of each arm to its reward. Note that the linearity assumption does not necessarily hold for any of the datasets, and that rewards are known to change over time for the Yahoo! data. This realism adds a robustness component to the evaluation.

We also add straightforward FairX-variants of the ϵ-greedy algorithms to the empirical analysis, which we call FairX-EG and FairX-LinEG. The algorithms are identical to their conventional ϵ-greedy counterparts, except that they construct their policies according to $\pi_t(a) = \frac{f(\hat{\mu}_{t,a})}{\sum_{a'} f(\hat{\mu}_{t,a'})}$ or $\pi_t(a) = \frac{f(\hat{\mu}_t \cdot x_{t,a})}{\sum_{a'} f(\hat{\mu}_t \cdot x_{t,a})}$ where $\hat{\mu}_t$ is the estimated parameter at round t. While ϵ-greedy has weaker guarantees already in the conventional bandit setting, it is well known that it often performs well empirically and we thus add it as a reference for the more sophisticated algorithms. In addition to the FairX algorithms, we also include the fairness regret of conventional UCB, TS, LinUCB, and LinTS bandit algorithms.

We use merit functions of the form $f(\mu) = \exp(c\mu)$, since the choice of the constant c provides a straightforward way to explore how the algorithms perform for steeper vs. flatter merit functions. In particular, the choice of c varies the value of L/γ. For both FairX-UCB and FairX-LinUCB, we use projected gradient descent to solve the non-convex optimization problem each round. We set the learning rate to be 0.01 and the number of steps to be 10. For FairX-LinUCB, we use a fixed $\beta_t = \beta$ for all rounds.

In general, we use grid search to tune hyper-parameters to minimize fairness regret on the validation set and report the performance on the test set. We grid search w_0 for FairX-UCB and UCB; prior variance and reward variance for FairX-TS, TS, FairX-LinTS and LinTS; λ and β for FairX-LinUCB and LinUCB; ϵ for FairX-EG; ϵ and the regularization parameter of the ridge regression for FairX-LinEG. We run each experiment 10 times and report the mean and the standard deviation.

5.2. How unfair are conventional bandit algorithms?

We first verify that conventional bandit algorithms indeed violate merit-based fairness of exposure, and that our FairX algorithms specifically designed to ensure fairness do indeed perform better. Figure 1 shows the average exposure that each arm received across rounds under the conventional UCB and TS algorithms for a typical run, and it compares this to the exposure allocation under the FairX-UCB and FairX-TS algorithm. The plots show the average exposure after 2,000 (left) and 2,000,000 (right) rounds, and it also
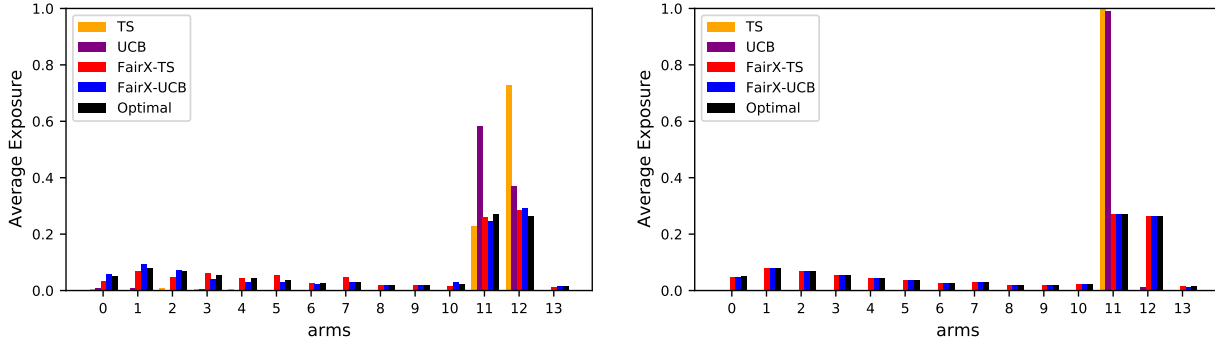
Figure 1. The average exposure distribution of different algorithms on the yeast dataset in the MAB setting after $2,000$ rounds (left) and $2,000,000$ rounds (right). ($c = 4$)

includes the optimally fair exposure allocation. Already after 2,000 rounds, the conventional algorithms under-expose many of the arms. After 2,000,000 rounds, they focus virtually all exposure on arm 11, even though arm 12 has only slightly lower merit. Both FairX-UCB and FairX-TS track the optimal exposure allocation substantially better, and they converge to the optimally fair solution. This verifies that FairX algorithms like FairX-UCB and FairX-TS are indeed necessary to enforce merit-based fairness of exposure. The following sections further show that conventional bandit algorithms consistently suffer from much larger fairness regret compared to FairX algorithms across different datasets and merit functions in both MAB and linear bandits setting.

5.3. How do the FairX algorithms compare in the MAB setting?
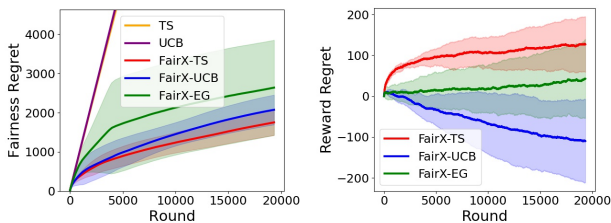

Figure 2. Fairness regret and reward regret of different MAB algorithms on the yeast dataset. ($c = 10$)

Figure 2 compares the performance of the bandit algorithms on the yeast dataset. The fairness regret converges roughly at the rate predicted by the bounds for FairX-UCB and FairX-TS, and FairX-EG shows a similar behavior as well. In terms of reward regret, all FairX algorithms perform substantially better than their worst-case bounds suggest. Note that FairX-UCB does particularly well in terms of reward regret, but also note that part of this is due to violating fairness more than FairX-TS. Specifically, in the FairX setting, an unfair policy can get better reward than the

optimal fair policy, making a negative reward regret possible. While FairX-EG wins neither on fairness regret nor on reward regret, it nevertheless does surprisingly well given the simplicity of the exploration scheme. We conjecture that FairX-EG benefits from the implicit exploration that results from the stochasticity of the fair policies. Results for other merit functions are given in Appendix B, and we find that the algorithms perform more similarly the flatter the merit function.

5.4. How do the FairX algorithms compare in the linear bandits setting?
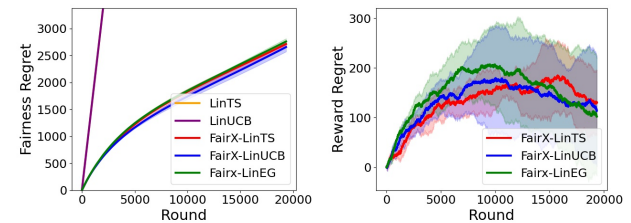

Figure 3. Fairness regret and reward regret of different linear bandit algorithms on the yeast dataset. ($c = 3$)

We show the fairness regret and the reward regret of the bandit algorithms on the yeast dataset in Figure 3. Results for other merit functions are in Appendix B. Similar to the MAB setting, there is no clear winner between the three FairX algorithms. Again we see some trade-offs between reward regret and fairness regret, but all three FairX algorithms show a qualitatively similar behavior. One difference is that the fairness regret no longer seems to converge. This can be explained with the misspecification of the linear model, as the estimated "optimal" policy that we use in our computation of regret may differ from the policy learned by the algorithms due to selection biases. Nevertheless, we conclude that the fairness achieved by the FairX algorithms is still highly preferable to that of the conventional bandit
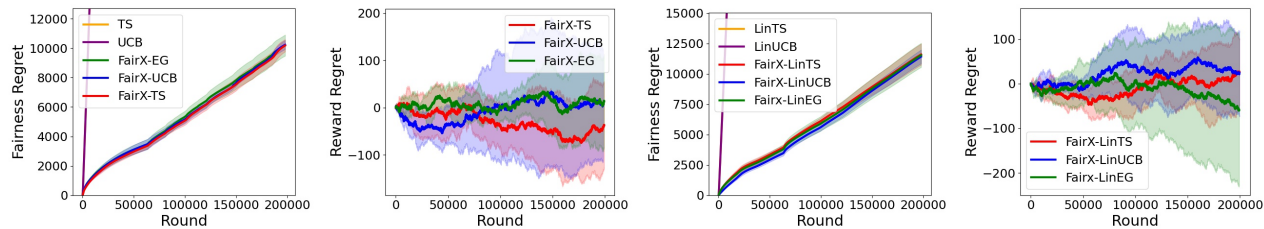
Figure 4. Experiment results on the Yahoo! dataset for both the MAB and the linear bandits setting ($c = 10$ for both settings).

algorithms.

5.5. How do the FairX algorithms compare on the real-world data?

To validate the algorithms on a real-world application, Figure 4 provides fairness and reward regret on the Yahoo! dataset for both the MAB and the linear bandits setting. Again, all three types of FairX algorithms perform comparably and have reward regret that converges quickly. Note that even the MAB setting now includes some misspecification of the model, since the reward distribution changes over time. This explains the behavior of the fairness regret. However, all FairX algorithms perform robustly in both settings, even though the real data does not exactly match the model assumptions.

6. Conclusions

We introduced a new bandit setting that formalizes merit-based fairness of exposure for both the stochastic MAB and the linear bandits setting. In particular, we define fairness regret and reward regret with respect to the optimal fair policy that fulfills the merit-based fairness of exposure, develop UCB and Thompson sampling algorithms for both settings, and prove bounds on their fairness and reward regret. An empirical evaluation shows that these algorithms provide substantially better fairness of exposure to the items, and that they are effective across a range of settings.

Acknowledgements

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., and Pizzato, L. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30(1):127–158, 2020.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Bechavod, Y., Ligett, K., Roth, A., Waggoner, B., and Wu, S. Z. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems*, pp. 8974–8984, 2019.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2212–2220, 2019.

Biega, A. J., Gummadi, K. P., and Weikum, G. Equity of attention: Amortizing individual fairness in rankings. In Collins-Thompson, K., Mei, Q., 0001, B. D. D., 0001, Y. L., and Yilmaz, E. (eds.), *SIGIR*, pp. 405–414. ACM, 2018. URL http://dl.acm.org/citation.cfm?id=3209978.

Blum, A. and Lykouris, T. Advancing subgroup fairness via sleeping experts. *arXiv preprint arXiv:1909.08375*, 2019.

Blum, A., Gunasekar, S., Lykouris, T., and Srebro, N. On preserving non-discrimination when combining expert advice. In *Advances in Neural Information Processing Systems*, pp. 8376–8387, 2018.

Brams, S. J. and Taylor, A. D. *Fair Division: From cake-cutting to dispute resolution*. Cambridge University Press, 1996.

Bretagnolle, J. and Huber, C. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Celis, L. E., Kapoor, S., Salehi, F., and Vishnoi, N. K. An algorithmic framework to control bias in bandit-based personalization. *arXiv preprint arXiv:1802.08674*, 2018.

Chen, Y., Cuellar, A., Luo, H., Modi, J., Nemlekar, H., and Nikolaidis, S. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pp. 181–190. PMLR, 2020.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

Claure, H., Chen, Y., Modi, J., Jung, M., and Nikolaidis, S. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 299–308, 2020.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Annual Conference on Learning Theory (COLT)*, 2008.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Epstein, R. and Robertson, R. E. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

Geyik, S. C., Ambler, S., and Kenthapadi, K. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2221–2231, 2019.

Gillen, S., Jung, C., Kearns, M., and Roth, A. Online learning with an unknown fairness metric. In *Advances in neural information processing systems*, pp. 2600–2609, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Heidari, H. and Krause, A. Preventing disparate treatment in sequential decision making. In *IJCAI*, pp. 2248–2254, 2018.

Horton, P. and Nakai, K. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Ismb*, volume 4, pp. 109–115, 1996.

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559*, 2016a.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 325–333, 2016b.

Joseph, M., Kearns, M., 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*, pp. 158–163, 2018.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS*, volume 67 of *LIPIcs*, pp. 43:1–43:23, 2017.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, F., Liu, J., and Ji, B. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 2019.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., and Parkes, D. C. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., and Diaz, F. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*, pp. 2243–2251, 2018.

