The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Achieving Fairness in the Stochastic Multi-Armed Bandit Problem

Vishakha Patil, Ganesh Ghalme, Vineet Nair, Y. Narahari
Indian Institute of Science, Bengaluru, India
{patilv, ganeshg, vineet, narahari}@iisc.ac.in

Abstract

We study an interesting variant of the stochastic multi-armed bandit problem, which we call the FAIR-MAB problem, where, in addition to the objective of maximizing the sum of expected rewards, the algorithm also needs to ensure that at any time, each arm is pulled at least a pre-specified fraction of times. We investigate the interplay between *learning* and *fairness* in terms of a pre-specified vector denoting the fractions of guaranteed pulls. We define a *fairness-aware regret*, which we call r-Regret, that takes into account the above fairness constraints and extends the conventional notion of regret in a natural way. Our primary contribution is to obtain a complete characterization of a class of FAIR-MAB algorithms via two parameters: the unfairness tolerance and the learning algorithm used as a black-box. For this class of algorithms, we provide a fairness guarantee that holds uniformly over time, irrespective of the choice of the learning algorithm. Further, when the learning algorithm is UCB1, we show that our algorithm achieves constant r-Regret for a large enough time horizon. Finally, we analyze the *cost of fairness* in terms of the conventional notion of regret. We conclude by experimentally validating our theoretical results.

1 Introduction

The multi-armed bandit (MAB) problem is a classic framework for sequential decision-making in uncertain environments. Starting with the seminal work of Robbins (1952), over the years, a significant body of work has been developed to address both theoretical aspects and practical applications of this problem; see (Bubeck and Cesa-Bianchi 2012; Lattimore and Szepesvári 2018; Slivkins 2019) for textbook expositions of the MAB problem. Indeed, the study of the MAB problem and its numerous variants continues to be a central pursuit in multiple fields such as online learning and reinforcement learning. In the MAB setup, at every round a decision maker (an online algorithm) is faced with k choices, which correspond to unknown (to the algorithm) reward distributions. Each choice is referred to as an arm and when the decision maker pulls a specific arm she receives a reward drawn from the corresponding (a priori unknown) distribution[1]. The goal of the decision maker is to maximize the cumulative reward in expectation accrued through a sequence of arm pulls, i.e. if the process repeats for T rounds then in each round the decision maker selects an arm with the objective of maximizing the total expected reward[2].

Several variations of the MAB problem have been extensively studied in the literature. Various papers study MAB problems with additional constraints which include bandits with knapsack constraints (Badanidiyuru, Kleinberg, and Slivkins 2013), bandits with budget constraints (Xia et al. 2015), sleeping bandits (Kleinberg, Niculescu-Mizil, and Sharma 2010; Chatterjee et al. 2017), etc. In this paper we consider FAIR-MAB, a variant of the MAB problem where, in addition to maximizing the cumulative expected reward, the algorithm also needs to ensure that uniformly (i.e., at the end of every round) each arm is pulled at least a pre-specified fraction of times. This imposes an additional constraint on the algorithm which we refer to as a *fairness constraint*, specified in terms of a vector $r \in \mathbb{R}^k$.

Formally, each component r_i of the given vector r specifies a *fairness-quota* for arm i and the online algorithm must ensure that for all time steps t (i.e. uniformly), each arm i is pulled at least $\lfloor r_i \cdot t \rfloor$ times in t rounds. The goal of the online algorithm is to minimize expected regret while satisfying the fairness requirement of each arm. The expected regret in this setting, which we call r-Regret, is computed with respect to the optimal *fair* policy (see Definition 4). We note that the difficulty of this problem is in satisfying these fairness constraints at the end of every round, which in particular ensures fairness even when the time horizon is unknown to the algorithm beforehand. It is relevant to note that the current work contributes to the long line of work in constrained variants of the MAB problem (Badanidiyuru, Kleinberg, and Slivkins 2013; Kleinberg, Niculescu-Mizil, and Sharma 2010; Xia et al. 2015).

[1] The arms which are not pulled do not give any reward.

[2] We study the standard setup in which T is not known upfront to the online algorithm.

The fairness constraints described above naturally capture many real-world settings wherein the arm pulls correspond to allocation of resources among agents with specified entitlements (quotas). The objective of ensuring a certain minimum allocation guarantee to each individual is, at times, at odds with the objective of maximizing efficiency, the classical goal of any learning algorithm. However, in many applications the allocation rules must consider such constraints in order to ensure fairness. The minimum entitlement over available resources secures the prerogative of individuals. For concreteness, we next present a motivating example.

The US Department of Housing and Urban Development recently sued Facebook for engaging in housing discrimination by targeting ads based on attributes such as gender, race, religion, etc. which are protected classes under the US law[3]. Facebook's algorithm that decides which ad should be shown to a particular user, inadvertently ends up discriminating because of the objective that it is trying to optimize. For example if the algorithm learns that it can generate more revenue by displaying an ad to more number of men as compared to women, then it would end up discriminating against women. The proposed FAIR-MAB model ensures that both men and women are shown the ad for at least a pre-specified fraction of the total number of ad displays, thereby preserving the fundamental right of equal access to opportunities. In a way, the minimum fraction guarantee also provides a moral justification to the chosen allocation rule by evaluating it to be fair under the veil of ignorance (Rawls 1971) in which an allocation rule is considered as a hypothetical agreement among free and equal individuals unaware of the natural capabilities and circumstantial advantages and biases they might have i.e. a socially agreed upon allocation in the original position (refer to (Freeman 2019; Heidari et al. 2018) for a detailed discussion).

