Fairness of Exposure in Stochastic Bandits

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 shown to 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 setup and present algorithms for both stochastic multi-armed bandits and stochastic linear bandits. We prove that the algorithms achieve sublinear fairness regret and reward regret. Beyond the theoretical analysis, we also provide empirical evidence that these algorithms can 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. 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 users response (e.g. purchase, stream, click), and then updates its policy. Over time, the bandit algorithm thus learns to maximize the user responses, which is 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 user and the platform 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, and it is well understood that algorithms that maximize user responses can be unfair in how they allocate exposure to the items (Singh & Joachims, 2018) – and thus their ability to get purchased, streamed or read. In particular, two items with very similar merit (e.g. click probability) can 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. sellers retention (Mehrotra et al., 2018), anti-discrimination (Noble, 2018), anti-polarization (Epstein & Robertson, 2015)).

To illustrate the problem, consider a conventional (non-personalized) k-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 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 can polarize a user by quickly learning which type of article the user is most likely to click, and then exclusively recommending 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). 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 in each time step. Based on these quantities, we then formulate the reward regret and the fairness regret so that minimizing these two regrets corresponds 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 fair linear TS algorithm in the stochastic linear bandit setting. We prove that all algorithms achieve fairness regret

and reward regret with sublinear dependence on the number of rounds, while the TS algorithms have computational advantages. The fairness regrets of these algorithms also depend 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 effectively allocate exposure to different arms while maximizing utility.

2. Related Work

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

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 best arm while giving zero exposure to all the other arms. Another type of fairness definition in bandits is to ensure a minimum and maximum amount of exposure to each arm or arm group (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) where the probability that any two arms are selected is bounded by the distance between their context vectors in adversarial linear bandit. They require additional feedback of fairness constraints violations. We work in the stochastic bandit setting and we do not require any additional feedback beyond the reward. We also ensure similar items obtain similar exposure, but we focus on similarity that is defined by the how much two items are close in mean reward conditioned on context.

The most relevant work is (Liu et al., 2017) which considers fairness in stochastic MAB problems where the reward distribution is Bernoulli. They proposed calibrated fairness where each arm is selected with the probability equal to that of its utility being the largest. They proposed a Thompson Sampling based algorithm that achieves $T^{2/3}$ regret upper bound where T is the number of rounds. Our formulation is more general in a sense that we consider arbitrary reward distribution and merit function with their formulation as a special case. Also, we consider fairness regret across rounds instead of a fairness constraint for each around which enables us to achieve the $\sqrt{T}$ reward regret upper bound. In addition, we also 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 resource to different agents in a fair way. In our problem, we aim to allocate the user 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, fairness of exposure has been studied in ranking in the statistical learning framework (Singh & Joachims, 2018; 2019). We study fairness of exposure in the online learning setup.

3. Stochastic Multi-Armed Bandits in the FairX Setting

We begin by introducing the FairXsetting for stochastic MAB, presenting 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 by 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 actions based on the interaction history before round t. The learner then samples and executes an action $a_t \sim \pi_t$. In response to the action, the environment samples a reward $r_{t,a_t} \in \mathbb{R}$ 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, taken actions, and their associated rewards. Conventionally, the goal of learning is to maximize the cumulative expected reward $\sum_{t=1}^T \mathbb{E}_{a\sim\pi_t}\mu_{a_t}^\star$. 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 a merit-based fairness-of-exposure constraint (Singh & Joachims, 2018) into the bandits objective. Specifically, we aim to learn a policy $\pi^\star$ that ensures each arm receives an amount of exposure proportional to its merit. 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 provide a design choice that permits tailoring the fairness criterion to different applications. For any choice of $f(\cdot) > 0$, it is easy to see that only the policy $\pi^\star$

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

fulfills the fairness constraints and thus provides maximal reward subject to this constraint. We refer to $\pi^\star$ as the optimal fair policy. TJ: should we state this as a lemma?

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

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

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*, which measures the cumulative ℓ^1 distance between the deployed policy and the optimal fair policy at horizon T

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

The fairness regret and the reward regret depend on both the randomly sampled rewards, as well as the actions 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 which are necessary for designing FairX bandit algorithms with sublinear fairness regret.

Condition 3.1.1 (Minimum Merit)**.** *The merit of each arm is positive, i.e.* $\min_\mu f(\mu) \geq \gamma$ *for some positive constant* $\gamma > 0$.

Condition 3.1.2 (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.

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 3.1.3 (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* $\inf_\mu f(\mu) = 1/\sqrt{T}$*, such that for any bandit algorithm, there must exist a MAB instance such that the fairness regret is at least* $FR_T \geq 0.03T$.

Theorem 3.1.4 (Lower Bound on Fairness Regret is Linear without Bounded Lipschitz-Continuity Condition)**.** *For time horizon* $T > 0$*, there exists a* $4\sqrt{T}$*-Lipschitz continuous merit function* f *with minimum merit* 1*, such that for any bandit algorithm, there must exist a MAB instance such that the fairness regret is at least* $FR \geq 0.03T$.

All the proofs of the theorems in this paper are in Appendix A.

3.2. FairX-UCB Algorithm

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 like the conventional UCB algorithm. At each round t, the algorithm constructs a confidence region CR_t with a confidence width constant w_0 which contains the true parameter with high probability. Then the algorithm optimistically selects a parameter

Algorithm 1 FairX-UCB Algorithm

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

$\mu_t \in \mathbb{R}^K$ 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. Finally, we apply the constructed policy, observe the feedback, and update the confidence region estimate.

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

Theorem 3.2.1 (FairX-UCB Fairness Regret)**.** *Under Condition 3.1.1 and 3.1.2, suppose* $\forall t, a : r_{t,a} \in [-1, 1]$*, when* $T > K$*, the fairness regret of the FairX-UCB algorithm is* $FR_T = \widetilde{O}\left(L\sqrt{KT}/\gamma\right)$ *with high probability.*

Theorem 3.2.2 (FairX-UCB Reward Regret)**.** *Suppose* $\forall t, a : r_{t,a} \in [-1, 1]$*, when* $T > K$*, the reward regret of the FairX-UCB algorithm is* $RR_T = \widetilde{O}(\sqrt{KT})$ *with high probability.*

$\widetilde{O}$ ignores logarithmic factors in O. The reward regret of FairX-UCB does not depend on Condition 3.1.1 and 3.1.2 about the merit function f. Thus the reward regret lower bound developed for conventional bandit problem also holds for our setup because the conventional stochastic MAB problem that only minimizes the reward regret is a special case of our setup where we set the merit function f to be an infinitely steep increasing function. So reward regret upper bound of the FairX-UCB algorithm matches the well-known minimax lower bound on the reward regret of $\Omega\left(\sqrt{KT}\right)$ (Auer et al., 2002b) 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. And it further depends on the minimum merit constant γ and the Lipschitz continuity constant L we treat as absolute constants due to Theorem 3.1.3 and Theorem 3.1.4.

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.