Noble, S. U. *Algorithms of oppression: How search engines reinforce racism*. nyu Press, 2018.

Patil, V., Ghalme, G., Nair, V., and Narahari, Y. Achieving fairness in the stochastic multi-armed bandit problem. In *AAAI*, pp. 5379–5386, 2020.

Procaccia, A. D. Cake cutting: not just child's play. *Communications of the ACM*, 56(7):78–87, 2013.

Rahimi, A., Recht, B., et al. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 3, pp. 5, 2007.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

Schumann, C., Lang, Z., Mattei, N., and Dickerson, J. P. Group fairness in bandit arm selection. *arXiv preprint arXiv:1912.03802*, 2019.

Singh, A. and Joachims, T. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2219–2228, 2018.

Singh, A. and Joachims, T. Policy learning for fairness in ranking. In *Advances in Neural Information Processing Systems*, pp. 5426–5436, 2019.

Slivkins, A. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

Snoek, C. G., Worring, M., Van Gemert, J. C., Geusebroek, J.-M., and Smeulders, A. W. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th ACM international conference on Multimedia*, pp. 421–430, 2006.

Steihaus, H. The problem of fair division. *Econometrica*, 16:101–104, 1948.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Wen, M., Bastani, O., and Topcu, U. Fairness with dynamics. *arXiv preprint arXiv:1901.08568*, 2019.

A. Proofs of the Theorems

A.1. Proof of Theorem 3.1.1

Proof. First, the optimal fair policy satisfies the merit-based fairness of exposure constraints, since for any two arms $a, a' \in [K]$,

$$\frac{\pi^\star(a)}{f(\mu_a^\star)} = \frac{\pi^\star(a')}{f(\mu_{a'}^\star)} = \frac{1}{\sum_a f(\mu_a^\star)}.$$

Second, we show that the solution is unique. The merit-based fairness of exposure constraints on $\pi^\star$ correspond to $K-1$ linearly independent equations of $\pi^\star$. With an additional linear equation $\sum_a \pi^\star(a) = 1$ that is linearly independent of the other $K-1$ ones, we have K linearly independent equations on K unknowns in $\pi^\star$. Thus the solution to these equations is unique.

∎

A.2. Proof of Theorem 3.1.4

Proof. We prove the theorem by constructing two MAB instances and a 1-Lipschitz merit function. We show the sum of the expected fairness regrets of the two MAB instances are linear in T for any bandit algorithm under the merit function. Thus any bandit algorithm will have expected fairness regret linear in T for at least one of the two MAB instances.

Let us consider two MAB instances $v^1 = (P_1^1, P_2^1)$ and $v^2 = (P_1^2, P_2^2)$ where each instance has two arms, and each arm has reward distributions being Gaussian distributions with variance fixed to be $1/2$. The first instance's mean $\mu_1 = (\theta, 2\theta)$, i.e. $P_1^1 = \mathcal{N}(\theta, 1/2)$, $P_2^1 = \mathcal{N}(2\theta, 1/2)$ and the second instance's mean is $\mu_2 = (2\theta, 2\theta)$, i.e. $P_1^2 = \mathcal{N}(2\theta, 1/2)$, $P_2^2 = \mathcal{N}(2\theta, 1/2)$, where $\theta > 0$ is a positive constant. The merit function f is an identify function, i.e. $f(\cdot) = \cdot$. This means that the optimal fair policy for the first instance is $\pi^{\star,1} = [1/3, 2/3]$, while the optimal fair policy for the second instance is $\pi^{\star,2} = [1/2, 1/2]$. Let us consider any bandit algorithm $\mathcal{A}$ which at every round t, produces a policy π_t (may be in a randomized way), based on the history $\mathcal{H}_t = (\pi_1, a_1, r_{1,a_1}, \ldots, \pi_{t-1}, a_{t-1}, r_{t-1,a_{t-1}})$, i.e. $\pi_t \sim \mathcal{A}\left(\cdot | \pi_1, a_1, r_{1,a_1}, \ldots, \pi_{t-1}, a_{t-1}, r_{t-1,a_{t-1}}\right), a_t \sim \pi_t, r_t \sim P_{a_t}$.

Let us denote an outcome trajectory as $\tau = (\pi_1, a_1, r_{1,a_1}, \ldots, \pi_T, a_T, r_{T,a_T})$. Denote $\mathbb{P}^1$ as the distribution of τ of $\mathcal{A}$ interacting with the first MAB instance v^1, while $\mathbb{P}^2$ as the distribution of τ of $\mathcal{A}$ interacting with the second MAB instance v^2. The KL divergence between $\mathbb{P}^1$ and $\mathbb{P}^2$ can be upper bounded as follows:

$$\text{KL}\left(\mathbb{P}^1, \mathbb{P}^2\right) = \mathbb{E}_{\tau \sim \mathbb{P}^1}\left[\ln \frac{\mathbb{P}^1(\tau)}{\mathbb{P}^2(\tau)}\right] = \mathbb{E}_{\tau \sim \mathbb{P}^1}\left[\sum_{t=1}^T \ln \frac{P_{a_t}^1(r_{a_t})}{P_{a_t}^2(r_{a_t})}\right] = \sum_{t=1}^T \mathbb{E}_{\pi_t \sim \mathcal{A}^1} \mathbb{E}_{a_t \sim \pi_t} \text{KL}\left(P_{a_t}^1, P_{a_t}^2\right)$$

$$= \sum_{t=1}^T \mathbb{E}_{\pi_t \sim \mathcal{A}^1}\left[\pi_t(1)\text{KL}\left(P_1^1, P_1^2\right)\right] = \sum_{t=1}^T \mathbb{E}_{\pi_t \sim \mathcal{A}^1}\left[\pi_t(1)\theta^2\right] \leq \sum_{t=1}^T \mathbb{E}_{\pi_t \sim \mathcal{A}^1} \theta^2 = T\theta^2,$$

where $\pi_t \sim \mathcal{A}^1$ means that π_t is sampled from the process of $\mathcal{A}$ interacting with the first MAB instance. Namely, when $\theta \to 0$, it would be hard to distinguish between $\mathbb{P}^1$ and $\mathbb{P}^2$.

For any sequence of policies $\pi_1, \ldots, \pi_T$, we can lower bound the fairness regret for each instance as follows. For the fairness regret FR_T^1 of instance v^1, we have:

$$\mathbb{E}\left[\frac{1}{T}\text{FR}_T^1\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T (|\pi_t(1) - 1/3| + |\pi_t(2) - 2/3|)\right]$$

$$\geq \mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^T \pi_t(1) - 1/3\right| + \left|\frac{1}{T}\sum_{t=1}^T \pi_t(2) - 2/3\right|\right] = 2\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^T \pi_t(1) - 1/3\right|\right].$$

Similarly, for the fairness regret FR_T^2 of instance v^2, we have:

$$\mathbb{E}\left[\frac{1}{T}\text{FR}_T^2\right] \geq 2\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^T \pi_1(t) - 1/2\right|\right].$$

Thus we have:

$$\mathbb{E}\left[\mathrm{FR}_T^1/T\right] + \mathbb{E}\left[\mathrm{FR}_T^2/T\right] \geq \frac{1}{6}\mathbb{P}^1\left(\frac{1}{T}\sum_{t=1}^T \pi_t(1) > \frac{5}{12}\right) + \frac{1}{6}\mathbb{P}^2\left(\frac{1}{T}\sum_{t=1}^T \pi_t(1) \leq \frac{5}{12}\right)$$

$$\geq \frac{1}{12}\exp\left(-\mathrm{KL}\left(\mathbb{P}^1, \mathbb{P}^2\right)\right) \geq \frac{1}{12}\exp\left(-\theta^2 T\right)$$

where the second inequality applies the Bretagnolle-Huber inequality (Bretagnolle & Huber, 1979).

Set $\theta = 1/\sqrt{T}$, we prove that:

$$\mathbb{E}\left[\mathrm{FR}_T^1/T\right] + \mathbb{E}\left[\mathrm{FR}_T^2/T\right] \geq 0.03,$$

which implies that at least one instance suffers linear expected fairness regret.

This concludes the proof.

∎

A.3. Proof of Theorem 3.1.5

Proof. Similar to the proof of Theorem 3.1.4, we construct two MAB instances and a merit function with minimum merit 1. We show the sum of the expected fairness regrets of the two MAB instances are linear in T for any bandit algorithm under this merit function. Thus any bandit algorithm will have expected fairness regret linear in T for at least one of the two MAB instances.

Let us consider two MAB instances $v^1 = (P_1^1, P_2^1)$ and $v^2 = (P_1^2, P_2^2)$ where each instance has two arms, and each arm has reward distributions being Gaussian distributions with variance fixed to be $1/2$. The first instance's mean $\mu_1 = (\theta, 0)$, i.e. $P_1^1 = \mathcal{N}(\theta, 1/2)$, $P_2^1 = \mathcal{N}(0, 1/2)$ and the second instance's mean is $\mu_2 = (0, 0)$, i.e. $P_1^2 = \mathcal{N}(0, 1/2)$, $P_2^2 = \mathcal{N}(0, 1/2)$, where $\theta > 0$ is a positive constant to be set later. The merit function f with minimum merit 1 is a piece-wise linear function

$$f(\mu) = \begin{cases} 1 & \mu \leq 0 \\ L\mu + 1 & \mu > 0 \end{cases}$$

where $L > 0$ is a positive constant to be set later. This means that the optimal fair policy for the first instance is $\pi^{\star,1} = [(L\theta+1)/(L\theta+2), 1/(L\theta+2)]$, while the optimal fair policy for the second instance is $\pi^{\star,2} = [1/2, 1/2]$. Let us consider any algorithm $\mathcal{A}$ which at every round t, produces a policy π_t (may be in a randomized way), based on the history $\mathcal{H}_t = (\pi_1, a_1, r_{1,a_1}, \ldots, \pi_{t-1}, a_{t-1}, r_{t-1,a_{t-1}})$, i.e. $\pi_t \sim \mathcal{A}\left(\cdot|\pi_1, a_1, r_{1,a_1}, \ldots, \pi_{t-1}, a_{t-1}, r_{t-1,a_{t-1}}\right)$, $a_t \sim \pi_t, r_t \sim P_{a_t}$.

Let us denote an outcome trajectory as $\tau = (\pi_1, a_1, r_{1,a_1}, \ldots, \pi_T, a_T, r_{T,a_T})$. Denote $\mathbb{P}^1$ as the distribution of τ of $\mathcal{A}$ interacting with the first MAB instance v^1, while $\mathbb{P}^2$ as the distribution of τ of $\mathcal{A}$ interacting with the second MAB instance v^2. The KL divergence between $\mathbb{P}^1$ and $\mathbb{P}^2$ can be upper bounded as follows:

$$\mathrm{KL}\left(\mathbb{P}^1, \mathbb{P}^2\right) = \mathbb{E}_{\tau \sim \mathbb{P}^1}\left[\ln \frac{\mathbb{P}^1(\tau)}{\mathbb{P}^2(\tau)}\right] = \mathbb{E}_{\tau \sim \mathbb{P}^1}\left[\sum_{t=1}^T \ln \frac{P_{a_t}^1(r_{a_t})}{P_{a_t}^2(r_{a_t})}\right] = \sum_{t=1}^T \mathbb{E}_{\pi_t \sim \mathcal{A}^1}\mathbb{E}_{a_t \sim \pi_t}\mathrm{KL}\left(P_{a_t}^1, P_{a_t}^2\right)$$

$$= \sum_{t=1}^T \mathbb{E}_{\pi_t \sim \mathcal{A}^1}\left[\pi_t(1)\mathrm{KL}\left(P_1^1, P_1^2\right)\right] = \sum_{t=1}^T \mathbb{E}_{\pi_t \sim \mathcal{A}^1}\left[\pi_t(1)\theta^2\right] \leq \sum_{t=1}^T \mathbb{E}_{\pi_t \sim \mathcal{A}^1}\theta^2 = T\theta^2,$$

where $\pi_t \sim \mathcal{A}^1$ means that π_t is sampled from the process of $\mathcal{A}$ interacting with the first MAB instance.