The fairness model in this work naturally captures many resource allocation situations such as the sponsored ads on a search engine where each advertiser should be guaranteed a certain fraction of pulls in a bid to avoid monopolization of ad space; crowd-sourcing where each crowd-worker is guaranteed a fraction of tasks in order to induce participation; and a wireless communication setting where the receiver must ensure minimum quality of service to each sender. The work by (Li, Liu, and Ji 2019) contains a detailed discussion of these applications. We discuss other related works on fairness in Section 7.

Our contributions: We first define the FAIR-MAB problem in Section 2. Any FAIR-MAB algorithm is evaluated based on two criteria: the fairness guarantee it provides and its r-Regret. The fairness notion that we consider requires that the fairness constraints be satisfied after each round, and the r-Regret notion is a natural extension of the conventional notion of regret which is defined with respect to an optimal policy that satisfies the fairness constraints. The uniform time fairness guarantee that we seek ensures fairness even in *horizon-agnostic case*, that is when the time horizon T is unknown to the algorithm. We remark that, even when the horizon T is known, the intuitive approach of pulling each arm sufficiently many times to satisfy its fairness constraint does not guarantee fairness at the end of each round. The reader is referred to the Appendix of the extended version[4]. As our primary contribution, in Section 3, we define a class of FAIR-MAB algorithms, called FAIR-LEARN, characterized by two parameters: the unfairness tolerance and the learning algorithm used as a black-box. We prove that any algorithm in FAIR-LEARN satisfies the fairness constraints at any time t. Thus the fairness guarantee for FAIR-LEARN holds uniformly over time, independently of the choice of the learning algorithm. We note here that our meta-algorithm FAIR-LEARN, allows any MAB algorithm to be plugged-in as a black-box. This simple yet elegant framework can be implemented on top of any existing MAB algorithm to ensure fairness with quantifiable loss in terms of regret. The practical applicability of our algorithm is a notable feature of this work.

When the learning algorithm is UCB1, we prove a sub-logarithmic r-Regret bound for the FAIR-UCB algorithm. Additionally, for sufficiently large T we see that the FAIR-UCB incurs constant r-Regret. We then evaluate the cost of fairness in FAIR-MAB with respect to the *conventional* notion of regret in Section 4. We conclude by providing detailed experimental results to validate our theoretical guarantees in Section 6. In particular, we compare the performance of FAIR-UCB with LFG algorithm proposed in (Li, Liu, and Ji 2019), which is the work closest to the current paper. We remark here that we obtain a much stronger fairness guarantee that holds at any time, unlike the asymptotic fairness guarantee of LFG . We also prove a better regret bound with finer dependence on the problem instance parameters. Section 7 provides a detailed comparison.

2 The Model

In this section we formally define the FAIR-MAB problem, the notion of fairness, and the concept of r-regret used in this work.

The FAIR-MAB Problem

An instance of the FAIR-MAB problem is a tuple $\langle T, [k], (\mu_i)_{i\in[k]}, (r_i)_{i\in[k]}\rangle$, where T is the time horizon, $[k] = \{1, 2, \ldots, k\}$ is the set of arms, $\mu_i \in [0, 1]$ represents the mean of the reward distribution $\mathcal{D}_i$ associated with arm i, and $(r_i)_{i\in[k]}$ represents the fairness constraint vector. In the FAIR-MAB setting, the fairness constraints are exogenously specified to the algorithm in the form of a vector $r = (r_1, r_2, \ldots, r_k)$ where $r_i \in [0, 1/k)$, for all $i \in [k]$, and consequently $\sum_{i\in[k]} r_i < 1$ and r_i denotes the minimum fraction of times an arm $i \in [k]$ has to be pulled in T rounds, for any T. We consider $r_i \in [0, 1/k)$ to be consistent with the notion of *proportionality* wherein, guaranteeing any

[3]https://www.technologyreview.com/s/613274/facebook-algorithm-discriminates-ai-bias/

[4]Extended version can be found at https://arxiv.org/abs/1907.10516

arm a fraction greater than its proportional fraction, which is $1/k$, is *unfair* in itself.

In each round t, a FAIR-MAB algorithm pulls an arm $i_t \in [k]$ and collects the reward $X_{i_t} \sim \mathcal{D}_{i_t}$. We assume that the reward distributions are $Bernoulli(\mu_i)$ for each arm $i \in [k]$. This assumption holds without loss of generality since one can reduce the MAB problem with general distributions supported on [0,1] to a MAB problem with Bernoulli rewards using the extension provided in (Agrawal and Goyal 2012). Note that the true value of $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ is *unknown* to the algorithm. Throughout this paper we assume without loss of generality that $\mu_1 > \mu_2 > \ldots > \mu_k$ and arm 1 is called the *optimal* arm. Next, we formalize the notions of fairness and regret used in the paper.