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: Derive policy $\pi_t(a) = \frac{f(\mu_{t,a})}{\sum_{a'} f(\mu_{t,a'})}$
5: Sample and select $a_t \sim \pi_t$
6: Observe reward r_{t,a_t}
7: Update posterior $\mathcal{V}_{t+1} = \text{Update}(\mathcal{V}_1, \mathcal{H}_{t+1})$
8: **end for**

3.3. FairX-TS Algorithm

Another approach to designing stochastic bandit algorithms that have 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 Thompson Sampling 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 from the posterior, and constructs a fair policy from the sampled parameter to deploy. Finally, the algorithm observes the feedback and update the posterior distribution about the true parameter.

Following (Russo & Van Roy, 2014), we analyse 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

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

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

We show in the following two theorems that both the Bayesian reward regret and the Bayesian fairness regret of the FairX-TS algorithm are on the same order as the fairness and reward regret of the FairX-UCB algorithm.

Theorem 3.3.1 (FairX-TS Fairness Regret)**.** *Under Condition 3.1.1 and 3.1.2, suppose the prior distribution of the reward mean* $\mu_a^\star$ *of each arm* a *is 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 *BayesFR*$_T = \widetilde{O}\left(L\sqrt{KT}/\gamma\right)$.

Theorem 3.3.2 (FairX-TS Reward Regret)**.** *Suppose the prior distribution of the reward mean* $\mu_a^\star$ *of each arm* a

is standard normal distribution $\mathcal{N}(0,1)$, and $\forall t, a$ $r_{t,a} \sim \mathcal{N}(\mu^\star_a, 1)$, the Bayesian fairness regret of the FairX-TS algorithm at any round T BayesRR$_T = \widetilde{O}\left(\sqrt{KT}\right)$.

4. Stochastic Linear Bandits in the FairX Setting

In this section, we extend the two algorithms introduced in the MAB setup to the more general linear stochastic bandit setup where the learner is provided with contextual information for making decisions. We discuss how the two algorithms can be adapted to this setup 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 bandit instance $v = (P_{x_{t,a}})$ 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 of the reward distribution $P_{x_{t,a}}$ 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

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

is thus a martingale difference sequence, since

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

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

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

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

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

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

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

Algorithm 3 FairX-LinUCB Algorithm

input: β_t, f
initialization: $\Sigma_1 = \mathbf{I}_d$, $b_1 = \mathbf{0}_d$
for $t = 1$ to ∞ **do**
  Observe K context vectors $\mathcal{D}_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,K}) \subset \mathbb{R}^d$
  $\hat{\mu}_t = \Sigma_t^{-1} b_t$ {The ridge regression solution}
  $\text{CR}_t = (\mu : \|\mu - \hat{\mu}_t\|_{\Sigma_t} \leq \sqrt{\beta_t})$
  $\mu_t = \arg\max_{\mu\in\text{CR}_t} \sum_a \frac{f(\mu\cdot x_{t,a})\mu\cdot x_{t,a}}{\sum_{a'} f(\mu\cdot x_{t,a'})}$
  Construct policy $\pi_t(a) = \frac{f(\mu_t\cdot x_{t,a})}{\sum_{a'} f(\mu_t\cdot x_{t,a'})}$
  Sample an action a_t from policy π_t
  Observe reward r_{t,a_t}
  $\Sigma_{t+1} = \Sigma_t + x_{t,a_t}x_{t,a_t}^\top$
  $b_{t+1} = b_t + x_{t,a_t} r_{t,a_t}$
end for

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

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

The lower bounds on the fairness regret derived in Theorem 3.1.3 and Theorem 3.1.4 in the MAB setting also apply to the linear stochastic bandit setting, since we can easily convert a MAB instance into a stochastic linear bandit instance by constructing K K-dimensional basis vectors, each representing one arm. Thus we again employ Condition 3.1.1 and Condition 3.1.2 to design algorithms that have fairness regret with sublinear dependence on the horizon T.

4.2. FairX-LinUCB Algorithm

Similar to the FairX-UCB algorithm, the FairX-LinUCBalgorithm constructs a confidence region CR_t of the true parameter $\mu^\star$ at each round t. The mean of the confidence region 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. We prove the following upper bounds on the fairness regret and reward regret of the FairX-LinUCB algorithm.

Theorem 4.2.1 (Fair LinUCB Fairness Regret)**.** *Under Condition 3.1.1 and 3.1.2, suppose $\forall t, a$ $\|x_{t,a}\|_2 \leq 1$, η_t is 1 sub-Gaussian, $\|\mu^\star\|_2 \leq 1$, with proper choice of β_t, the fairness regret at any round $T > 0$ $\mathbb{E}[FR_T] = \widetilde{O}\left(Ld\sqrt{T}/\gamma\right)$ with high probability.*

Theorem 4.2.2 (Fair LinUCB Reward Regret)**.** *Suppose $\forall t, a$ $\|x_{t,a}\|_2 \leq 1$, η_t is 1 sub-Gaussian, $\|\mu^\star\|_2 \leq 1$, with proper choice of β_t, the reward regret at any round $T > 0$*

Algorithm 4 FairX-LinTSAlgorithm

input: $f, \mathcal{V}_1$
for $t = 1$ to ∞ **do**
  Observe K arms $\mathcal{D}_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,K}) \subset \mathbb{R}^d$
  Sample parameter from posterior $\mu_t \sim \mathcal{V}_t$
  Derive policy $\pi_t(a) = \frac{f(\mu_{t,a} \cdot x_{t,a})}{\sum_{a'} f(\mu_{t,a'} \cdot x_{t,a})}$
  Sample and select $a_t \sim \pi_t$
  Observe reward r_{t,a_t}
  Update posterior $\mathcal{V}_{t+1} = \text{Update}(\mathcal{V}_1, \mathcal{H}_{t+1})$
end for

$\mathbb{E}[RR_T] = \widetilde{O}\left(d\sqrt{T}\right)$ *with high probability.*

Note that the upper bound on the reward regret matches the lower bound on the reward regret $\Omega(d\sqrt{T})$ (Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011) up to logarithmic factors, where the decision set contains an infinite number of arms. However, it is worse than the reward regret bound of some variants of the LinUCB algorithms, which achieve $\widetilde{O}(\sqrt{dT})$ in the classical stochastic linear bandits with K arms (Auer et al., 2002a; Chu et al., 2011). LW: do we need explanation for this? 1. That is the price for fairness. 2. The variants that achieve $\sqrt{dT}$ are complex and are hardly used in practice. 3. The design of these variants use deterministic policies, which can ensure that either each selected action is very uncertain or we are very certain about every action. But we must use stochastic policies to achieve fairness.TJ: It is hard to compare the reward regret subject to fairness to the one without fairness, since they use different optimal policies. So, I would say the price of fairness is the difference between the optimal policies in the two settings more than the difference in convergence rate.

The optimization Step 7 in Algorithm 3, where we need to find a 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 find approximate solutions in our empirical evaluation.