For any sequence of policies $\pi_1, \ldots, \pi_T$, we can lower bound the fairness regret for each instance as follows. For the

fairness regret FR_T^1 of instance v^1, we have:

$$\mathbb{E}\left[\frac{1}{T}\text{FR}_T^1\right] = \frac{1}{T}\sum_{t=1}^T \left(|\pi_t(1) - (L\theta+1)/(L\theta+2)| + |\pi_t(2) - 1/(L\theta+2)|\right)$$

$$\geq \left|\frac{1}{T}\sum_{t=1}^T \pi_t(1) - (L\theta+1)/(L\theta+2)\right| + \left|\frac{1}{T}\sum_{t=1}^T \pi_t(2) - 1/(L\theta+2)\right|$$

$$= 2\left|\frac{1}{T}\sum_{t=1}^T \pi_t(1) - (L\theta+1)/(L\theta+2)\right|.$$

Similarly, for the fairness regret FR_T^2 of instance v^2, we have:

$$\mathbb{E}\left[\frac{1}{T}\text{FR}_T^2\right] \geq 2\left|\frac{1}{T}\sum_{t=1}^T \pi_1(t) - 1/2\right|.$$

Thus we have:

$$\mathbb{E}\left[\text{FR}_T^1/T\right] + \mathbb{E}\left[\text{FR}_T^2/T\right] \geq \frac{L\theta}{2L\theta+4}\mathbb{P}^1\left(\frac{1}{T}\sum_{t=1}^T \pi_t(1) \leq \frac{3L\theta+4}{4L\theta+8}\right) + \frac{L\theta}{2L\theta+4}\mathbb{P}^2\left(\frac{1}{T}\sum_{t=1}^T \pi_t(1) > \frac{3L\theta+4}{4L\theta+8}\right)$$

$$\geq \frac{L\theta}{4L\theta+8}\exp\left(-\text{KL}\left(\mathbb{P}^1, \mathbb{P}^2\right)\right) \geq \frac{L\theta}{4L\theta+8}\exp\left(-\theta^2 T\right)$$

where the second inequality applies the Bretagnolle-Huber inequality (Bretagnolle & Huber, 1979).

Set $\theta = 1/\sqrt{T}$ and $L = \sqrt{T}$, we prove that:

$$\mathbb{E}\left[\text{FR}_T^1/T\right] + \mathbb{E}\left[\text{FR}_T^2/T\right] \geq 0.03,$$

which implies that at least one instance suffers linear expected fairness regret.

This concludes the proof.

∎

A.4. Proof of Theorem 3.2.2

Lemma A.4.1. *For any $\delta \in (0,1)$, with probability at least $1-\delta/2$, $\forall\, t > K, a \in [K]$, $\mu^\star \in CR_t$.*

Proof. For any $t > K$ and $a \in [K]$, apply Hoeffding's inequality, we have with probability at least $1 - \delta/(2KT)$,

$$|\hat{\mu}_{t,a} - \mu_a^\star| \leq \sqrt{2\ln(4KT/\delta)/N_{t,a}}.$$

Apply union bound to $\forall\, t > K, a \in [K]$, we conclude the proof. ∎

Lemma A.4.2. *For any $\delta \in (0,1)$, with probability at least $1-\delta/2$,*

$$\left|\sum_{t=K+1}^T \mathbb{E}_{a\sim\pi_t}\sqrt{1/N_{t,a}} - \sum_{t=K+1}^T \sqrt{1/N_{t,a_t}}\right| \leq \sqrt{2T\ln(4/\delta)}.$$

Proof. The sequence

$$\sqrt{1/N_{t,a_t}} - \mathbb{E}_{a\sim\pi_t}\sqrt{1/N_{t,a}}$$

is a martingale difference sequence, and $\forall t > K$

$$\left|\sqrt{1/N_{t,a_t}} - \mathbb{E}_{a\sim\pi_t}\sqrt{1/N_{t,a}}\right| \leq 1.$$

We can apply the Azuma-Hoeffding's inequality to get with probability at least $1 - \delta/2$,

$$\left| \sum_{t=K+1}^{T} \mathbb{E}_{a \sim \pi_t} \sqrt{1/N_{t,a}} - \sum_{t=K+1}^{T} \sqrt{1/N_{t,a_t}} \right| \leq \sqrt{2T \ln(4/\delta)}.$$

This concludes the proof.

∎

Proof. (Theorem 3.2.2)

The reward regret can be upper bounded as follows:

$$
\begin{aligned}
\mathrm{RR}_T &= \sum_{t=1}^{T} \sum_{a} (\pi^\star(a) - \pi_t(a)) \mu_a^\star \\
&\leq 2K + \sum_{t=K+1}^{T} \sum_{a} \pi_t(a) \mu_{t,a} - \pi_t(a) \mu_a^\star \\
&= 2K + \sum_{t=K+1}^{T} \sum_{a} \pi_t(a) (\mu_{t,a} - \hat{\mu}_{t,a} + \hat{\mu}_{t,a} - \mu_a^\star) \\
&\leq 2K + \sum_{t=K+1}^{T} \sum_{a} \pi_t(a) 2 \sqrt{\frac{2 \ln(4TK/\delta)}{N_{t,a}}} \\
&= 2K + 2\sqrt{2 \ln(4TK/\delta)} \sum_{t=K+1}^{T} \mathbb{E}_{a \sim \pi_t} \sqrt{1/N_{t,a}} \\
&\leq 2K + 2\sqrt{2 \ln(4TK/\delta)} \left(\sqrt{2T \ln(4/\delta)} + \sum_{t=K+1}^{T} \sqrt{1/N_{t,a_t}} \right) \\
&\leq 2K + 2\sqrt{2 \ln(4TK/\delta)} \left(\sqrt{2T \ln(4/\delta)} + 2\sqrt{TK} \right).
\end{aligned}
$$

The first inequality comes from Line 7 in Algorithm 1. The second inequality comes from Lemma A.4.1. The third inequality comes from Lemma A.4.2 and The last inequality applies the AM-GM inequality. Thus when $T > K$, with probability at least $1 - \delta$,

$$\mathrm{RR}_T = \tilde{O}\left(\sqrt{TK}\right).$$

This concludes the proof.

∎

A.5. Proof of Theorem 3.2.1

Proof. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the events in Lemma A.4.1 and Lemma A.4.2 hold and at each round $t > K$,

$$
\sum_{a=1}^{K} |\pi_t(a) - \pi^\star(a)|
$$

$$
= \sum_{a=1}^{K} \left| \frac{f(\mu_{t,a})}{\sum_{a'=1}^{K} f(\mu_{t,a})} - \frac{f(\mu_a^\star)}{\sum_{a'=1}^{K} f(\mu_{a'}^\star)} \right|
$$

$$
= \sum_{a=1}^{K} \frac{\left| f(\mu_{t,a}) \sum_{a'=1}^{K} f(\mu_{a'}^\star) - f(\mu_a^\star) \sum_{a'=1}^{K} f(\mu_{t,a'}) \right|}{\sum_{a'=1}^{K} f(\mu_{t,a}) \sum_{a'=1}^{K} f(\mu_{a'}^\star)}
$$

$$
= \sum_{a=1}^{K} \frac{\left| f(\mu_{t,a}) \sum_{a'=1}^{K} f(\mu_{a'}^\star) - f(\mu_a^\star) \sum_{a'=1}^{K} f(\mu_{a'}^\star) + f(\mu_a^\star) \sum_{a'=1}^{K} f(\mu_{a'}^\star) - f(\mu_a^\star) \sum_{a'=1}^{K} f(\mu_{t,a'}) \right|}{\sum_{a'=1}^{K} f(\mu_{t,a}) \sum_{a'=1}^{K} f(\mu_{a'}^\star)}
$$

$$
\leq \frac{\sum_{a=1}^{K} |f(\mu_{t,a}) - f(\mu_a^\star)| \sum_{a'=1}^{K} f(\mu_{a'}^\star) + \sum_{a=1}^{K} f(\mu_a^\star) \sum_{a'=1}^{K} |f(\mu_{a'}^\star) - f(\mu_{t,a})|}{\sum_{a'=1}^{K} f(\mu_{t,a}) \sum_{a'=1}^{K} f(\mu_{a'}^\star)}
$$

$$
= \frac{2 \sum_{a=1}^{K} |f(\mu_{t,a}) - f(\mu_a^\star)|}{\sum_{a'=1}^{K} f(\mu_{t,a})}
$$

$$
= \frac{2 \sum_{a=1}^{K} \frac{f(\mu_{t,a})}{f(\mu_{t,a})} |f(\mu_{t,a}) - f(\mu_a^\star)|}{\sum_{a'=1}^{K} f(\hat{\mu}_{t,a})}
$$

$$
\leq \sum_{a=1}^{K} \frac{4 L \pi_t(a)}{\gamma} \sqrt{2 \ln(4TK/\delta)/N_{t,a_t}}
$$

$$
= \frac{4L \sqrt{2 \ln(4TK/\delta)}}{\gamma} \mathbb{E}_{a \sim \pi_t} \left[1/\sqrt{N_{t,a_t}} \right].
$$

The second inequality comes from lemma A.4.1. And by lemma A.4.2,

$$
\sum_{t=K+1}^{T} \sum_{a=1}^{K} |\pi_t(a) - \pi^\star(a)| \leq \frac{4L \sqrt{2 \ln(4TK/\delta)}}{\gamma} \sum_{t=K+1}^{T} \mathbb{E}_{a \sim \pi_t} \left[1/\sqrt{N_{t,a_t}} \right]
$$

$$
\leq \frac{4L \sqrt{2 \ln(4TK/\delta)}}{\gamma} \left(\sqrt{2T \ln(4/\delta)} + \sum_{t=K+1}^{T} \sqrt{1/N_{t,a_t}} \right)
$$

$$
\leq \frac{4L \sqrt{2 \ln(4TK/\delta)}}{\gamma} \left(\sqrt{2T \ln(4/\delta)} + 2\sqrt{TK} \right).
$$

So when $T > K$, with probability at least $1 - \delta$, the fairness regret

$$
\mathrm{FR}_T \leq 2K + \frac{4L \sqrt{2 \ln(4TK/\delta)}}{\gamma} \left(\sqrt{2T \ln(4/\delta)} + 2\sqrt{TK} \right) = \tilde{O} \left(L\sqrt{TK}/\gamma \right),
$$

which concludes the proof.

■

A.6. Proof of Theorem 4.2.1

Proposition A.6.1. *(Confidence) For $\delta \in (0, 1)$, assume $\|\mu^\star\|_2 \leq W$, set $\beta_t = \left(W + \sqrt{d \ln(1 + t/d) + 2 \ln(\pi^2 t^2/3\delta)} \right)^2$ with probability at least $1 - \delta/2$, $\forall t$, $\mu^\star \in CR_t$.*

Section A.6.1 is devoted to establishing this confidence analysis.

Proposition A.6.2. *Let*

$$fr_t = \sum_a |\pi_t^\star(a) - \pi_t(a)| \tag{9}$$

denote the instantaneous fairness regret acquired by the algorithm at round t. For the FairX-LinUCB algorithm, if $\mu^\star \in CR_t$ for all $t \leq T$, then with probability at least $1 - \delta/2$

$$\sum_{t=1}^T fr_t \leq \frac{4L\sqrt{\beta_T}}{\gamma}\sqrt{2Td\ln(1+\frac{T}{d})} + \frac{4L\sqrt{\beta_T}}{\gamma}\sqrt{2T\ln(4/\delta)}. \tag{10}$$

Lemma A.6.3 (Lemma 7 in (Dani et al., 2008)). *For the FairX-LinUCB algorithm, if $\mu \in CR_t$, then for any $x \in \mathbb{R}^d$*

$$|(\mu - \hat{\mu}_t) \cdot x| \leq \sqrt{\beta_t x^\top \Sigma_t^{-1} x}.$$

Define

$$w_{t,a} := \sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}}$$

which we interpret as the "normalized width" at time t for action a.