Notion of Fairness

Let $N_{i,t}$ denote the number of times arm i is pulled in t rounds. We first present the definition of fairness proposed by (Li, Liu, and Ji 2019) and then define the stronger notion of fairness considered in this paper.

Definition 1. (Li, Liu, and Ji 2019) A FAIR-MAB algorithm $\mathcal{A}$ is called (asymptotically) fair if $\liminf_{t\to\infty} \mathbb{E}_{\mathcal{A}}\big[r_i - \frac{N_{i,t}}{t}\big] \leq 0$ for all $i \in [k]$.

We refer to the above notion of fairness as *asymptotic fairness*. We now define a much stronger notion of fairness that holds over all rounds and is parameterized by the *unfairness tolerance* allowed in the system which is denoted by a constant $\alpha \geq 0$.

Definition 2. Given an *unfairness tolerance* $\alpha \geq 0$, a FAIR-MAB algorithm $\mathcal{A}$ is said to be α-fair if $\lfloor r_i t \rfloor - N_{i,t} \leq \alpha$ for all $t \leq T$ and for all arms $i \in [k]$.

In particular, if the above guarantee holds for $\alpha = 0$, then we call the FAIR-MAB algorithm *fair*. Note that our fairness guarantee holds uniformly over the time horizon and for any sequence of arm pulls $(i_t)_{t\leq T}$ by the algorithm. Hence it is much stronger than the guarantee in (Li, Liu, and Ji 2019) which only guarantees asymptotic fairness (Definition 1). Notice that for any given constant $\alpha \geq 0$, α-fairness (Definition 2) implies asymptotic fairness.

Notions of Regret

In the MAB setting, the optimal policy is the one which pulls the optimal arm in every round. The regret of a MAB algorithm is defined as the difference between the cumulative reward of the optimal policy and that of the algorithm.

Definition 3. The expected regret of a MAB algorithm $\mathcal{A}$ after T rounds is defined as:

$$\mathcal{R}_{\mathcal{A}}(T) = \sum_{i\in[k]} \Delta_i \cdot \mathbb{E}[N_{i,T}] \tag{1}$$

where $\Delta_i = \mu_1 - \mu_i$ and $N_{i,T}$ denotes the number of pulls of an arm $i \in [k]$ by $\mathcal{A}$ in T rounds.

We call an algorithm optimal if it attains zero regret. It is easy to see that the above notion of regret does not adequately quantify the performance of a FAIR-MAB algorithm as the optimal policy here does not account for the fairness constraints. Also, note that the conventional regret in the FAIR-MAB setting can be $O(T)$ (see Section 4 for further details). Hence, we first state the fairness-aware optimal policy that we consider as a baseline.

Observation 1. A FAIR-MAB algorithm $\mathcal{A}$ is optimal iff $\mathcal{A}$ satisfies the following: if $\lfloor r_i T \rfloor - \alpha > 0$ then $N_{i,T} = \lfloor r_i T \rfloor - \alpha$, else $N_{i,T} = 0$, for all $i \neq 1$.

From Observation 1 we have that an optimal FAIR-MAB algorithm that knows the value of μ must play sub-optimal arms exactly $\lfloor r_i \cdot T \rfloor - \alpha$ times in order to satisfy the fairness constraint and play the optimal arm (arm 1) for the rest of the rounds i.e. for $T - \sum_{i\neq 1} \lfloor r_i \cdot T \rfloor + (k-1)\alpha$ rounds. The regret of an algorithm is compared with such an optimal policy that satisfies the fairness constraints in the FAIR-MAB setting.

Definition 4. Given a fairness constraint vector $r = (r_1, r_2, \ldots, r_k)$ and the *unfairness tolerance* $\alpha \geq 0$, the fairness-aware r-Regret of a FAIR-MAB algorithm $\mathcal{A}$ is defined as:

$$\mathcal{R}^r_{\mathcal{A}}(T) = \sum_{i\in[k]} \Delta_i \cdot \Big(\mathbb{E}[N_{i,T}] - \max\big(0, \lfloor r_i \cdot T \rfloor - \alpha\big)\Big) \tag{2}$$

The $\max(0, \lfloor r_i \cdot T \rfloor - \alpha)$ in the above definition accounts for the number of pulls of arm i made by the optimal algorithm to satisfy its fairness constraint. Also the r-Regret of an algorithm that is not α-fair could be negative but this is an infeasible solution. A learning algorithm that pulls a sub-optimal arm i for more than $\lfloor r_i T \rfloor - \alpha$ rounds, incurs a regret of $\Delta_i = \mu_1 - \mu_i$ for each extra pull. The technical difficulties in designing an optimal algorithm for the FAIR-MAB problem are the conflicting constraints on the quantity $N_{i,T} - \lfloor r_i T \rfloor$ for a sub-optimal arm $i \neq 1$: at any time T for the algorithm to be fair we want $N_{i,T} - \lfloor r_i T \rfloor$ to be at least α whereas to minimize the regret we want $N_{i,T} - \lfloor r_i T \rfloor$ to be close to α.