4.3. FairX-LinTSAlgorithm

LW: do we write down the prior and posterior update in the algorithms? 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 from the posterior distribution of the true parameter and derives a fair policy from the sampled parameter. Then the algorithm deploys the policy and observes the feedback for the taken 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. ???), as opposed to the non-convex optimization problem in FairX-LinUCB.

Appropriately extending our definition of Bayesian reward regret and fairness regret

$$\text{BayesRR}_T = \mathbb{E}_{\mu^\star}\left[\mathbb{E}[\text{RR}_T|\mu^\star]\right] \tag{7}$$

$$\text{BayesFR}_T = \mathbb{E}_{\mu^\star}\left[\mathbb{E}[\text{FR}_T|\mu^\star]\right] \tag{8}$$

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

Theorem 4.3.1 (Fair LinTS Fairness Regret)**.** *Under Condition 3.1.1 and 3.1.2, suppose each dimension of the true parameter $\mu^\star$ is 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 fair LinTS algorithm BayesFR* $= \widetilde{O}\left(L\sqrt{dT}/\gamma\right)$.

Theorem 4.3.2 (Fair LinTS Reward Regret)**.** *Suppose each dimension of the true parameter $\mu^\star$ is 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 fair LinTS algorithm 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 improves an order of $\sqrt{d}$ compared to the FairX-LinUCB algorithm. LW: Reasons? Bayesian regret is weaker, we assume a particular prior?

5. Empirical Evaluation

We conduct experiments on two both simulated data from multi-class multi-label classification datasets and a real world news article recommendation dataset to show how the proposed algorithms perform.

1. Epsilon greedy is a very strong baseline. It has the same fairness regret as the other algorithms TJ: comes out of the blue. Though it might have low TJ: ?? reward regret theoretically, it empirically works pretty good.

2. Calibrated fairness baseline. (Liu et al., 2017). LW: 1. do we calculate the exact probability of each arm being selected? or do we calculate the policy constructed from randomly sampled reward? 2. They only work for identity merit function. Do we also compare the algorithms in the identity merit function?

   (a) Do some random exploration.
   (b) Sample mean of each arm from beta posterior.

(c) sample binary reward from Bernoulli distribution with sampled mean.

(d) Randomly select an arm with reward 1.

3. Compare with conventional algorithms that only maximize reward?

Experiment Plan

experiment 1 MAB yeast histogram of each arm selected.

experiment 1

1. Setting:MAB, Linear Bandit
2. Algorithms:
   - MAB: fair UCB, fair TS, fair epsilon greedy, fair CB?, conventional UCB or conventional TS or both?
   - linear bandit: fair LinUCB, fair LinTS, fair Lin epsilon greedy, conventional LinUCB or conventional TS?
3. Merit function: exp? identity?
4. Datasets: yahoo(both MAB and Lin bandit?) yeast, mediamill
5. features:
   - Yahoo: outer product?
   - yeast and mediamill: random fourier feature?

experiment 2

1. setting: MAB and Linear bandit.
2. When we change the weight in the merit function.
3. algorithms: same as exp1.
4. merit function exp(cμ)
5. data: yeast.
6. features: same as before

5.1. Experiments Setup

This subsection gives the details for the experiments.

5.1.1. Simulation Data

In our experiments, We use two multi-class multi-label classification datasets yeast (Horton & Nakai, 1996) and mediamill datasets (Snoek et al., 2006) to simulate bandit data for learning. Dataset details...

How to convert data: We regard each label as an arm. At every round, one data point's feature is revealed to the agent, then the agent selects an arm based on history and its algorithm. The reward is assigned based on this data point's ground truth label, if the data point belongs to the arm (label) selected by the agent, the reward is 1.0 otherwise 0.0. Note that for MAB setting, the data point's feature is not utilized by the agent, but the reward is assigned w.r.t this data point. The data point is showed to the agent sequentially, and if the number of data point is not enough for the total simulation rounds we required, we shuffle the dataset and repeat. Following the procedure above, we can simulate a (linear) MAB setting using multi-label classification dataset. The ground truth...

We randomly split the data into two sets, 20% as validation set to tune hyper-parameter and 80% as test set to test the performance of different algorithms. For the context vectors for the arms, we use 50-dimensional random Fourier features (Rahimi et al., 2007) generated from the outer product between the data feature given by the dataset and on hot representation of each arm.

5.1.2. News Article Recommendation

We test our algorithms on a real world new article recommendation dataset where the stochastic assumption might be violated. Yahoo dataset. (Li et al., 2010). We split 50% for validation and 50% for test temporally. Following (Li et al., 2010), we use the outer product between the user feature and the article feature as the context feature for each article. The ground truth...For evaluation, we use the unbiased evaluation method proposed in (Li et al., 2011).

5.1.3. Experiment details

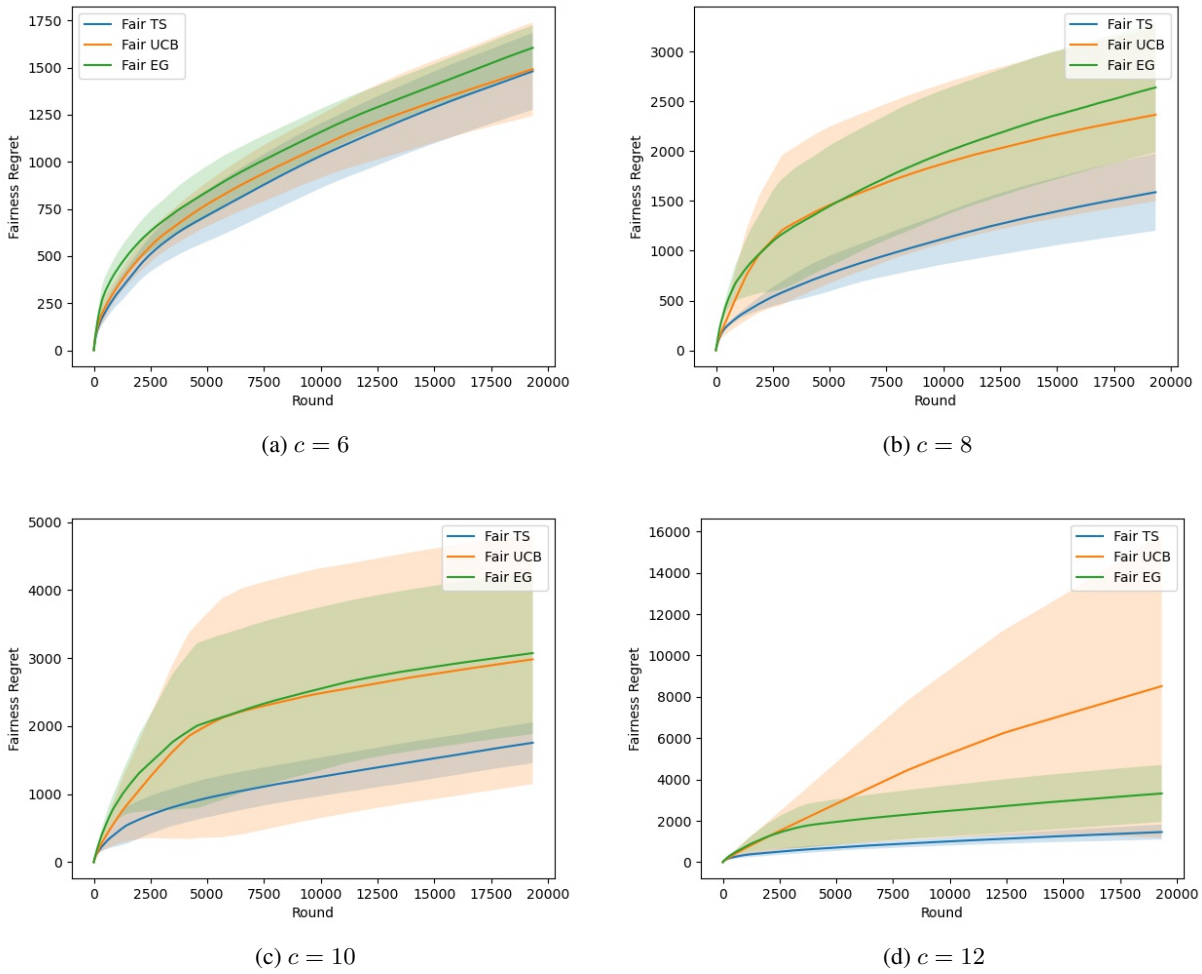
Merit function $f(\mu) = \exp(c\mu)$ where c is a constant to control how steep the merit function is. c determines the quantity of L/γ. We vary c to see how different algorithms perform under different L/γ.