Lemma A.6.4. *For the FairX-LinUCB algorithm, with probability $1 - \delta/2$,*

$$\left| \sum_{t=1}^T w_{t,a_t} - \sum_{t=1}^T \mathbb{E}_{a \sim \pi_t} w_{t,a} \right| \leq \sqrt{2T\ln(4/\delta)}.$$

Proof. The sequence

$$w_{t,a_t} - \mathbb{E}_{a \sim \pi_t} w_{t,a}$$

is a martingale difference sequence and $\forall t$

$$w_{t,a} = \|x_{t,a_t}\|_{\Sigma_t^{-1}} \leq \sqrt{\lambda_{\max}(\Sigma_t^{-1})} \|x_{t,a_t}\|_2 \leq 1,$$

where $\lambda_{max}(\cdot)$ denotes the largest eigenvalue of a matrix. Using Azuma-Hoeffding's inequality, with probability at least $1 - \delta/2$

$$\left| \sum_{t=1}^T w_{t,a_t} - \sum_{t=1}^T \mathbb{E}_{a \sim \pi_t} w_{t,a} \right| \leq \sqrt{2T\ln(4/\delta)}.$$

∎

Lemma A.6.5. *For the FairX-LinUCB algorithm , if $\forall t$, $\mu^\star \in CR_t$, then with probability at least $1 - \delta/2$*

$$\sum_{t=1}^T fr_t \leq \frac{4L\sqrt{\beta_t}}{\gamma}\sum_{t=1}^T w_{t,a_t} + \frac{4L\sqrt{\beta_t}}{\gamma}\sqrt{2T\ln(4/\delta)}$$

Proof.

$$fr_t$$

$$=\sum_a \left| \frac{f(\mu^\star \cdot x_{t,a})}{\sum_{a'} f(\mu^\star \cdot x_{t,a'})} - \frac{f(\mu_t \cdot x_{t,a})}{\sum_{a'} f(\mu_t \cdot x_{t,a'})} \right|$$

$$=\sum_a \left| \frac{f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu_t \cdot x_{t,a'}) - f(\mu_t \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'})}{\sum_{a'} f(\mu^\star \cdot x_{t,a'}) \sum_{a'} f(\mu_t \cdot x_{t,a'})} \right|$$

$$=\sum_a \left| \frac{f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu_t \cdot x_{t,a'}) - f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'}) + f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'}) - f(\mu_t \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'})}{\sum_{a'} f(\mu^\star \cdot x_{t,a'}) \sum_{a'} f(\mu_t \cdot x_{t,a'})} \right|$$

$$\leq \sum_a \frac{|f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu_t \cdot x_{t,a'}) - f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'})| + |f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'}) - f(\mu_t \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'})|}{\sum_{a'} f(\mu^\star \cdot x_{t,a'}) \sum_{a'} f(\mu_t \cdot x_{t,a'})}$$

$$\leq \frac{2 \sum_a |f(\mu^\star \cdot x_{t,a}) - f(\mu_t \cdot x_{t,a})|}{\sum_{a'} f(\mu_t \cdot x_{t,a'})}$$

$$=2 \sum_a \frac{\pi_t(a)}{f(\mu_t \cdot x_{t,a})} |f(\mu^\star \cdot x_{t,a}) - f(\hat{\mu}_t \cdot x_{t,a}) + f(\hat{\mu}_t \cdot x_{t,a}) - f(\mu_t \cdot x_{t,a})|$$

$$\leq \frac{2L}{\gamma} \mathbb{E}_{a \sim \pi_t} \left[\|\mu^\star - \hat{\mu}_t\|_{\Sigma_t} \|x_{t,a}\|_{\Sigma_t^{-1}} + \|\mu_t - \hat{\mu}_t\|_{\Sigma_t} \|x_{t,a}\|_{\Sigma_t^{-1}} \right]$$

$$\leq \frac{4L\sqrt{\beta_t}}{\gamma} \mathbb{E}_{a \sim \pi_t} w_{t,a}.$$

So by lemma A.6.4 and that β_t is increasing,

$$\sum_{t=1}^T fr_t \leq \frac{4L\sqrt{\beta_T}}{\gamma} \sum_{t=1}^T w_{t,a_t} + \frac{4L\sqrt{\beta_T}}{\gamma} \sqrt{2T \ln(4/\delta)}.$$

∎

Lemma A.6.6 (Lemma 10 in (Dani et al., 2008)). *We have $\forall t$*

$$\det \Sigma_t = \prod_{\tau=1}^{t-1} (1 + w_{t,a_t}^2).$$

Lemma A.6.7. $\forall t$, $\det \Sigma_{t+1} \leq (1 + t/d)^d$.

Proof.

$$\text{Trace } \Sigma_{t+1} = \text{Trace } \left(I + \sum_{\tau=1}^t x_{\tau,a_\tau} x_{\tau,a_\tau}^\top \right)$$

$$= d + \sum_{\tau=1}^{t=1} \text{Trace } \left(x_{t,a_t} x_{t,a_t}^\top \right) \tag{11}$$

$$= d + \sum_{\tau=1}^t \|x_{t,a_t}\|_2^2$$

$$\leq d + t.$$

Now, recall that Trace Σ_t equals the sum of the eigenvalues of Σ_t. On the other hand, $\det(\Sigma_t)$ equals the product of the eigenvalues. Since Σ_t is positive definite, its eigenvalues are all positive. Subject to these constraints, by AM-GM inequality, $\det(\Sigma_t)$ is maximized when all the eigenvalues are equal; the desired bound follows. ∎

Lemma A.6.8. *We have for all t,*

$$\sum_{\tau=1}^{t} w_{t,a_t}^2 \leq 2d \ln\left(1 + t/d\right).$$

Proof. Using the fact that for $0 \leq y \leq 1$, $ln(1+y) \geq y/2$, we have

$$
\begin{aligned}
\sum_{\tau=1}^{t} w_{t,a}^2 &\leq 2 \sum_{\tau=1}^{t} \ln(1 + w_{t,a_t}^2) \\
&= 2 \ln(\det(\Sigma_{t+1})) \\
&\leq 2d \ln(1 + t/d)
\end{aligned}
$$

by the previous two lemmas. $\blacksquare$

Proof. (Proof of Proposition A.6.2)

$$
\begin{aligned}
\sum_{t=1}^{T} fr_t &\leq \frac{4L\sqrt{\beta_T}}{\gamma} \sum_{t=1}^{T} w_{t,a_t} + \frac{4L\sqrt{\beta_T}}{\gamma} \sqrt{2T \ln(4/\delta)} \\
&\leq \frac{4L\sqrt{\beta_T}}{\gamma} \sqrt{T \sum_{t=1}^{T} w_{t,a_t}^2} + \frac{4L\sqrt{\beta_T}}{\gamma} \sqrt{2T \ln(4/\delta)} \\
&\leq \frac{4L\sqrt{\beta_T}}{\gamma} \sqrt{2Td \ln(1 + \frac{T}{d})} + \frac{4L\sqrt{\beta_T}}{\gamma} \sqrt{2T \ln(4/\delta)}.
\end{aligned}
$$

$\blacksquare$

Proof. (Proof of theorem 4.2.1)

By Proposition A.6.1 and Proposition A.6.2, with probability at least $1 - \delta$,

$$\text{FR}_T = \sum_{t=1}^{T} fr_t \leq \frac{4L\sqrt{\beta_T}}{\gamma} \sqrt{2Td \ln(1 + \frac{t}{d})} + \frac{4L\sqrt{\beta_T}}{\gamma} \sqrt{2T \ln(4/\delta)} = \tilde{O}\left(Ld\sqrt{T}/\gamma\right).$$

$\blacksquare$

A.6.1. CONFIDENCE ANALYSIS

In this section, we prove Proposition A.6.1, which states that with high probability, the true parameter $\mu^\star$ lies in the confidence Region CR_t for all t.

Proof. (Proof of Proposition A.6.1)

Since $r_{\tau,a_\tau} = \mu^\star \cdot x_{\tau,a_\tau} + \eta_\tau$, we have

$$\hat{\mu}_t - \mu^\star = \Sigma_t^{-1} \sum_{\tau=1}^{t-1} r_{\tau,a_\tau} x_{\tau,a_\tau} - \mu^\star = \Sigma_t^{-1} \mu^\star + \Sigma_t^{-1} \sum_{\tau=1}^{t-1} \eta_\tau x_{\tau,a_\tau}$$

For any $0 < \delta_t < 1$, using self-normalized bound for vector-valued martingales (Abbasi-Yadkori et al., 2011) and

Lemma A.6.7, we have with probability at least $1 - \delta_t$,

$$\sqrt{(\hat{\mu}_t - \mu^\star)^\top \Sigma_t (\hat{\mu}_t - \mu^\star)} = \|\Sigma_t^{1/2}(\hat{\mu}_t - \mu^\star)\|_2$$

$$\leq \|\Sigma_t^{-1/2}\mu^\star\|_2 + \|\Sigma_t^{-1/2}\sum_{\tau=1}^{t-1}\eta_\tau x_{\tau,a_\tau}\|_2$$

$$\leq \|\mu^\star\|_2 + \sqrt{\ln(\det(\Sigma_t)\det(\Sigma_1)^{-1}/\delta_t^2)}$$

$$\leq W + \sqrt{d\ln(1 + t/d) + 2\ln(1/\delta_t)}.$$

Let $\delta_t = \frac{3\delta/\pi^2}{t^2}$,

$$\mathbb{P}(\forall t\ \mu^\star \in \mathrm{CR}_t) \geq 1 - \sum_{t=1}^{\infty}(\delta/t^2)(3/\pi^2) = 1 - \frac{\delta}{2}.$$

∎

A.7. Proof of Theorem 4.2.2

Proof. With probability at least $1 - \delta$, the events in Proposition A.6.1 and Lemma A.6.4 hold and

$$\mathrm{RR}_T$$

$$= \sum_{t=1}^{T}\mu^\star \cdot \mathbb{E}_{a\sim\pi_t^\star}[x_{t,a}] - \mu^\star \cdot \mathbb{E}_{a\sim\pi_t}[x_{t,a}]$$

$$\leq \sum_{t=1}^{T}\mu_t \cdot \mathbb{E}_{a\sim\pi_t}[x_{t,a}] - \mu^\star \cdot \mathbb{E}_{a\sim\pi_t}[x_{t,a}]$$

$$= \sum_{t=1}^{T}(\mu_t - \hat{\mu}_t) \cdot \mathbb{E}_{a\sim\pi_t}[x_{t,a}] + (\hat{\mu}_t - \mu^\star) \cdot \mathbb{E}_{a\sim\pi_t}[x_{t,a}]$$

$$\leq \sum_{t=1}^{T}\mathbb{E}_{a\sim\pi_t}[2\sqrt{\beta_t}w_{t,a}]$$

$$\leq 2\sqrt{\beta_T}\sum_{t=1}^{T}\mathbb{E}_{a\sim\pi_t}[w_{t,a}]$$

$$\leq 2\sqrt{\beta_T}\left(\sum_{t=1}^{T}w_{t,a_t} + 2\sqrt{2T\ln(4/\delta)}\right)$$

$$\leq 2\sqrt{\beta_T}\left(\sqrt{2Td\ln(1 + \frac{T}{d})} + 2\sqrt{2T\ln(4/\delta)}\right)$$

$$= \tilde{O}\left(d\sqrt{T}\right).$$

The first inequality comes from the algorithm. The second inequality comes from Lemma A.6.3. The third inequality comes from the fact that β_t is increasing. The fourth inequality comes from lemma A.6.4. And the last inequality comes from Lemma A.6.8.