3 A Framework for Fair MAB Algorithms

In this section, we provide the framework of our proposed class of FAIR-MAB algorithms. Our meta-algorithm FAIR-LEARN is given in Algorithm 1. The key result in this work is the following theorem, which guarantees that FAIR-LEARN is α-fair (see Definition 2), independent of the choice of the learning algorithm LEARN($\cdot$). Note that the fairness guarantee holds uniformly over the time horizon, for any sequence of arm pulls by FAIR-LEARN.

Theorem 1. *For a given $\alpha \geq 0$ and for any given fairness constraint vector $r = (r_1, r_2, \ldots, r_k)$ where $r_i \in [0, \frac{1}{k})$ for all $i \in [k]$,* FAIR-LEARN *is α-fair irrespective of the choice of the learning algorithm* LEARN($\cdot$).

The proof of Theorem 1 is given in Section 5. The guarantee in the above theorem also holds when $\alpha = 0$ and hence

Algorithm 1: Fair-Learn

Input: $[k], (r_i)_{i\in[k]}, \alpha \geq 0$, Learn$(\cdot)$

Initialize:

$N_{i,0} = 0$ for all $i \in [k]$

$S_{i,0} = 0$ for all $i \in [k]$, where $S_{i,t}$ = total reward of arm i in t rounds

for $t = 1, 2, \ldots$ **do**

  Define : $A(t) = \left\{ i \mid r_i \cdot (t-1) - N_{i,t-1} > \alpha \right\}$

  Pull arm

$$i_t = \begin{cases} \text{argmax}_{i\in[k]} \\ \quad (r_i \cdot (t-1) - N_{i,t-1}) & \text{If} A(t) \neq \emptyset \\ \text{Learn}(N_t, S_t) & \text{Otherwise} \end{cases}$$

  Update parameters N_t and S_t

end

Fair-Learn with $\alpha = 0$ is *fair*. In particular, when the learning algorithm Learn$(\cdot)$ = UCB1, we call this algorithm Fair-ucb. We provide the r-Regret bound for Fair-ucb.

Theorem 2. *The r-Regret of* Fair-ucb *is given by*

$$\mathcal{R}^r_{\text{Fair-ucb}}(T) \leq \left(1 + \frac{\pi^2}{3}\right) \cdot \sum_{i\in[k]} \Delta_i + \sum_{\substack{i\in S(T)\\ i\neq 1}} \Delta_i \cdot \left(\frac{8 \ln T}{\Delta_i^2} - \left(r_i \cdot T - \alpha \right) \right)$$

where $S(T) = \left\{ i \in [k] \,\middle|\, r_i \cdot T - \alpha < \frac{8 \ln T}{\Delta_i^2} \right\}$. *In particular for large enough* T, $\mathcal{R}^r_{\text{Fair-ucb}}(T) \leq \left(1 + \frac{\pi^2}{3}\right) \cdot \sum_{i\in[k]} \Delta_i$.

Theorem 2 is proved in Section 5. Observe that if $S(T) \neq \emptyset$ the r-Regret of Fair-ucb is sub-logarithmic and if $S(T) = \emptyset$ then the r-Regret is constant. We prove the distribution-free regret of Fair-ucb in Theorem 3 [5].

Theorem 3. *The distribution-free r-Regret of* Fair-ucb *is* $O(\sqrt{T \ln T})$.

We conclude this section by observing that as the fairness guarantees of Fair-Learn hold without any loss in Learn$(\cdot)$, this framework can easily be made operational in practice.

4 Cost of Fairness

Our regret guarantees until now have been in terms of r-Regret, but now we evaluate the *cost of fairness* in terms of the conventional notion of regret. In particular, we show the trade-off between the conventional regret and fairness in terms of the *unfairness tolerance* .

[5] The proofs of Theorems 3 and 4 can be found in the extended version of this paper.

Theorem 4. *The expected regret of* Fair-ucb *is given by*

$$\mathcal{R}(T) \leq \sum_{i\in S(T)} (r_i \cdot T - \alpha) \cdot \Delta_i + \sum_{\substack{i\in S(T)\\ i\neq 1}} 8 \ln T / \Delta_i + \sum_{i\in[k]} (1 + \pi^2/3) \cdot \Delta_i$$

(where $S(T) = \{ i \mid (r_i \cdot T - \alpha) < 8 \ln T / \Delta_i^2 \}$.)

Theorem 4 captures the explicit trade-off between regret and fairness in terms of the *unfairness tolerance* parameter α. If $S(T) = \emptyset$ we have that the regret is $O(\ln T)$. This implies that if $\alpha > r_i T - 8 \ln T / \Delta_i^2$ for all $i \neq 1$, then the regret is $O(\ln T)$. However, if $S(T) \neq \emptyset$ then for each $i \in S(T)$, an additional regret equal to $r_i T - \alpha$ is incurred in which case the regret is $O(T)$. We complement these results with simulations in Section 6.