We grid search some hyper-parameters on the validation set for fixed T.

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.



(a) $c = 6$ (b) $c = 8$

(c) $c = 10$ (d) $c = 12$

Figure 1. Fairness regret of different FairX MAB algorithms for different c in the merit function $f(\mu) = \exp(c\mu)$ on Yeast dataset.

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., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

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

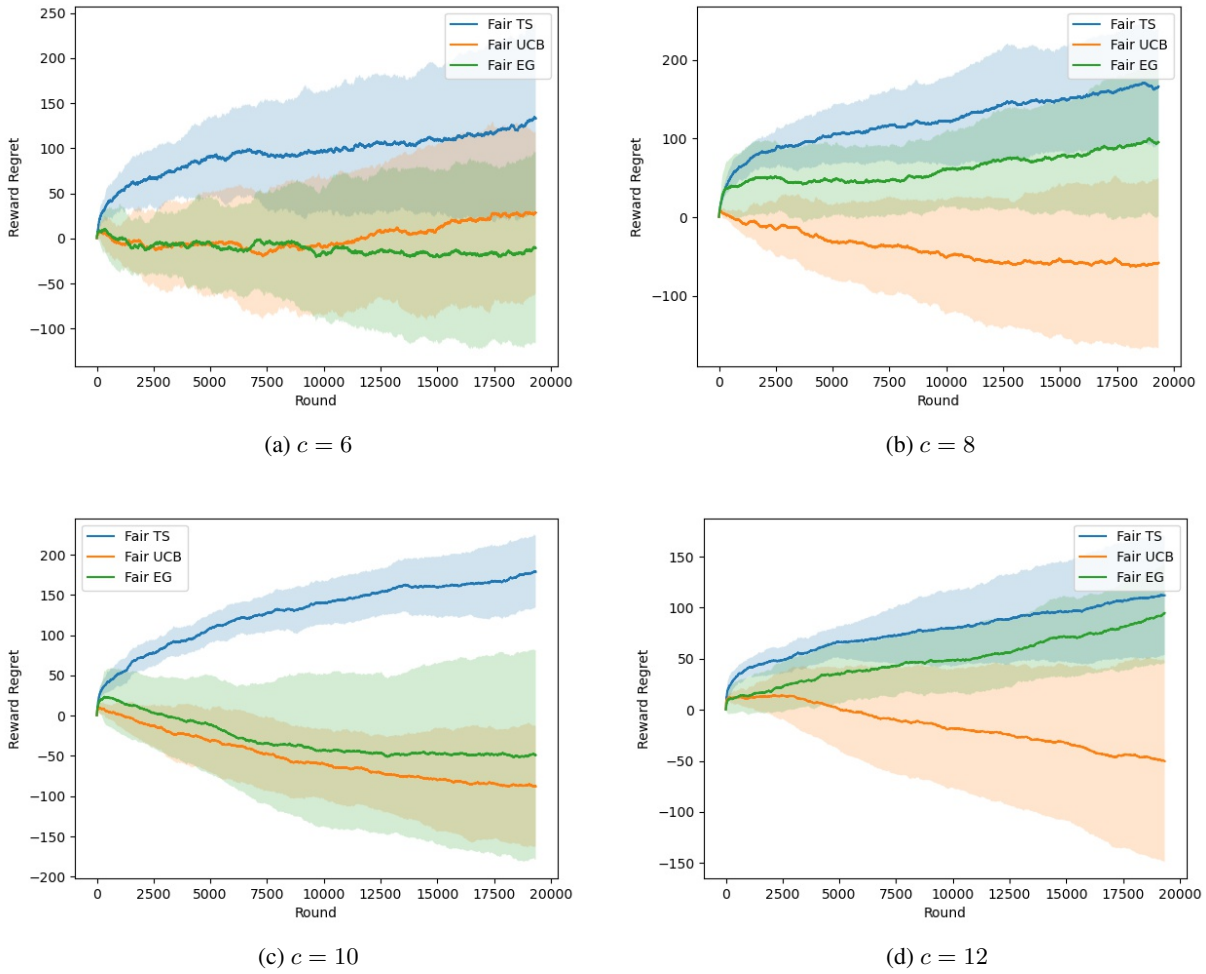
Bechavod, Y., Ligett, K., Roth, A., Waggoner, B., and Wu, S. Z. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems*, pp. 8974–8984, 2019.

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.

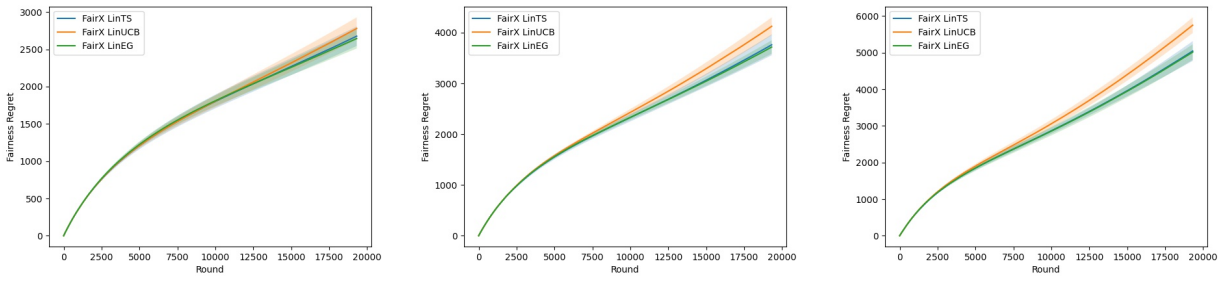
Blum, A., Gunasekar, S., Lykouris, T., and Srebro, N. On preserving non-discrimination when combining expert



(a) $c = 6$ (b) $c = 8$

(c) $c = 10$ (d) $c = 12$

Figure 2. Reward regret of different FairX MAB algorithms for different c in the merit function $f(\mu) = \exp(c\mu)$ on Yeast dataset.