∎

A.8. Proof of Theorem 4.3.1

Proof. Denote the posterior distribution of $\mu^\star$ conditioned on $\mathcal{H}_t$ as $p(\mu^\star|\mathcal{H}_t)$ and the corresponding conditional expectation as $\mathbb{E}[\cdot|\mathcal{H}_t]$. In stochastic linear bandits, the posterior distribution is a Gaussian distribution: $p(\cdot|\mathcal{H}_t) := \mathcal{N}\left(\hat{\mu}_t, \Sigma_t^{-1}\right)$. We notice that our μ_t and $\mu^\star$ are identically distributed from $p(\cdot|\mathcal{H}_t)$.

First, We can follow the same step we had above in the proof of fairness regret of FairX-LinUCB to upper bound the instantaneous fairness regret as follows (conditioned on history $\mathcal{H}_t$):

$$\mathbb{E}[fr_t] \leq \frac{2}{\gamma} \mathbb{E}_{\mathcal{H}_t} \left[\mathbb{E}_{\mu_t, \mu^\star} [\mathbb{E}_{a \sim \pi_t} |f(\mu_t \cdot x_{t,a}) - f(\mu^\star \cdot x_{t,a})| \mid \mathcal{H}_t] \right] \leq \frac{2L}{\gamma} \mathbb{E}_{\mathcal{H}_t} \left[\mathbb{E}_{\mu_t, \mu^\star} [\mathbb{E}_{a \sim \pi_t} |\mu_t \cdot x_{t,a} - \mu^\star \cdot x_{t,a}| \mid \mathcal{H}_t] \right]$$

Note that π_t is fully determined by μ_t and is independent of $\mu^\star$ given $\mathcal{H}_t$. In the following, We use π_{μ_t} to denote π_t to stress the dependence of π_t on μ_t. Hence, taking expectation with respect to the randomness of μ_t and $\mu^\star$:

$$\mathbb{E}_{\mu_t, \mu^\star} \left[\sum_a \pi_{\mu_t}(a) |(\mu_t - \mu^\star) \cdot x_{t,a}| \, \Big| \, \mathcal{H}_t \right] = \mathbb{E}_{\mu_t} \left[\sum_a \pi_{\mu_t}(a) \mathbb{E}_{\mu^\star} [|(\mu_t - \mu^\star) \cdot x_{t,a}| \mid \mathcal{H}_t, \mu_t] \, \Big| \, \mathcal{H}_t \right]$$

Note that for any $x_{t,a}$, conditioned on μ_t, we have:

$$(\mu_t - \mu^\star) \cdot x_{t,a} \sim \mathcal{N} \left(\mu_t \cdot x_{t,a} - \hat{\mu}_t \cdot x_{t,a}, x_{t,a}^\top \Sigma_t^{-1} x_{t,a} \right),$$

which means that:

$$\mathbb{E}_{\mu^\star} [|(\mu_t - \mu^\star) \cdot x_{t,a}| \mid \mathcal{H}_t, \mu_t] \leq \sqrt{\mathbb{E}_{\mu^\star} \left[|(\mu_t - \mu^\star) \cdot x_{t,a}|^2 \mid \mathcal{H}_t, \mu_t \right]} = \sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a} + ((\mu_t - \hat{\mu}_t) \cdot x_{t,a})^2}$$

$$\leq \sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}} + |(\mu_t - \hat{\mu}_t) \cdot x_{t,a}|$$

Denote the random variable $z_{t,a} := (\mu_t - \hat{\mu}_t) \cdot x_{t,a} / \sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}}$. Given $\mathcal{H}_t$, $z_{t,a}$ is a random variable and is only dependent on μ_t. Note that $z_{t,a} \sim \mathcal{N}(0, 1)$. Thus by the CDF of normal distribution, we have with probability at least $1 - \delta'$:

$$|(\mu_t - \hat{\mu}_t) \cdot x_{t,a}| \leq \sqrt{2 \ln(1/\delta')} \sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}}.$$

Allow union bound over all a and all T, we get with probability at least $1 - \delta'$:

$$\forall t \in [T], x_{t,a} \in \mathcal{D}_t : |(\mu_t - \hat{\mu}_t) \cdot x_{t,a}| \leq \sqrt{2 \ln(KT/\delta')} \sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}}.$$

Denote the above inequality at episode t as event $\mathcal{E}_t$ (note that $\mathcal{E}_t$ only depends on the random variable μ_t). We have

$$\mathbb{E}_{\mu_t} \left[\sum_a \pi_{\mu_t}(a) \mathbb{E}_{\mu^\star} [|(\mu_t - \mu^\star) \cdot x_{t,a}| \mid \mathcal{H}_t, \mu_t] \, \Big| \, \mathcal{H}_t \right]$$

$$= \mathbb{E}_{\mu_t} \left[\mathbf{1}\{\mathcal{E}_t\} \sum_a \pi_{\mu_t}(a) \mathbb{E}_{\mu^\star} [|(\mu_t - \mu^\star) \cdot x_{t,a}| \mid \mathcal{H}_t, \mu_t] \, \Big| \, \mathcal{H}_t \right] + \mathbb{E}_{\mu_t} \left[\mathbf{1}\{\overline{\mathcal{E}_t}\} \sum_a \pi_{\mu_t}(a) \mathbb{E}_{\mu^\star} [|(\mu_t - \mu^\star) \cdot x_{t,a}| \mid \mathcal{H}_t, \mu_t] \, \Big| \, \mathcal{H}_t \right]$$

$$\leq \mathbb{E}_{\mu_t} \left[\mathbb{E}_{a \sim \pi_{\mu_t}} \left((1 + \sqrt{2 \ln(KT/\delta')}) \sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}} \right) \, \Big| \, \mathcal{H}_t \right] + \underbrace{\mathbb{E}_{\mu_t} \left[\mathbf{1}\{\overline{\mathcal{E}_t}\} \sum_a \pi_{\mu_t}(a) \mathbb{E}_{\mu^\star} [|(\mu_t - \mu^\star) \cdot x_{t,a}| \mid \mathcal{H}_t, \mu_t] \, \Big| \, \mathcal{H}_t \right]}_{\text{term b}}.$$

Below we bound term b above. First note that:

$$\mathbb{E}_{\mu^\star} \left[|(\mu_t - \mu^\star) \cdot x_{t,a}|^2 | \mathcal{H}_t, \mu_t \right] \leq 2((\mu_t - \hat{\mu}_t) \cdot x_{t,a})^2 + 2\mathbb{E}_{\mu^\star} \left[((\mu^\star - \hat{\mu}_t) \cdot x_{t,a})^2 | \mathcal{H}_t, \mu_t \right]$$

$$= 2((\mu_t - \hat{\mu}_t) \cdot x_{t,a})^2 + 2 x_{t,a}^\top \Sigma_t^{-1} x_{t,a} \leq 2((\mu_t - \hat{\mu}_t) \cdot x_{t,a})^2 + 2,$$

where the first inequality uses the fact that $(a + b)^2 \leq 2a^2 + 2b^2$, the first equality uses the fact that $(\mu^\star - \hat{\mu}_t) \cdot x_{t,a} \sim \mathcal{N}(0, x_{t,a}^\top \Sigma_t^{-1} x_{t,a})$, and in the last inequality we use $\|x_{t,a}\|_2 \leq 1$ and $\det(\Sigma_t^{-1}) \leq 1$.

For term b above, we now can upper bound it as:

$$\text{term b} \le \mathbb{E}_{\mu_t}\left[\mathbf{1}\{\overline{\mathcal{E}_t}\}\sqrt{\sum_a \pi_{\mu_t}(a)^2}\sqrt{\sum_a \mathbb{E}_{\mu^\star}\left[((\mu_t-\mu^\star)\cdot x_{t,a})^2 \mid \mathcal{H}_t,\mu_t\right]}\,\Big|\mathcal{H}_t\right]$$

$$\le \mathbb{E}_{\mu_t}\left[\mathbf{1}\{\overline{\mathcal{E}_t}\}\sqrt{\sum_a \mathbb{E}_{\mu^\star}\left[((\mu_t-\mu^\star)\cdot x_{t,a})^2 \mid \mathcal{H}_t,\mu_t\right]}\,\Big|\mathcal{H}_t\right]$$

$$\le 2\mathbb{E}_{\mu_t}\left[\mathbf{1}\{\overline{\mathcal{E}_t}\}\sqrt{\sum_a ((\mu_t-\hat{\mu}_t)\cdot x_{t,a})^2 + K}\,\Big|\mathcal{H}_t\right]$$

$$\le 2\mathbb{E}_{\mu_t}\left[\mathbf{1}\{\overline{\mathcal{E}_t}\}\sqrt{\sum_a ((\mu_t-\hat{\mu}_t)\cdot x_{t,a})^2}\,\Big|\mathcal{H}_t\right] + 2\mathbb{E}_{\mu_t}\left[\mathbf{1}\{\overline{\mathcal{E}_t}\}\sqrt{K}\right].$$

Note that we can further upper bound the first term on the RHS of the above inequality as follows:

$$\mathbb{E}_{\mu_t}\left[\mathbf{1}\{\overline{\mathcal{E}_t}\}\sqrt{\sum_a ((\mu_t-\hat{\mu}_t)\cdot x_{t,a})^2}\,\Big|\mathcal{H}_t\right] \le \sqrt{\mathbb{E}_{\mu_t}\left[\mathbf{1}\{\mu_t\in\overline{\mathcal{E}_t}\} \mid \mathcal{H}_t\right]}\sqrt{\mathbb{E}_{\mu_t}\left[\sum_a ((\mu_t-\hat{\mu}_t)\cdot x_{t,a})^2\,\Big|\mathcal{H}_t\right]},$$

where we use the inequality that $\mathbb{E}[uv] \le \sqrt{\mathbb{E}[u^2]}\sqrt{\mathbb{E}[v^2]}$. Also note that:

$$\mathbb{E}_{\mu_t}\left[\sum_a ((\mu_t-\hat{\mu}_t)\cdot x_{t,a})^2\,\Big|\mathcal{H}_t\right] = \sum_a x_{t,a}^\top \Sigma_t^{-1} x_{t,a} \le K,$$

since $(\mu_t-\hat{\mu}_t)\cdot x_{t,a} \sim \mathcal{N}\left(0, x_{t,a}^\top \Sigma_t^{-1} x_{t,a}\right)$. Hence, we have:

$$\mathbb{E}_{\mu_t}\left[\mathbf{1}\{\overline{\mathcal{E}_t}\}\sqrt{\sum_a ((\mu_t-\hat{\mu}_t)\cdot x_{t,a})^2}\,\Big|\mathcal{H}_t\right] \le \sqrt{\mathbb{E}_{\mu_t}\left[\mathbf{1}\{\mu_t\in\overline{\mathcal{E}_t}\} \mid \mathcal{H}_t\right]}\sqrt{K}.$$

This implies that for term b, we have:

$$\text{term b} \le 2\sqrt{\mathbb{E}_{\mu_t}\left[\mathbf{1}\{\mu_t\in\overline{\mathcal{E}_t}\} \mid \mathcal{H}_t\right]}\sqrt{K} + 2\mathbb{E}_{\mu_t}\left[\mathbf{1}\{\overline{\mathcal{E}_t}\}\sqrt{K}\right] = 2\left(\sqrt{\mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)} + \mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)\right)\sqrt{K}.$$

Sum over T episodes, we have:

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{\mu_t}\left[\sum_a \pi_{\mu_t}(a)\mathbb{E}_{\mu^\star}\left[|(\mu_t-\mu^\star)\cdot x_{t,a}| \mid \mathcal{H}_t,\mu_t\right]\Big|\mathcal{H}_t\right]\right]$$

$$\le \frac{4L}{\gamma}\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\mathbb{E}_{\mu_t}\left[\mathbb{E}_{a\sim\pi_{\mu_t}}\left((1+\sqrt{2\ln(KT/\delta')})\sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}}\right)\Big|\mathcal{H}_t\right] + \left(\sqrt{\mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)} + \mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)\right)\sqrt{K}\right]$$

$$= \frac{4L}{\gamma}(1+\sqrt{2\ln(KT/\delta')})\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{a\sim\pi_{\mu_t}}\sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}}\right] + \frac{4L}{\gamma}\sqrt{K}\left(\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\sqrt{\mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)} + \mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)\right]\right).$$