5 Proof of Theoretical Results

Proof of Theorem 1

After each round t (and before round $t+1$), we consider the $k+1$ sets, $M_{1,t}, M_{2,t}, \ldots, M_{k,t}$, and S_t, as defined below:

- arm $i \in M_{j,t} \iff \alpha + \frac{(k-j)}{k} \leq r_i t - N_{i,t} < \alpha + \frac{(k-j+1)}{k}, \forall j \in [k]$
- arm $i \in S_t \iff r_i t - N_{i,t} < \alpha$

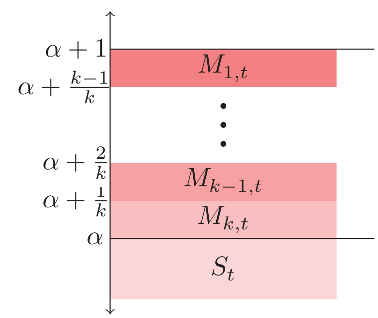

Figure 1: Partition of the arms

Let $V_{j,t} = \uplus_{\ell=1}^{j} M_{\ell,t}$, for all $j \in [k]$. Then the following lemma guarantees the fairness of the algorithm and is at the heart of the proof. The proof of the theorem is immediate from the proof of the lemma.

Lemma 1. *For $t \geq 1$, we have*

1. $V_{k,t} \uplus S_t = [k]$
2. $|V_{j,t}| \leq j$, *for all* $j \in [k]$

Condition 1 in Lemma 1 ensures that at any time $t \geq 1$, the $k+1$ sets $M_{1,t}, M_{2,t}, \ldots, M_{k,t}, S_t$ form a partition of the set $[k]$ of arms. Hence the arm pulled at the $(t+1)$-th round by the algorithm is from one of these $k+1$ sets. As a part of the proof of Lemma 1, in Observation 2 we show that if i_{t+1} is the arm pulled at the $(t+1)$-th round then after $t+1$

rounds $i_{t+1} \in M_{k,t+1} \uplus S_{t+1}$. Also in Observation 3 we show that if an arm $i \in M_{j,t}$ is not pulled in the $(t+1)$-th round then after $t+1$ rounds arm $i \in M_{j-1,t+1} \uplus M_{j,t+1}$ for all $j \in [2,k]$. We note that the two conditions in Lemma 1 are true after the first round, and then the Observations 2 and 3 together ensure that these conditions remain true for all $t > 1$. Hence, all arms $i \in [k]$ satisfy $r_i t - N_{i,t} < \alpha + 1$ for all $t \geq 1$, which implies $\lfloor r_i t \rfloor - N_{i,t} \leq \alpha$. In particular, we have $\lfloor r_i t \rfloor - N_{i,t} \leq \alpha$, for all $t \geq 1$, for all $i \in [k]$, which by Definition 2 proves that FAIR-LEARN is α-fair . $\square$

Proof of Lemma 1: We begin with two complementary observations and then prove the lemma by induction.

Observation 2. Let i be the arm pulled by FAIR-LEARN in round $t+1$.

1. if $i \in S_t$, then $i \in S_{t+1}$
2. if $i \in M_{j,t}$ for some $j \in [k]$, then $i \in M_{k,t+1} \uplus S_{t+1}$

Proof. Case 1: $i \in S_t \implies r_i t - N_{i,t} < \alpha$. Then after round $t+1$, we have

$$\begin{aligned} r_i(t+1) - N_{i,t+1} &= r_i t + r_i - N_{i,t} - 1 \\ &< \alpha - (1 - r_i) \\ &< \alpha \qquad \text{(Since } 1 - r_i > 0) \end{aligned}$$

$$\implies i \in S_{t+1}$$

Case 2: $i \in M_{j,t}$ for some $j \in [k] \implies r_i t - N_{i,t} < \alpha + \frac{(k-j+1)}{k}$. Then after round $t+1$, we have

$$\begin{aligned} r_i(t+1) - N_{i,t+1} &= r_i t + r_i - N_{i,t} - 1 \\ &< \alpha + \frac{(k-j+1)}{k} - (1 - r_i) \\ &< \alpha - \frac{j}{k} + \frac{1}{k} + r_i \\ &< \alpha + r_i \;<\; \alpha + \frac{1}{k} \quad \text{(Since } r_i < \tfrac{1}{k}) \\ \implies i \in M_{k.t+1} \uplus S_{t+1} \end{aligned}$$ $\square$

Observation 3. Let $i \in [k]$ be any arm not pulled at time $t+1$.

1. If $i \in S_t$, then $i \in S_{t+1} \uplus M_{k,t+1}$
2. If $i \in M_{j,t}$ for $j \in [2,k]$, then $i \in M_{j-1,t+1} \uplus M_{j,t+1}$

Proof. Case 1: $i \in S_t \implies r_i t - N_{i,t} < \alpha$. Then after round $t+1$, we have

$$\begin{aligned} r_i(t+1) - N_{i,t+1} &= r_i t - N_{i,t} + r_i \qquad (N_{i,t+1} = N_{i,t}) \\ &< \alpha + r_i \;<\; \alpha + \frac{1}{k} \quad \text{(Since } r_i < \tfrac{1}{k}) \\ &\implies i \in S_{t+1} \uplus M_{k,t+1} \end{aligned}$$

Case 2: $i \in M_{j,t}$ for some $j \in [2,k] \implies \alpha + \frac{k-j}{k} \leq r_i t - N_{i,t} < \alpha + \frac{(k-j+1)}{k}$. Then after round $t+1$, we have

$$\begin{aligned} r_i(t+1) - N_{i,t+1} &= r_i t - N_{i,t} + r_i \\ &< \alpha + \frac{(k-j+1)}{k} + r_i \\ &< \alpha + \frac{(k-j+1)}{k} + \frac{1}{k} \\ &= \alpha + \frac{(k-(j-1)+1)}{k} \end{aligned}$$

and $r_i t - N_{i,t} + r_i \geq \alpha + \frac{k-j}{k} + r_i \;\geq\; \alpha + \frac{k-j}{k}$

$$\implies i \in M_{j-1,t+1} \uplus M_{j,t+1}$$ $\square$

With the above observation we complete the proof of the lemma using induction.