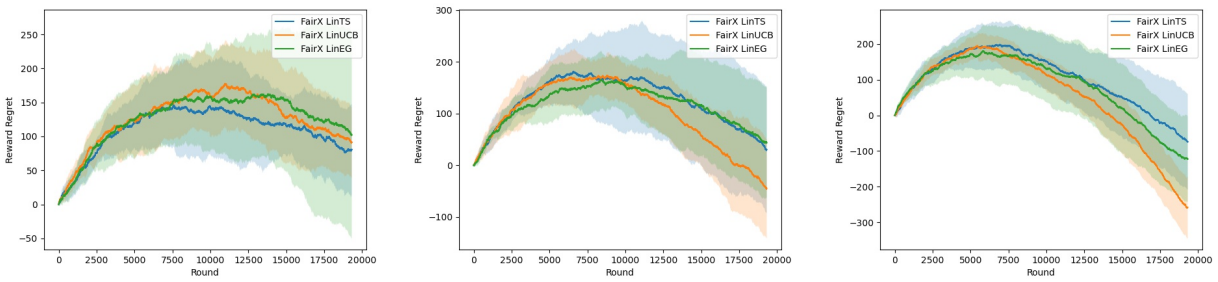
(a) $c = 3$ (b) $c = 4$ (c) $c = 5$

Figure 3. Fairness regret of different FairX linear bandit algorithms for different c in the merit function $f(\mu) = \exp(c\mu)$ on Yeast dataset.

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.



(a) $c = 3$ (b) $c = 4$ (c) $c = 5$

Figure 4. Reward regret of different FairX linear bandit algorithms for different c in the merit function $f(\mu) = \exp(c\mu)$ on Yeast dataset.

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.

Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, 2011.

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 *NIPS*, volume 3, pp. 5. Citeseer, 2007.

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

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

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 3.1.3

Proof. 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 $\mu_1 = (\lambda, 2\lambda)$, i.e. $P_1^1 = \mathcal{N}(\lambda, 1/2)$, $P_2^1 = \mathcal{N}(2\lambda, 1/2)$ and the second instance's mean is $\mu_2 = (2\lambda, 2\lambda)$, i.e. $P_1^2 = \mathcal{N}(2\lambda, 1/2)$, $P_2^2 = \mathcal{N}(2\lambda, 1/2)$, where $\lambda > 0$ 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 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:

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}^1, \mathbb{P}^2\right) &= \mathbb{E}_{\tau\sim\mathbb{P}^1}\left[\ln\frac{\mathbb{P}^1(\tau)}{\mathbb{P}^2(\tau)}\right] = \mathbb{E}_{\tau\sim\mathbb{P}^1}\left[\sum_{t=1}^{T}\ln\frac{P_{a_t}^1(r_{a_t})}{P_{a_t}^2(r_{a_t})}\right] = \sum_{t=1}^{T}\mathbb{E}_{\pi_t\sim\mathcal{A}^1}\mathbb{E}_{a_t\sim\pi_t}\mathrm{KL}\left(P_{a_t}^1, P_{a_t}^2\right) \\
&= \sum_{t=1}^{T}\mathbb{E}_{\pi_t\sim\mathcal{A}^1}\left[\pi_t(1)\mathrm{KL}\left(P_1^1, P_1^2\right)\right] = \sum_{t=1}^{T}\mathbb{E}_{\pi_t\sim\mathcal{A}^1}\left[\pi_t(1)\lambda^2\right] \le \sum_{t=1}^{T}\mathbb{E}_{\pi_t\sim\mathcal{A}^1}\lambda^2 = T\lambda^2,
\end{aligned}$$

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 $\lambda \to 0$, 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 instance v^1, we have:

$$\begin{aligned}
\frac{1}{T}\mathrm{FR}^1(\pi_1, \ldots, \pi_T) &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(|\pi_t(1) - 1/3| + |\pi_t(2) - 2/3|\right)\right] \\
&\ge \mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T}\pi_t(1) - 1/3\right| + \left|\frac{1}{T}\sum_{t=1}^{T}\pi_t(2) - 2/3\right|\right] = 2\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T}\pi_t(1) - 1/3\right|\right].
\end{aligned}$$

Similarly, for instance v^2, we have:

$$\frac{1}{T}\mathrm{FR}^2(\pi_1, \ldots, \pi_T) \ge 2\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T}\pi_1(t) - 1/2\right|\right].$$

Thus we have:

$$\begin{aligned}
\mathrm{FR}^1/T + \mathrm{FR}^2/T &\ge \frac{1}{6}\mathbb{P}^1\left(\frac{1}{T}\sum_{t=1}^{T}\pi_t(1) > \frac{5}{12}\right) + \frac{1}{6}\mathbb{P}^2\left(\frac{1}{T}\sum_{t=1}^{T}\pi_t(1) \le \frac{5}{12}\right) \\
&\ge \frac{1}{12}\exp\left(-\mathrm{KL}\left(\mathbb{P}^1, \mathbb{P}^2\right)\right) \ge \frac{1}{12}\exp\left(-\lambda^2 T\right)
\end{aligned}$$

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

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

$$\mathrm{Regret}_1/T + \mathrm{Regret}_2/T \ge 0.03,$$

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

This concludes the proof. ■

A.2. Proof of Theorem 3.1.4

Proof. 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 $\mu_1 = (\epsilon, 0)$, i.e. $P_1^1 = \mathcal{N}(\epsilon, 1/2)$, $P_2^1 = \mathcal{N}(0, 1/2)$ 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 $\epsilon > 0$ 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 $\pi^{\star,1} = [(L\epsilon+1)/(L\epsilon+2), 1/(L\epsilon+2)]$, 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:

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}^1, \mathbb{P}^2\right) &= \mathbb{E}_{\tau\sim\mathbb{P}^1}\left[\ln\frac{\mathbb{P}^1(\tau)}{\mathbb{P}^2(\tau)}\right] = \mathbb{E}_{\tau\sim\mathbb{P}^1}\left[\sum_{t=1}^T \ln\frac{P^1_{a_t}(r_{a_t})}{P^2_{a_t}(r_{a_t})}\right] = \sum_{t=1}^T \mathbb{E}_{\pi_t\sim\mathcal{A}^1}\mathbb{E}_{a_t\sim\pi_t}\mathrm{KL}\left(P^1_{a_t}, P^2_{a_t}\right) \\
&= \sum_{t=1}^T \mathbb{E}_{\pi_t\sim\mathcal{A}^1}\left[\pi_t(1)\mathrm{KL}\left(P_1^1, P_1^2\right)\right] = \sum_{t=1}^T \mathbb{E}_{\pi_t\sim\mathcal{A}^1}\left[\pi_t(1)\epsilon^2\right] \leq \sum_{t=1}^T \mathbb{E}_{\pi_t\sim\mathcal{A}^1}\epsilon^2 = T\epsilon^2,
\end{aligned}$$