For $\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\sqrt{\mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)} + \mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)\right]$, we have:

$$\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\sqrt{\mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)} + \mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)\right] = \sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\sqrt{\mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)} + \sum_{t=1}^T \mathbb{P}(\overline{\mathcal{E}_t}) \le \sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\sqrt{\mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)} + \delta'$$

$$\le \sqrt{T}\sqrt{\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\mathbb{P}(\overline{\mathcal{E}_t} \mid \mathcal{H}_t)} + \delta' \le \sqrt{T\delta'} + \delta'.$$

Hence, we have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\mu_t}\left[\sum_a \pi_{\mu_t}(a)\mathbb{E}_{\mu^\star}\left[|(\mu_t - \mu^\star)\cdot x_{t,a}| \mid \mathcal{H}_t, \mu_t\right]\Big|\mathcal{H}_t\right]\right]$$

$$\leq \frac{4L}{\gamma}(1 + \sqrt{2\ln(KT/\delta')})\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{a\sim\pi_{\mu_t}}\sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}}\right] + \frac{4L}{\gamma}\sqrt{K}\left(\sqrt{T\delta'} + \delta'\right).$$

From Lemma A.6.4, we have for any $\delta \in (0, 1)$,

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{a\sim\pi_{\mu_t}}\sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}}\right] \leq \delta T + \sqrt{2T\ln(4/\delta)} + \sum_{t=1}^{T} w_{t,a_t}$$

$$\leq \delta T + \sqrt{2T\ln(4/\delta)} + \sqrt{2Td\ln(1 + \frac{T}{d})}.$$

Let $\delta = 1/T$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{a\sim\pi_{\mu_t}}\sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}}\right] \leq 1 + \sqrt{2T\ln(4T)} + \sqrt{2Td\ln(1 + \frac{T}{d})}.$$

Hence, we have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\mu_t}\left[\sum_a \pi_{\mu_t}(a)\mathbb{E}_{\mu^\star}\left[|(\mu_t - \mu^\star)\cdot x_{t,a}| \mid \mathcal{H}_t, \mu_t\right]\Big|\mathcal{H}_t\right]\right]$$

$$\leq \frac{4L}{\gamma}(1 + \sqrt{2\ln(KT/\delta')})\left(\sqrt{2Td\ln(1 + \frac{T}{d})} + \sqrt{2T\ln(4T)} + 1\right) + \frac{4L}{\gamma}\sqrt{K}(\sqrt{T\delta'} + \delta').$$

Set $\delta' = 1/(KT)$, we have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\mu_t}\left[\sum_a \pi_{\mu_t}(a)\mathbb{E}_{\mu^\star}\left[|(\mu_t - \mu^\star)\cdot x_{t,a}| \mid \mathcal{H}_t, \mu_t\right]\Big|\mathcal{H}_t\right]\right] = \widetilde{O}(L\sqrt{Td}/\gamma).$$

$\blacksquare$

A.8.1. PROOF OF THEOREM 4.3.2

Lemma A.8.1 (Adapted from Proposition 1 from (Russo & Van Roy, 2014)). *For any UCB sequence $(U_t : t \in \mathbb{N})$, the Bayesian reward regret of FairX-LinTS can be decomposed as follows:*

$$BayesRR_T = \mathbb{E}\sum_{t=1}^{T}\left[\mathbb{E}_{a\sim\pi_t}\left[U_{t,a} - \mu^\star \cdot x_{t,a}\right]\right] + \mathbb{E}\sum_{t=1}^{T}\left[\mathbb{E}_{a\sim\pi_t^\star}\left[\mu^\star \cdot x_{t,a} - U_{t,a}\right]\right].$$

Proof. Note that at any round t, conditioned on history $\mathcal{H}_t$, the optimal fair policy $\pi_t^\star$ and the deployed policy π_t selected by posterior sampling are identically distributed. In addition, U_t is deterministic and fully determined by the history $\mathcal{H}_t$. Hence $\mathbb{E}\left[\mathbb{E}_{a\sim\pi_t}\left[U_{t,a}\right] \mid \mathcal{H}_t\right] = \mathbb{E}\left[\mathbb{E}_{a\sim\pi_t^\star}\left[U_{t,a}\right] \mid \mathcal{H}_t\right]$. Therefore

$$\mathbb{E}[\mathbb{E}_{a\sim\pi_t^\star}[\mu^\star \cdot x_{t,a}] - \mathbb{E}_{a\sim\pi_t}[\mu^\star \cdot x_{t,a}]]$$

$$= \mathbb{E}_{\mathcal{H}_t}[\mathbb{E}[\mathbb{E}_{a\sim\pi_t^\star}[\mu^\star \cdot x_{t,a}] - \mathbb{E}_{a\sim\pi_t}[\mu^\star \cdot x_{t,a}]|\mathcal{H}_t]]$$

$$= \mathbb{E}_{\mathcal{H}_t}[\mathbb{E}[\mathbb{E}_{a\sim\pi_t^\star}[\mu^\star \cdot x_{t,a}] - \mathbb{E}_{a\sim\pi_t^\star}U_{t,a} + \mathbb{E}_{a\sim\pi_t}U_{t,a} - \mathbb{E}_{a\sim\pi_t}[\mu^\star \cdot x_{t,a}]|\mathcal{H}_t]]$$

$$= \mathbb{E}[\mathbb{E}_{a\sim\pi_t}[U_{t,a} - \mu^\star \cdot x_{t,a}]] + \mathbb{E}[\mathbb{E}_{a\sim\pi^\star}[\mu^\star \cdot x_{t,a} - U_{t,a}]].$$

Summing over T steps concludes the proof. $\blacksquare$

Indeed, the above lemma holds for any U_t that is fully determined by the history $\mathcal{H}_t$ which does not have to be a valid upper bound.

Proof. (proof of Theorem 4.3.2)

Since Lemma A.8.1 holds for any confidence sequences, we construct confidence sequences from the confidence analysis of the FairX-LinUCB algorithm in Proposition A.6.1 for this proof. The upper bounds of an arm a across rounds are constructed as $U_{t,a} := \max_{\mu \in \text{CR}_t} \mu \cdot x_{t,a}$ and the lower bounds are constructed as $B_{t,a} := \min_{\mu \in \text{CR}_t} \mu \cdot x_{t,a}$.

We bound the reward regret under different conditions on the following three events. The first event is that the norm of the true parameter $\mu^\star$ sampled from the prior normal distribution is not too large. We denote the event that $\|\mu^\star\|_2 \leq W$ as $\mathcal{E}_1$. Since each dimension of $\mu^\star$ is independently sampled from the standard normal distribution, $\mathbb{P}(\overline{\mathcal{E}}_1) \leq 2d \exp(-W^2/2d)$. The second event is that for all rounds, the true parameter is within the confidence region of the FairX-LinUCB algorithm, which is the event in Proposition A.6.1. We denote the event $\forall t, \mu^\star \in \text{CR}_t$ as $\mathcal{E}_2$. By proposition A.6.1, we know that $\mathbb{P}(\overline{\mathcal{E}}_2) \leq \delta/2$. The third event is the event in Lemma A.6.4. We denote $\left| \sum_{t=1}^T w_{t,a_t} - \sum_{t=1}^T \mathbb{E}_{a \sim \pi_t} w_{t,a} \right| \leq \sqrt{2T \ln(4/\delta)}$ as $\mathcal{E}_3$ and $\mathbb{P}(\overline{\mathcal{E}}_3) \leq \delta/2$.

By Lemma A.8.1, we have that

BayesRR_T

$$= \mathbb{E}\left[\sum_{t=1}^T \sum_a (\pi_t^\star(a) - \pi_t(a))\mu^\star \cdot x_{t,a} \right]$$

$$= \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_2 \text{ and } \mathcal{E}_3\} \sum_{t=1}^T \sum_a (\pi_t^\star(a) - \pi_t(a))\mu^\star \cdot x_{t,a} \right] + \underbrace{\mathbb{E}\left[\mathbf{1}\{\overline{\mathcal{E}}_2 \text{ or } \overline{\mathcal{E}}_3\} \sum_{t=1}^T \sum_a (\pi_t^\star(a) - \pi_t(a))\mu^\star \cdot x_{t,a} \right]}_{\text{term c}}$$

$$= \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_2 \text{ and } \mathcal{E}_3\} \sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}[U_{t,a} - \mu^\star \cdot x_{t,a}] \right] + \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_2 \text{ and } \mathcal{E}_3\} \sum_{t=1}^T \mathbb{E}_{a \sim \pi^\star}[\mu^\star \cdot x_{t,a} - U_{t,a}] \right] + \text{term c}$$

$$\leq \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_2 \text{ and } \mathcal{E}_3\} \sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}[U_{t,a} - \mu^\star \cdot x_{t,a}] \right] + \text{term c}$$

$$= \underbrace{\mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ and } \mathcal{E}_3\} \sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}[U_{t,a} - \mu^\star \cdot x_{t,a}] \right]}_{\text{term a}} + \underbrace{\mathbb{E}\left[\mathbf{1}\{\overline{\mathcal{E}}_1 \text{ and } \mathcal{E}_2 \text{ and } \mathcal{E}_3\} \sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}[U_{t,a} - \mu^\star \cdot x_{t,a}] \right]}_{\text{term b}} + \text{term c.}$$

The inequality holds because under $\mathcal{E}_2$, $\mu^\star \cdot x_{t,a} \leq U_{t,a}$. We now bound the three terms as follows.

$$\text{term a} = \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ and } \mathcal{E}_3\} \sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}[U_{t,a} - \mu^\star \cdot x_{t,a}] \right]$$

$$\leq \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1 \text{ and } \mathcal{E}_2 \text{ and } \mathcal{E}_3\} \sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}[U_{t,a} - B_{t,a}] \right]$$

$$\leq 2\left(W + \sqrt{d \ln(1 + T/d) + 2\ln(T^2\pi^2/3\delta)} \right) \left(\sqrt{2Td \ln(1 + \frac{T}{d})} + 2\sqrt{2T \ln(4/\delta)} \right).$$

The first inequality holds because under $\mathcal{E}_2$, $B_{t,a}$ is a lower bound of $\mu^\star x_{t,a}$. The derivation of the second inequality is the same as the proof of the reward regret of the FairX-LinUCB algorithm.

To bound term c, we upper bound on the expected instantaneous reward regret under the event $\mathcal{E}_1$ and the event $\overline{\mathcal{E}}_1$

respectively. Under event $\mathcal{E}_1$, we have that at each round t, the expected instantaneous reward regret

$$\mathbb{E}\left[\mathbf{1}\{\mathcal{E}_1\}\sum_a (\pi_t^\star(a) - \pi_t(a))\mu^\star \cdot x_{t,a}\right] \leq \mathbb{E}\left[\mathbf{1}\{\|\mu^\star\|_2 \leq W\}\sum_a |\pi_t^\star(a) - \pi_t(a)|\,|\mu^\star \cdot x_{t,a}|\right]$$

$$\leq \mathbb{E}\left[\mathbf{1}\{\|\mu^\star\|_2 \leq W\}\sum_a |\pi_t^\star(a) - \pi_t(a)|\,\|\mu^\star\|_2 \cdot \|x_{t,a}\|_2\right]$$

$$\leq 2W.$$

Under event $\overline{\mathcal{E}}_1$, we have that at each round t, the instantaneous reward regret

$$\mathbb{E}\left[\mathbf{1}\{\overline{\mathcal{E}}_1\}\sum_a (\pi_t^\star(a) - \pi_t(a))\mu^\star \cdot x_{t,a}\right] \leq \mathbb{E}\left[\mathbf{1}\{\|\mu^\star\|_2 > W\}\sum_a |\pi_t^\star(a) - \pi_t(a)|\,\|\mu^\star\|_2 \cdot \|x_{t,a}\|_2\right]$$

$$\leq 2\mathbb{E}\left[\mathbf{1}\{\|\mu^\star\|_2 > W\}\|\mu^\star\|_2\right]$$

$$= 2\mathbb{E}_{\mu^\star \sim \mathcal{N}(0,I)}\left[\mathbf{1}\{\|\mu^\star\|_2 > W\}\|\mu^\star\|_2\right]$$

$$\leq 2\sqrt{\mathbb{E}_{\mu^\star \sim \mathcal{N}(0,I)}\mathbf{1}\{\|\mu^\star\|_2 \geq W\}}\sqrt{\mathbb{E}_{\mu^\star \sim \mathcal{N}(0,I)}\|\mu^\star\|_2^2}$$

$$\leq 2\sqrt{2d\exp\left(-W^2/2d\right)}\sqrt{\mathbb{E}_{\mu^\star \sim \mathcal{N}(0,I)}\left[\sum_{i,j}\mu_i^\star \mu_j^\star\right]}$$

$$\leq 2\sqrt{2d\exp\left(-W^2/2d\right)}\sqrt{d}$$

$$= 2d\sqrt{2\exp\left(-W^2/2d\right)}.$$

Here the third inequality uses $\mathbb{E}[ab] \leq \sqrt{\mathbb{E}[a^2]}\sqrt{\mathbb{E}[b^2]}$. Thus

$$\text{term c} = \mathbb{E}\left[\mathbf{1}\{\overline{\mathcal{E}}_2 \text{ or } \overline{\mathcal{E}}_3\}\sum_{t=1}^T \sum_a (\pi_t^\star(a) - \pi_t(a))\mu^\star \cdot x_{t,a}\right]$$

$$\leq \delta T(2d\sqrt{2\exp\left(-W^2/2d\right)} + 2W),$$

since no matter $\mathcal{E}_1$ or $\overline{\mathcal{E}}_1$, the expected instantaneous reward regret is less than the sum of the two situations.