<u>Induction base case ($t = 1$)</u>: Let i_1 be the arm pulled at $t = 1$. Then

$$\begin{aligned} r_{i_1} t - N_{i_1,1} = r_{i_1} - 1 < 0 \leq \alpha \\ \implies i_1 \in S_1 \end{aligned}$$

For all $i \neq i_1$, we have $r_i t - N_{i,1} = r_i < \frac{1}{k} \leq \alpha + \frac{1}{k} \implies i \in S_1 \uplus M_{k,1}$. Hence, $V_{k,1} \uplus S_1 = [k]$, $|V_{k,1}| \leq k-1$, and $|V_{j,1}| = 0$ for all $j \in [k-1]$. Thus, conditions (1) and (2) of the lemma hold.

<u>Inductive Step</u>: Assuming the conditions in the lemma hold after round t, we show that they hold after round $t+1$.

Case 1: $i_{t+1} \in S_t$. From Observation 2, we know $i_{t+1} \in S_{t+1}$. From Observation 3, we know that for any arm $i \neq i_{t+1}$, $i \in S_{t+1} \uplus M_{k,t+1}$. Hence, $V_{k,t+1} \uplus S_{t+1} = [k]$, $|V_{j,t+1}| = 0$ for all $j \in [k-1]$, and $|V_{k,t+1}| \leq k-1$. Thus, Conditions (1) and (2) in the lemma hold after round $t+1$.

Case 2: $i_{t+1} \in M_{a,t}$, for some $a \in [k]$.

$$\begin{aligned} i_{t+1} \in M_{a,t} &\implies i_{t+1} \in V_{a,t} \\ \implies |V_{j,t}| = 0 & \qquad \text{for all } j \in [1, a-1] \text{ if } a > 1 \end{aligned}$$

From Observation 2, we know $i_{t+1} \in S_{t+1} \uplus M_{k,t+1}$, and from Observation 3, we infer that $V_{j-1,t+1} = V_{j,t} \setminus \{i_{t+1}\}$ for all $j \in [2,k]$. Also,

$$\begin{aligned} |V_{j,t} \setminus \{i_{t+1}\}| \leq j - 1 & \qquad \text{for all } j \in [a,k] \\ \implies |V_{j,t+1}| \leq j & \qquad \text{for all } j \in [k] \end{aligned}$$

Also, $V_{k,t+1} \uplus S_{t+1} = [k]$. Hence, Conditions (1) and (2) of the lemma hold after round $t+1$. $\square$

Proof of Theorem 2

The regret analysis of FAIR-UCB builds on the regret analysis of UCB1 which we give in the Appendix of the extended version. In Appendix A we also introduce the notations used in this proof. The UCB1 estimate of the mean of arm i denoted as $\bar{\mu}_i(t) = \hat{\mu}_{i,N_{i,t-1}}(t-1) + c_{t,N_{i,t-1}}$, where $\hat{\mu}_{i,N_{i,t-1}}(t-1)$ is the empirical estimate of the mean of arm i when it is pulled $N_{i,t-1}$ times in $t-1$ rounds and $c_{t,N_{i,t-1}} = \sqrt{\frac{2\ln t}{N_{i,t-1}}}$ is the confidence interval of the arm i at round t. Similar to the analysis of the UCB1 algorithm,

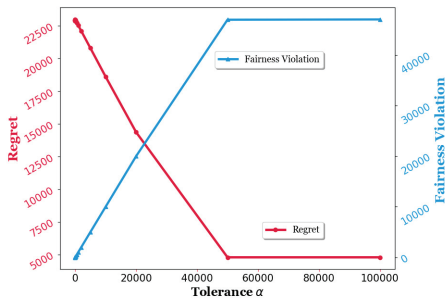

Figure 2: Cost of Fairness

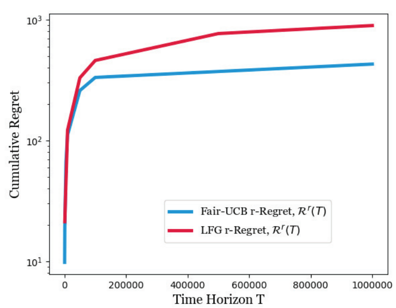

Figure 3: r-Regret Guarantee

we upper bound the expected number of times a sub-optimal arm is pulled. We do this by considering two cases dependent on the number of times the sub-optimal arm is required to be pulled for satisfying its fairness constraint.