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 instance v^1, we have:

$$\begin{aligned}
\frac{1}{T}\mathrm{FR}^1(\pi_1, \ldots, \pi_T) &= \frac{1}{T}\sum_{t=1}^T \left(|\pi_t(1) - (L\epsilon+1)/(L\epsilon+2)| + |\pi_t(2) - 1/(L\epsilon+2)|\right) \\
&\geq \left|\frac{1}{T}\sum_{t=1}^T \pi_t(1) - (L\epsilon+1)/(L\epsilon+2)\right| + \left|\frac{1}{T}\sum_{t=1}^T \pi_t(2) - 1/(L\epsilon+2)\right| \\
&= 2\left|\frac{1}{T}\sum_{t=1}^T \pi_t(1) - (L\epsilon+1)/(L\epsilon+2)\right|.
\end{aligned}$$

Similarly, for instance v^2, we have:

$$\frac{1}{T}\mathrm{FR}^2(\pi_1, \ldots, \pi_T) \geq 2\left|\frac{1}{T}\sum_{t=1}^T \pi_1(t) - 1/2\right|.$$

Thus we have:

$$\begin{aligned}
\mathrm{FR}^1/T + \mathrm{FR}^2/T &\geq \frac{L\epsilon}{2L\epsilon+4}\mathbb{P}^1\left(\frac{1}{T}\sum_{t=1}^T \pi_t(1) \leq \frac{3L\epsilon+4}{4L\epsilon+8}\right) + \frac{L\epsilon}{2L\epsilon+4}\mathbb{P}^2\left(\frac{1}{T}\sum_{t=1}^T \pi_t(1) > \frac{3L\epsilon+4}{4L\epsilon+8}\right) \\
&\geq \frac{L\epsilon}{4L\epsilon+8}\exp\left(-\mathrm{KL}\left(\mathbb{P}^1, \mathbb{P}^2\right)\right) \geq \frac{L\epsilon}{4L\epsilon+8}\exp\left(-\epsilon^2 T\right)
\end{aligned}$$

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

Set $\epsilon = 1/\sqrt{T}$ and $L = 4\sqrt{T}$, we prove that:

$$\mathrm{Regret}_1/T + \mathrm{Regret}_2/T \geq 0.03,$$

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

This concludes the proof. ■

A.3. Proof of Theorem 3.2.2

Lemma A.3.1. *For $\delta \in (0,1)$, 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.3.2. *For $\delta \in (0,1)$, 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)

For $\delta \in (0,1)$, with probability at least $1-\delta$, the events in Lemma A.3.1 and Lemma A.3.2 hold and

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

The first inequality comes from Line 7 in Algorithm 1. The second inequality comes from Lemma A.3.1. The third inequality comes from Lemma A.3.2 and The last inequality applies the AM-GM inequality.

So the expected reward regret

$$
\mathbb{E}[\mathrm{RR}_T] \le 2\delta T + 2K + 2\sqrt{2\ln(4TK/\delta)}\left(\sqrt{2T\ln(4/\delta)} + 2\sqrt{TK}\right)
$$

Let $\delta = \frac{1}{T}$, we have when $T > K$

$$
\mathbb{E}[\mathrm{RR}_T] = O\left(\sqrt{TK\ln(TK)} + \sqrt{T\ln(T)\ln(TK)}\right) = \widetilde{O}\left(\sqrt{TK}\right)
$$

■

A.4. Proof of Theorem 3.2.1

Proof. For $\delta \in (0,1)$, with probability at least $1-\delta$, the events in Lemma A.3.1 and Lemma A.3.2 hold and

At each time step $t > K$

$$
\begin{aligned}
&\sum_{a=1}^{K} |\pi_t(a) - \pi^\star(a)| \\
=&\sum_{a=1}^{K} \left| \frac{f(\mu_{t,a})}{\sum_{a'=1}^{K} f(\mu_{t,a})} - \frac{f(\mu_a^\star)}{\sum_{a'=1}^{K} f(\mu_{a'}^\star)} \right| \\
=&\sum_{a=1}^{K} \frac{\left| f(\mu_{t,a}) \sum_{a'=1}^{K} f(\mu_{a'}^\star) - f(\mu_a^\star) \sum_{a'=1}^{K} f(\mu_{t,a'}) \right|}{\sum_{a'=1}^{K} f(\mu_{t,a}) \sum_{a'=1}^{K} f(\mu_{a'}^\star)} \\
=&\sum_{a=1}^{K} \frac{\left| f(\mu_{t,a}) \sum_{a'=1}^{K} f(\mu_{a'}^\star) - f(\mu_a^\star) \sum_{a'=1}^{K} f(\mu_{a'}^\star) + f(\mu_a^\star) \sum_{a'=1}^{K} f(\mu_{a'}^\star) - f(\mu_a^\star) \sum_{a'=1}^{K} f(\mu_{t,a'}) \right|}{\sum_{a'=1}^{K} f(\mu_{t,a}) \sum_{a'=1}^{K} f(\mu_{a'}^\star)} \\
\leq&\frac{\sum_{a=1}^{K} |f(\mu_{t,a}) - f(\mu_a^\star)| \sum_{a'=1}^{K} f(\mu_{a'}^\star) + \sum_{a=1}^{K} f(\mu_a^\star) \sum_{a'=1}^{K} |f(\mu_{a'}^\star) - f(\mu_{t,a})|}{\sum_{a'=1}^{K} f(\mu_{t,a}) \sum_{a'=1}^{K} f(\mu_{a'}^\star)} \\
=&\frac{2 \sum_{a=1}^{K} |f(\mu_{t,a}) - f(\mu_a^\star)|}{\sum_{a'=1}^{K} f(\mu_{t,a})} \\
=&\frac{2 \sum_{a=1}^{K} \frac{f(\mu_{t,a})}{f(\mu_{t,a})} |f(\mu_{t,a}) - f(\mu_a^\star)|}{\sum_{a'=1}^{K} f(\hat{\mu}_{t,a})} \\
\leq&\sum_{a=1}^{K} \frac{4L\pi_t(a)}{\gamma} \sqrt{2\ln(4TK/\delta)/N_{t,a_t}} \\
=&\frac{4L\sqrt{2\ln(4TK/\delta)}}{\gamma} \mathbb{E}_{a\sim\pi_t} \left[1/\sqrt{N_{t,a_t}} \right]
\end{aligned}
$$

The second inequality comes from lemma A.3.1. And by lemma A.3.2,

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

So the expected fairness regret

$$
\mathbb{E}\left[\mathrm{FR}_T\right] \leq 2\delta T + 2K + \frac{4L\sqrt{2\ln(4TK/\delta)}}{\gamma} \left(\sqrt{2T\ln(4/\delta)} + 2\sqrt{TK} \right)
$$

Let $\delta = \frac{1}{T}$, we have for $T > K$

$$
\mathbb{E}[\mathrm{FR}_T] = O\left(\frac{L}{\gamma} \sqrt{TK\ln(TK)} + \frac{L}{\gamma} \sqrt{T\ln(T)\ln(TK)} \right) = \widetilde{O}\left(L\sqrt{TK}/\gamma \right)
$$

■

A.5. Proof of Theorem 4.2.1

Proposition A.5.1. *(Confidence) For $\delta \in (0,1)$, with probability at least $1 - \delta/2$, $\forall t$, $\mu^\star \in CR_t$.*

Section A.5.1 is devoted to establishing this confidence bound.