For term b, denote $\mu_{t,a}^U$ the parameter that achieves the upper confidence bound $U_{t,a}$ in the confidence region CR_t.

$$\text{term b} = \mathbb{E}\left[\mathbf{1}\{\overline{\mathcal{E}}_1 \text{ and } \mathcal{E}_2 \text{ and } \mathcal{E}_3\}\sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}[U_{t,a} - \mu^\star \cdot x_{t,a}]\right]$$

$$= \mathbb{E}\left[\mathbf{1}\{\overline{\mathcal{E}}_1 \text{ and } \mathcal{E}_2 \text{ and } \mathcal{E}_3\}\sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}[\mu_{t,a}^U \cdot x_{t,a} - \mu^\star \cdot x_{t,a}]\right]$$

$$= \mathbb{E}\left[\mathbf{1}\{\overline{\mathcal{E}}_1 \text{ and } \mathcal{E}_2 \text{ and } \mathcal{E}_3\}\sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}[\mu_{t,a}^U \cdot x_{t,a} - \hat{\mu}_t \cdot x_{t,a} + \hat{\mu}_t \cdot x_{t,a} - \mu^\star \cdot x_{t,a}]\right]$$

$$\leq \mathbb{E}\left[\mathbf{1}\{\overline{\mathcal{E}}_1 \text{ and } \mathcal{E}_2 \text{ and } \mathcal{E}_3\}\sum_{t=1}^T \mathbb{E}_{a \sim \pi_t}\left[\|\mu_{t,a}^U - \hat{\mu}_t\|_{\Sigma_t} \cdot \|x_{t,a}\|_{\Sigma_t^{-1}} + \|\hat{\mu}_t - \mu^\star\|_{\Sigma_t} \cdot \|x_{t,a}\|_{\Sigma_t^{-1}}\right]\right]$$

$$\leq 2T\mathbb{E}_{\mu^\star \sim \mathcal{N}(0,I)}\left[\mathbf{1}\{\overline{\mathcal{E}}_1 \text{ and } \mathcal{E}_2 \text{ and } \mathcal{E}_3\}(\sqrt{d\ln(1 + T/d) + 2\ln(T^2\pi^2/3\delta)} + \|\mu^\star\|_2)\right]$$

$$\leq 4Td\exp(-W^2/2d)\sqrt{d\ln(1 + T/d) + 2\ln(T^2\pi^2/3\delta)} + 2Td\sqrt{2\exp(-W^2/2d)}.$$

The second inequality holds because $\|x_{t,a}\|_{\Sigma^{-1}} \leq 1$ and by the results of the confidence analysis in Section A.6.1. The last inequality comes from the analysis of term c.

So

$$
\begin{aligned}
\text{BayesRR}_T \leq &2\left(W + \sqrt{d\ln(1+T/d) + 2\ln(T^2\pi^2/3\delta)}\right)\left(\sqrt{2Td\ln(1+\frac{T}{d})} + 2\sqrt{2T\ln(4/\delta)}\right) \\
&+ 4Td\exp(-W^2/2d)\sqrt{d\ln(1+T/d) + 2\ln(T^2\pi^2/3\delta)} + 2Td\sqrt{2\exp(-W^2/2d)} \\
&+ \delta T(2d\sqrt{2\exp(-W^2/2d)} + 2W)
\end{aligned}
$$

Let $W = \sqrt{2d\ln(Td)}$ and $\delta = 1/T$, we have that

$$
\text{BayesRR}_T = \widetilde{O}\left(d\sqrt{T}\right)
$$

■

A.9. Proof of Theorem 3.3.1

Proof. We can convert a stochastic MAB instance into a stochastic linear bandit instance by constructing K K-dimensional basis vectors, each representing an arm. Then the upper bounds derived for linear bandits also hold for MAB. The $\widetilde{O}\left(L\sqrt{KT}/\gamma\right)$ fairness regret upper bound follows. ■

A.10. Proof of Theorem 3.3.2

Proof. We can use the confidence sequence in the FairX-UCB algorithm and apply Lemma A.8.1 to get the $\widetilde{O}\left(\sqrt{KT}\right)$ reward regret upper bound similarly as in the proof of Theorem 4.3.2. ■

B. Additional Experiments

In this section, we present additional experiment results to illustrate the effectiveness of different algorithms across datasets and merit functions.

B.1. Additional Experiment Setup

We conducted experiments on simulation data from the mediamill dataset (Snoek et al., 2006) in addtion to the yeast dataset. The mediamill dataset consists of $43,907$ examples. Each example has 120 features and belongs to one or multiple of the 101 classes. Similar to preparing the yeast dataset, we randomly split the dataset into two sets, 20% as the validation set to tune hyper-parameters and 80% as the test set to test the performance of different algorithms. All the details of the experiments are the same as the experiments on the yeast dataset as introduced in Section 5.1.

The ranges of the hyper-parameters we searched for each bandit algorithm on the two simulation datasets are as follows. For FairX-UCB and UCB, we grid search $w_0 \in [1e{-}5, 2e{-}5, 5e{-}5, 1e{-}4, 2e{-}4, 5e{-}4, 1e{-}3, 2e{-}3, 5e{-}3, 1e{-}2, 2e{-}2, 5e{-}2, 1e{-}1]$. For FairX-TSand TS, we grid search the normal prior standard deviation in $[1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 1e{-}1, 1]$ and reward standard deviation in $[1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 1e{-}1, 1]$. For FairX-EG, we grid search $\epsilon \in [0, 1e{-}4, 1e{-}3, 1e{-}2]$. For FairX-LinUCB and LinUCB, we grid search $\beta \in [0.01, 0.1, 1.0, 10.0]$ and $\lambda \in [1, 1e1, 1e2, 1e3, 1e4, 1e5]$. For FairX-LinTS and LinTS, we grid search prior standard deviation in $[1e{-}6, 1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 1e{-}1]$ and reward standard deviation in $["1e{-}6", "1e{-}5", "1e{-}4", "1e{-}3", "1e{-}2", "1e{-}1"]$. For FairX-EG, we grid search $\epsilon \in [0, 1e{-}4, 1e{-}3, 1e{-}2]$ and the regularization parameter of the ridge regression in $[1e1, 1e2, 1e3, 1e4, 1e5]$.

For the experiments on the Yahoo! dataset, in addition to the results where we select hyper-paramters on the logs in the first day and test the performance on the second day, we here present the results where we select hyper-parameters on the first 5 days and test the performance on the rest 5 days. The whole Yahoo! dataset contains $45,811,883$ events and $1,633,488$ clicks on 271 articles.

The ranges of the hyper-parameters we searched for each bandit algorithm on the Yahoo! dataset are as follows. For FairX-UCB and UCB, we grid search $w_0 \in [1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 1e{-}1]$. For FairX-TS and TS, we grid search

the normal prior standard deviation in $[0.01, 0.1]$ and reward standard deviation in $[1e-5, 1e-4, 1e-3, 1e-2, 1e-1]$. For FairX-EG, we grid search $\epsilon \in [0, 1e-3, 1e-2]$. For FairX-LinUCB and LinUCB, we grid search $\beta \in [1, 10, 100]$ and $\lambda \in [1e-6, 1e-4, 1e-2, 1]$. For FairX-LinTS and LinTS, we grid search prior standard deviation in $[1e-4, 1e-3, 1e-2]$ and reward standard deviation in $[1e-4, 1e-3, 1e-2]$. For FairX-LinEG, we grid search $\epsilon \in [0, 1e-4, 1e-3]$ and the regularization parameter of the ridge regression in $[1e-6, 1e-4, 1e-2, 1]$.
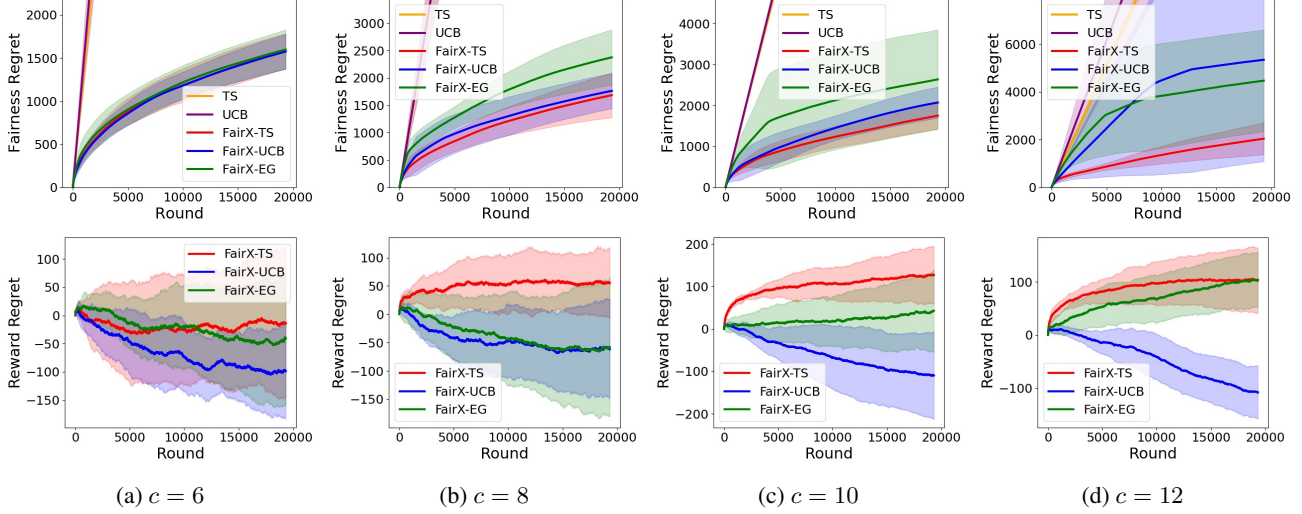
Figure 5. Experiment results on the yeast dataset with varying merit function parameter c for different MAB algorithms.
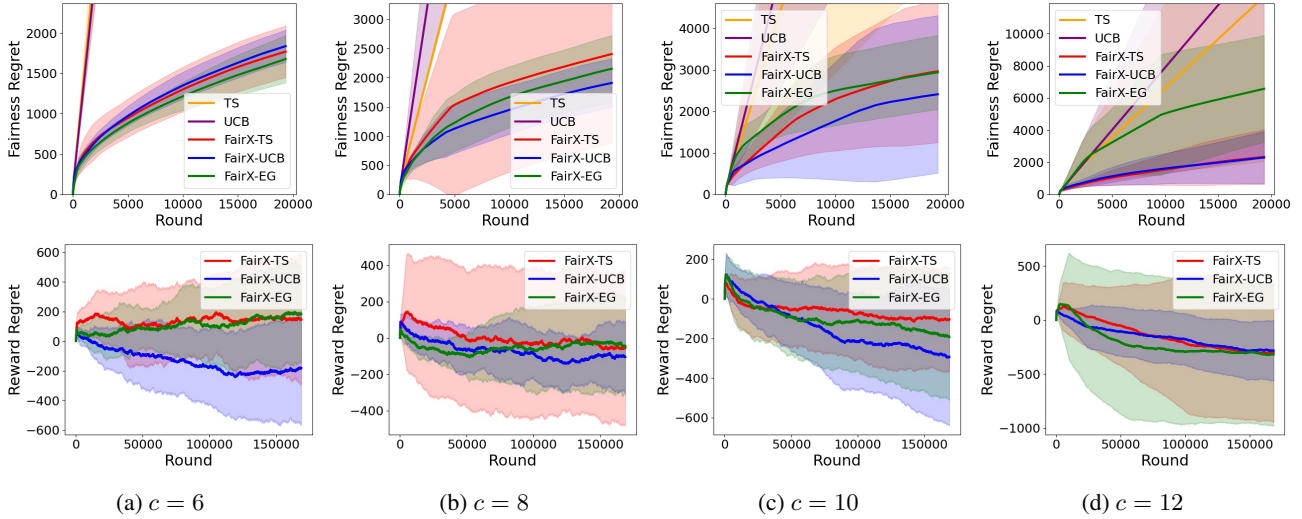
Figure 6. Experiment results on the mediamill dataset with varying merit function parameter c for different MAB algorithms.