Case 1: Let $i \neq 1$ and $r_i \cdot T - \alpha \geq \frac{8 \ln T}{\Delta_i^2}$. Then

$$\begin{aligned}\mathbb{E}[N_{i,T}] \leq & \big(r_i \cdot T - \alpha\big) + \sum_{t=1}^{T} \mathbb{1}\{i_t = i, N_{i,t-1} \geq r_i \cdot T - \alpha\} \\ \leq & \big(r_i \cdot T - \alpha\big) \\ & + \sum_{t=1}^{\infty} \sum_{s_1=1}^{t} \sum_{s_i = r_i \cdot T - \alpha}^{t} \mathbb{1}\Big\{\hat{\mu}_{1,s_1}(t) + c_{t,s_1} \leq \hat{\mu}_{1,s_i}(t) + c_{t,s_i}\Big\}\end{aligned}$$

(Follows from Appendix A, Theorem 6)

Since $r_i \cdot T - \alpha \geq \frac{8 \ln T}{\Delta_i^2}$, it follows from the proof of Theorem 6 in Appendix A that $\mathbb{E}[N_{i,T}] \leq r_i \cdot T - \alpha + \Big(1 + \frac{\pi^2}{3}\Big)$. Hence, $\mathbb{E}[N_{i,T}] - \big(r_i \cdot T - \alpha\big) \leq \Big(1 + \frac{\pi^2}{3}\Big)$.

Case 2: Let $i \neq 1$ and $r_i \cdot T < \frac{8 \ln T}{\Delta_i^2}$

Then the proof of Theorem 6 in Appendix A can be appropriately adapted to show that $\mathbb{E}[N_{i,T}] \leq \frac{8 \ln T}{\Delta_i^2} + \Big(1 + \frac{\pi^2}{3}\Big)$. Hence

$$\mathbb{E}[N_{i,T}] - \big(r_i \cdot T - \alpha\big) \leq \frac{8 \ln T}{\Delta_i^2} + \Big(1 + \frac{\pi^2}{3}\Big) - \big(r_i \cdot T - \alpha\big)$$

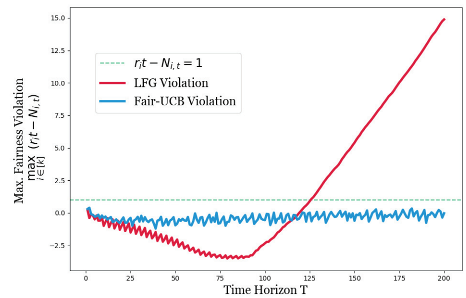

Figure 4: Fairness Guarantee

Suppose $S(T) = \Big\{ i \in [k] \,\Big|\, r_i \cdot T - \alpha < \frac{8 \ln T}{\Delta_i^2} \Big\}$. Then from the two cases discussed above, we can conclude that

$$\begin{aligned}\mathcal{R}^r_{\text{FAIR-UCB}}(T) \leq & \Big(1 + \frac{\pi^2}{3}\Big) \cdot \sum_{i \in [k]} \Delta_i \\ & + \sum_{i \in S(T), i \neq 1} \Delta_i \cdot \Big(\frac{8 \ln T}{\Delta_i^2} - \big(r_i \cdot T - \alpha\big)\Big)\end{aligned}$$

Hence, $\mathcal{R}^r_{\text{FAIR-UCB}}(T) = O(\sum_{i \neq 1} \frac{\ln T}{\Delta_i})$. □

6 Experimental Results

In this section we show the results of simulations that validate our theoretical findings. First, we represent the cost of fairness by showing the trade-off between regret and fairness with respect to the *unfairness tolerance* α. Second, we evaluate the performance of our algorithms in terms of r-Regret and fairness guarantee by comparing them with the algorithm by (Li, Liu, and Ji 2019), called Learning with Fairness Guarantee(LFG), as a baseline. Note that in Figure 3, cumulative regret is plotted on a logarithmic scale. The rationale behind the choice of instance parameters is discussed in the extended version of the paper.

Trade-off: Fairness vs. Regret

We consider the following FAIR-MAB instance: $k = 10$, $\mu_1 = 0.8$, and $\mu_i = \mu_1 - \Delta_i$, where $\Delta_i = 0.01i$, and $r = (0.05, 0.05, \ldots, 0.05) \in [0,1]^k$. We show the results for $T = 10^6$. Figure 2 shows the trade-off between regret in terms of the conventional regret and maximum fairness violation equal to $\max_{i \in [k]} r_i t - N_{i,t}$, with respect to α, and this in particular captures the *cost of fairness*. As can be seen, the regret decreases, and maximum fairness violation increases respectively as α increases till a threshold for α is reached. For values of α less than this threshold the fairness constraints cause some sub-optimal arms to be pulled more than the number of times required to determine its mean reward with sufficient confidence. On the other hand, for values of α more than this threshold, the regret reduces drastically, and we recover logarithmic regret as could be expected from the classical UCB1 algorithm. Note that the threshold for α in this case is problem-dependent.

Comparison: Fair-ucb vs. LFG: The work closest to ours is the one by (Li, Liu, and Ji 2019) and their algorithm, which is called *Learning with Fairness Guarantee* (LFG), is used as a baseline in the following simulation results. The simulation parameters that we consider for comparing r-Regret are the same as in the previous instance. Figure 3 shows the plot of time vs. r-Regret for Fair-ucb and LFG. Note that Fair-ucb and LFG perform comparably in terms of the r-Regret suffered by the algorithm. Also, the simulation results validate our theoretical result of logarithmic r-Regret bound.