Proposition A.5.2. *Let*

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

denote the instantaneous reward regret acquired by the algorithm on round t. For the fair 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.5.3 (Lemma 7 in (Dani et al., 2008))**.** *For the fair 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

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

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

Lemma A.5.4. *For the fair 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

$$x_{t,a_{t,a}} - \mathbb{E}_{a\sim\pi_t} x_{t,a}$$

is a martingale difference sequence and $\forall t$

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

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.5.5. *For the fair 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.

$$
\begin{aligned}
& fr_t \\
=& \sum_a \left| \frac{f(\mu^\star \cdot x_{t,a})}{\sum_{a'} f(\mu^\star \cdot x_{t,a'})} - \frac{f(\mu_t \cdot x_{t,a})}{\sum_{a'} f(\mu_t \cdot x_{t,a'})} \right| \\
=& \sum_a \left| \frac{f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu_t \cdot x_{t,a'}) - f(\mu_t \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'})}{\sum_{a'} f(\mu^\star \cdot x_{t,a'}) \sum_{a'} f(\mu_t \cdot x_{t,a'})} \right| \\
=& \sum_a \left| \frac{f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu_t \cdot x_{t,a'}) - f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'}) + f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'}) - f(\mu_t \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'})}{\sum_{a'} f(\mu^\star \cdot x_{t,a'}) \sum_{a'} f(\mu_t \cdot x_{t,a'})} \right| \\
\leq& \sum_a \frac{|f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu_t \cdot x_{t,a'}) - f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'})| + |f(\mu^\star \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'}) - f(\mu_t \cdot x_{t,a}) \sum_{a'} f(\mu^\star \cdot x_{t,a'})|}{\sum_{a'} f(\mu^\star \cdot x_{t,a'}) \sum_{a'} f(\mu_t \cdot x_{t,a'})} \\
\leq& \frac{2 \sum_a |f(\mu^\star \cdot x_{t,a}) - f(\mu_t \cdot x_{t,a})|}{\sum_{a'} f(\mu_t \cdot x_{t,a'})} \\
=& 2 \sum_a \frac{\pi_t(a)}{f(\mu_t \cdot x_{t,a})} |f(\mu^\star \cdot x_{t,a}) - f(\hat{\mu}_t \cdot x_{t,a}) + f(\hat{\mu}_t \cdot x_{t,a}) - f(\mu_t \cdot x_{t,a})| \\
\leq& \frac{4L}{\gamma} \mathbb{E}_{a \sim \pi_t} \left[\|\mu_t - \hat{\mu}_t\|_{\Sigma_t} \|x_{t,a}\|_{\Sigma_t^{-1}} + \|\mu - \hat{\mu}_t\|_{\Sigma_t} \|x_{t,a}\|_{\Sigma_t^{-1}} \right] \\
\leq& \frac{4L\sqrt{\beta_t}}{\gamma} \mathbb{E}_{a \sim \pi_t} w_{t,a}
\end{aligned}
$$

So by lemma A.5.4,

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

■

Lemma A.5.6 (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.5.7. $\forall t$, $\det \Sigma_{t+1} \leq (1 + t/d)^d$

Proof.

$$
\begin{aligned}
\text{Trace } \Sigma_{t+1} &= \text{Trace } \left(I + \sum_{\tau=1}^{t} x_{\tau,a_\tau} x_{\tau,a_\tau}^\top \right) \\
&= d + \sum_{\tau=1}^{t=1} \text{Trace } \left(x_{t,a_t} x_{t,a_t}^\top \right) \\
&= d + \sum_{\tau=1}^{t} \|x_{t,a_t}\|_2^2 \\
&\leq d + t
\end{aligned}
\tag{11}
$$

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.5.8. *We have for all t,*

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

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. ■

Proof. (Proof of Proposition A.5.2)

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

■

Proof. (Proof of theorem 4.2.1)

By Proposition A.5.1 and Proposition A.5.2, with probability at least $1-\delta$,

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

So set $\delta = 1/T$, we have

$$\mathbb{E}[\text{FR}_T] \leq 2T/T + \frac{4L\sqrt{\beta_T}}{\gamma}\sqrt{2Td\ln(1+\frac{t}{d})} + \frac{4L\sqrt{\beta_T}}{\gamma}\sqrt{2T\ln(4/\delta)} = \widetilde{O}\left(Ld\sqrt{T}/\gamma\right)$$

■

A.5.1. CONFIDENCE ANALYSIS

In this section, we prove Proposition A.5.1, 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.5.1)

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), we have with

probability at least $1-\delta_t$,

$$\begin{aligned}\sqrt{(\hat{\mu}_t-\mu^\star)^\top\Sigma_t(\hat{\mu}_t-\mu^\star)} &= \|\Sigma_t^{1/2}(\hat{\mu}_t-\mu^\star)\|_2 \\ &\le \|\Sigma_t^{-1/2}\mu^\star\|_2 + \|\Sigma_t^{-1/2}\sum_{\tau=1}^{t-1}\eta_\tau x_{\tau,a_\tau}\|_2 \\ &\le \|\mu^\star\|_2 + \sqrt{\ln(\det(\Sigma_t)\det(\Sigma_1)^{-1}/\delta_t^2)} \\ &\le 1+\sqrt{d\ln(1+t/d)+2\ln(1/\delta_t)}\end{aligned}$$

LW: $\le W+\sqrt{d\ln(1+t/d)+2\ln(1/\delta_t)}$

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

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

■

A.6. Proof of Theorem 4.2.2

Proof. With probability at least $1-\delta$, the events in Proposition A.5.1 and Lemma A.5.4 hold and

$$\begin{aligned}&\mathrm{RR}_T \\ =&\sum_{t=1}^{T}\mu^\star\cdot\mathbb{E}_{a\sim\pi_t^\star}[x_{t,a}]-\mu^\star\cdot\mathbb{E}_{a\sim\pi_t}[x_{t,a}] \\ \le&\sum_{t=1}^{T}\mu_t\cdot\mathbb{E}_{a\sim\pi_t}[x_{t,a}]-\mu^\star\cdot\mathbb{E}_{a\sim\pi_t}[x_{t,a}] \\ =&\sum_{t=1}^{T}(\mu_t-\hat{\mu}_t)\cdot\mathbb{E}_{a\sim\pi_t}[x_{t,a}]+(\hat{\mu}_t-\mu^\star)\cdot\mathbb{E}_{a\sim\pi_t}[x_{t,a}] \\ \le&\sum_{t=1}^{T}\mathbb{E}_{a\sim\pi_t}[2\sqrt{\beta_t}w_{t,a}] \\ \le&2\sqrt{\beta_T}\sum_{t=1}^{T}\mathbb{E}_{a\sim\pi_t}[w_{t,a}] \\ \le&2\sqrt{\beta_T}\left(\sum_{t=1}^{T}w_{t,a_t}+2\sqrt{2T\ln(4/\delta)}\right) \\ \le&2\sqrt{\beta_T}\left(\sqrt{2Td\ln(1+\frac{T}{d})}+2\sqrt{2T\ln(4/\delta)}\right)\end{aligned}$$

The first inequality comes from the algorithm. The second inequality comes from Lemma A.5.3. The third inequality comes from the fact that β_t is increasing. The fourth equality comes from lemma A.5.4. And the last inequality comes from Lemma A.5.8.