B.2. Additional Experiment Results

We show the experiment results with varying merit function parameter c in the FairX MAB setting and FairX linear bandits setting on the yeast and mediamill datasets in Figure 5, Figure 6, Figure 7, and Figure 8. All the FairX algorithms can effectively control merit-based fairness of exposure while achieving low reward regret. As the merit function becomes steeper, the variance of the runs and the difference between algorithms become larger.

The experiment results on the Yahoo! dataset are shown in Figure 9. Though the regrets occasionally jump a bit (which might be due to the change of candidate articles or user interests), all the FairX algorithms can robustly control merit-based fairness of exposure while maintaining low reward regret.
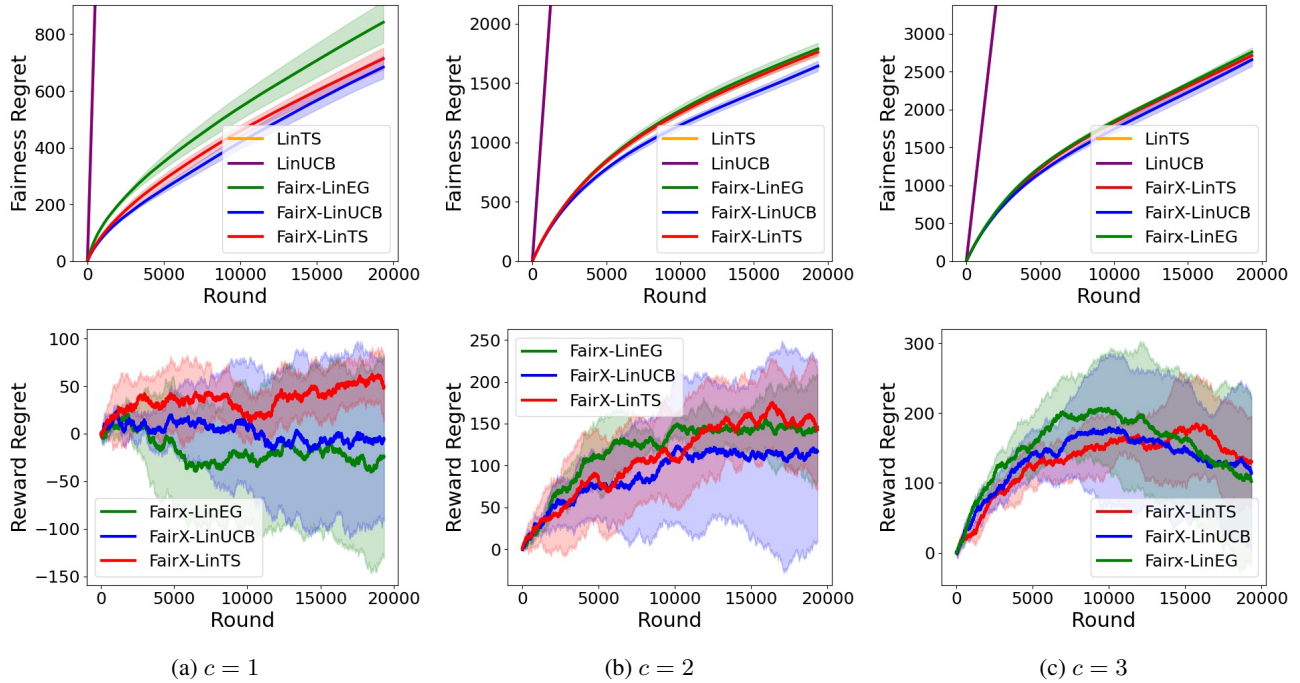
Figure 7. Experiment results on the yeast dataset with varying merit function parameter c for different linear bandit algorithms.
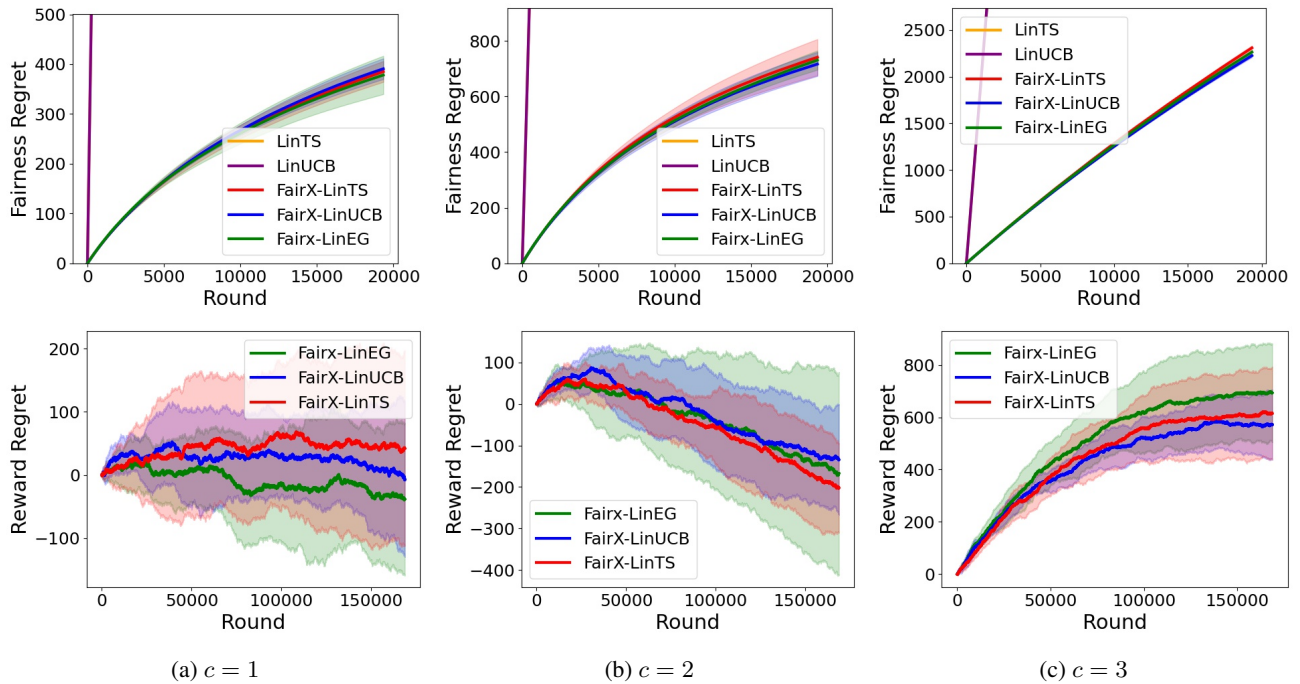

Figure 8. Experiment results on the mediamill dataset with varying merit function parameter c for different linear bandit algorithms.

B.3. Experiment for Linear Bandits with a Well-Specified Linear Model

As discussed in the main paper, the fairness regret of FairX linear bandit algorithms do not seem to converge because the linearity assumption does not necessarily hold for any of the datasets. In this section, we will create a dataset with a well-specified linear model and see how these bandit algorithms perform on this data. We perform the experiments on the
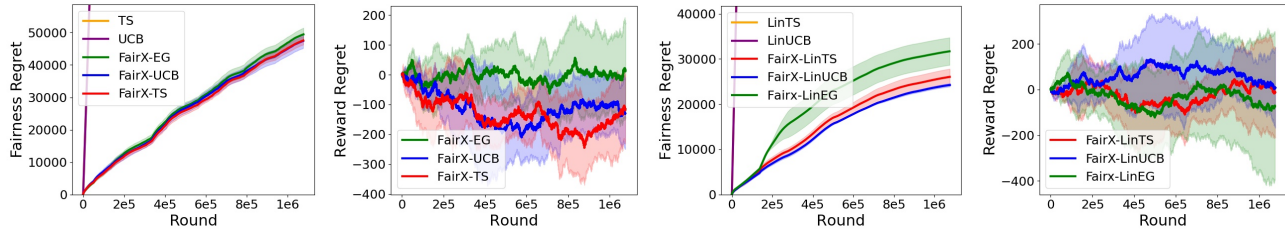
Figure 9. Experiment results on the Yahoo! dataset (select hyper-parameters on data from the first 5 days and report the performance on the data from the last 5 days) for both MAB and linear bandits setting. ($c = 10$ for both settings)
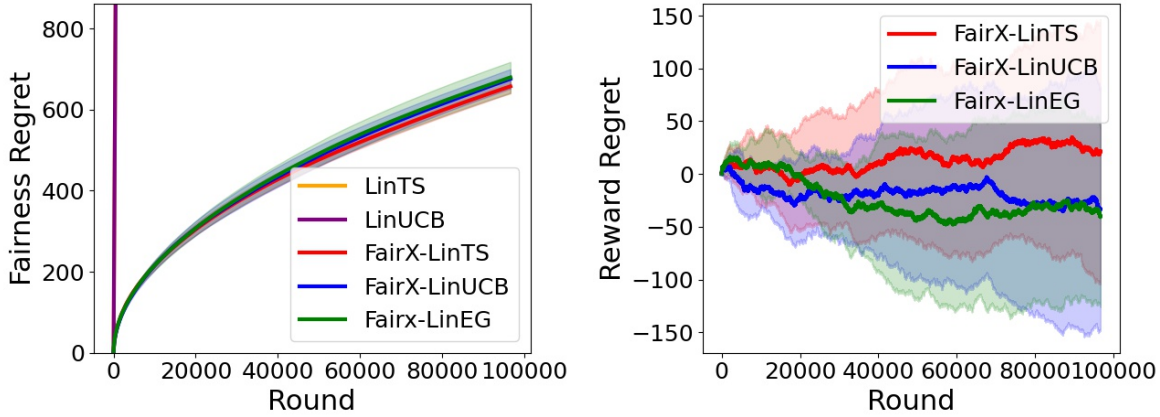

Figure 10. Experiment results on the yeast dataset with a well-specified linear model for linear bandit algorithms. ($c = 2$)

yeast dataset. To remove model misspecification, we do not use the class labels as rewards, but instead inpute the linear least-squares solution on the full-information data as rewards (plus Gaussian noise $\mathcal{N}(0, 0, 1)$). The results are shown in Figure 10. We can see that the fairness regret of the FairX algorithms converge better and follow the $\sqrt{T}$ bound predicted by the theoretical analysis in the absence of model mis-specification.