Set $\delta=1/T$, the expected reward regret

$$\mathbb{E}[\mathrm{RR}_T]\le 2T/T+2\sqrt{\beta_T}\left(\sqrt{2Td\ln(1+\frac{T}{d})}+2\sqrt{2T\ln(4T)}\right)=\widetilde{O}\left(d\sqrt{T}\right)$$

■

A.7. 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 bandit, the posterior distribution is a Gaussian distribution: $p(\cdot|\mathcal{H}_t) \coloneqq \mathcal{N}\left(\hat{\mu}_t, \Sigma_t^{-1}\right)$. LW: $p(\cdot|\mathcal{H}_t) \coloneqq TN\left(\hat{\mu}_t, \Sigma_t^{-1}, \mathbf{a}_t, \mathbf{b}_t\right)$ 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 LinUCB to upper bound the instantaneous fairness regret as follows (conditioned on history $\mathcal{H}_t$):

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

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 $\mu^\star$:

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

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:

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

Denote the random variable $z_{t,a} \coloneqq (\mu_t - \hat{\mu}_t)\cdot x_{t,a} / \sqrt{x_{t,a}^\top \Sigma_t^{-1} x_{t,a}}$. Given $\mathcal{H}_t$, $z_{t,a}$ is a random variable and is only dependent on μ_t. Note that $z_{t,a} \sim \mathcal{N}(0,1)$, which by the CDF of normal distribution, means that for any $\epsilon > 0$:

$$\mathbb{P}\left(|z_{t,a}| \geq \epsilon\right) \leq \exp(-\epsilon^2/2).$$

Set $\exp(-\epsilon^2/2) = \delta'$, we get that with probability at least $1-\delta'$, we have:

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

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

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

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

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

Below we bound term b above. First note that:

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

where the first inequality uses the fact that $(a+b)^2 \le 2a^2+2b^2$, 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 $\|x_{t,a}\|_2 \le 1$ and $\det(\Sigma_t^{-1}) \le 1$.

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

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

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

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

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

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

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:

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

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:

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

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:

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

Hence, we have:

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

where the last step can be easily derived from the fairness regret bound of LinUCB.

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

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

■

A.7.1. Proof of Theorem 4.3.2

Lemma A.7.1 (Adapted from Proposition 1 from (Russo & Van Roy, 2014))**.** *For any UCB sequence $(U_t : t\in\mathbb{N})$, the Bayesian reward regret of the fair Thompson sampling reward regret*

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

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

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

Summing over T steps concludes the proof. ■

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) We can use the confidence sequence in the proof of LinUCB algorithm where $U_{t,a}\coloneqq\max_{\mu\in\text{CR}_t}\mu\cdot x_{t,a}$ and $B_{t,a}\coloneqq\min_{\mu\in\text{CR}_t}\mu\cdot x_{t,a}$.

LW: Since each dimension of $\mu^\star$ is standard normal distribution, $\mathbb{P}(\|\mu^\star\|_2 \geq W) \leq d\exp(-W^2/d)$

LW: new proof

Denote the event $\|\mu_2^\star\|_2 < W$ as $\mathcal{E}$

By Lemma A.7.1, we have that

$$\begin{aligned}
\text{BayesRR}_T &\leq \mathbb{E}\left[\mathbf{1}\{\mathcal{E}\}\sum_{t=1}^{T}\left[\mathbb{E}_{a\sim\pi_t}\left[U_{t,a} - \mu^\star \cdot x_{t,a}\right]\right]\right] + \mathbb{E}\left[\mathbf{1}\{\mathcal{E}\}\sum_{t=1}^{T}\left[\mathbb{E}_{a\sim\pi_t^\star}\left[\mu^\star \cdot x_{t,a} - U_{t,a}\right]\right]\right] + \mathbb{E}\left[\mathbf{1}\{\bar{\mathcal{E}}\}2T\right] \\
&\leq \mathbb{E}\sum_{t=1}^{T}\left[\mathbf{1}\{\mathcal{E}\}\mathbb{E}_{a\sim\pi_t}\left[U_{t,a} - B_{t,a}\right]\right] + 2\sum_{t=1}^{T}\mathbb{P}(\mu^\star \in \text{CR}_t^c) + 2Td\exp(-W^2/d) \\
&\leq 2(W + \sqrt{d\ln(1+T/d) + 2\ln(T^3\pi^2/3)})\left(\sqrt{2Td\ln(1+\frac{T}{d})} + 2\sqrt{2T\ln(4T)}\right) + 2 + 2Td\exp(-\epsilon^2/d)
\end{aligned}$$

Let $W = \sqrt{d\ln(Td)}$, we have that

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

LW: end new proof

By Lemma A.7.1, we have that

$$\begin{aligned}
\text{BayesRR}_T &= \mathbb{E}\sum_{t=1}^{T}\left[\mathbb{E}_{a\sim\pi_t}\left[U_{t,a} - \mu^\star \cdot x_{t,a}\right]\right] + \mathbb{E}\sum_{t=1}^{T}\left[\mathbb{E}_{a\sim\pi_t^\star}\left[\mu^\star \cdot x_{t,a} - U_{t,a}\right]\right] \\
&\leq \mathbb{E}\sum_{t=1}^{T}\left[\mathbb{E}_{a\sim\pi_t}\left[U_{t,a} - B_{t,a}\right]\right] + 2\sum_{t=1}^{T}\mathbb{P}(\mu^\star \in \text{CR}_t^c) \\
&\leq 2\sqrt{\beta_T}\left(\sqrt{2Td\ln(1+\frac{T}{d})} + 2\sqrt{2T\ln(4T)}\right) + 2 \\
&= \widetilde{O}\left(d\sqrt{T}\right)
\end{aligned}$$

■

A.8. Proof of Theorem 3.3.1

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

A.9. Proof of Theorem 3.3.2

Proof. We can use the confidence sequence in the FairX-UCB algorithm and apply Lemma A.7.1 to get the $\widetilde{O}\left(\sqrt{KT}\right)$ reward regret upper bound similarly as in the proof of Theorem 4.3.2. ■