We next compare fairness guarantee of Fair-ucb with that of LFG. We consider an instance with $k = 3$, $\mu = (0.7, 0.5, 0.4)$, $r = (0.2, 0.3, 0.25)$ and, $\alpha = 0$. Figure 4 shows the plot of time vs. maximum fairness violation. Observe that the fairness guarantee of Fair-ucb holds uniformly over the time horizon T. Note that, though the fairness violation for LFG appears to be increasing, it does reduce at some point and go to zero which guarantees asymptotic fairness. To summarize, the simulation result reaffirm our theoretical guarantees for both fairness and r-Regret of Fair-Learn in general, and Fair-ucb in particular.

7 Related Work

There has been a surge in research efforts aimed at ensuring fairness in decision making by machine learning algorithms such as classification algorithms (Agarwal et al. 2018; Narasimhan 2018; Zafar et al. 2017a; 2017b), regression algorithms (Berk et al. 2017; Rezaei et al. 2019), ranking and recommendation systems (Singh and Joachims 2019; Beutel et al. 2019; Singh and Joachims 2018; Celis, Straszak, and Vishnoi 2017; Zehlike et al. 2017), etc. This is true even in the context of online learning, particularly in the MAB setting. We state these relevant works below.

(Joseph et al. 2016) propose a variant of the UCB algorithm that ensures what they call meritocratic fairness i.e. an arm is never preferred over a better arm irrespective of the algorithm's confidence over the mean reward of each arm. This guarantees individual fairness (see (Dwork et al. 2012)) for each arm while achieving efficiency in terms of sublinear regret. The work by (Liu et al. 2017) aims at ensuring "treatment equality", wherein similar individuals are treated similarly. (Gillen et al. 2018) consider individual fairness guarantees with respect to an unknown fairness metric.

The papers discussed above combine the conventional goal of maximizing cumulative reward with that of simultaneously satisfying some additional constraints. MAB problem with other added constraints have been considered. For example, (Badanidiyuru, Kleinberg, and Slivkins 2013; Immorlica et al. 2018) study the MAB with knapsack constraints, where the number of times that a particular arm can be pulled is limited by some budget. The works of (Xia et al. 2015; Amin et al. 2012; Tran-Thanh et al. 2014) consider the MAB problem in which there is some cost associated with pulling each arm, and the learner has a fixed budget. The work by (Lattimore, Crammer, and Szepesvári 2014; 2015; Talebi and Proutiere 2018) investigates bandit optimization problems with resource allocation constraints.

Comparison with (Li, Liu, and Ji 2019): In addition to proving a $O(\sqrt{T \ln T})$ distribution-free r-Regret bound as in (Li, Liu, and Ji 2019), we show a $O(\ln T)$ r-Regret bound with finer dependence on the instance parameters. Our fairness guarantee holds uniformly over time and hence is much stronger than the asymptotic fairness guarantee of LFG. Moreover, as our fairness guarantee is independent of the learning algorithm used in Fair-Learn, it holds for the setting considered in (Li, Liu, and Ji 2019).

Comparison with (Celis et al. 2018): A recent work by (Celis et al. 2018) considers a personalized news feed setting, where at any time t, for a given context (user), the arm (i.e. ad to be displayed) is sampled from a distribution p^t over the set $[k]$ of arms (ads) and fairness is achieved by ensuring a pre-specified probability mass on each arm which restricts the allowable set of distributions to a subset of the simplex. The algorithm in (Celis et al. 2018) when applied to the classical stochastic multi-armed bandit setting considered by us, ensures any-time fairness only in *expectation* over the random pulls of arms by the algorithm. In contrast, our algorithm (Theorem 1) provides much stronger deterministic any-time fairness guarantee. Further, we also provide an explicit trade-off (in terms of the unfairness tolerance α), between fairness and regret. Also, the computational overhead of our algorithm is just $O(1)$, whereas the algorithms in (Celis et al. 2018) need to solve LPs in each round. We also note that our model can directly be adapted to capture the setting in (Celis et al. 2018).

8 Discussion and Future Work

The constraints considered in this paper capture fairness by guaranteeing a minimum fraction of pulls to each arm at all times. There are many situations where such fairness constraints are indispensable, and in such cases the r-Regret notion compares the expected loss of any online algorithm with the expected loss of an optimal algorithm that also satisfies such fairness constraints. An important feature of our proposed meta algorithm Fair-Learn is the uniform time fairness guarantee that it provides independent of the learning algorithm used. We also elucidate the cost of satisfying such fairness constraints by evaluating the trade-off between the conventional regret and fairness in terms of an unfairness tolerance parameter. Several notions of fairness such as disparate impact, statistical parity, equalized odds, etc. (Barocas, Hardt, and Narayanan 2018) are extensively studied in literature. Incorporating such fairness notions in online learning framework, as done by (Blum et al. 2018; Blum and Lykouris 2019; Bechavod et al. 2019), is an exciting future direction.

Acknowledgement

We thank Prof. Siddharth Barman for useful insights and helping us improve the presentation of the paper. We also thank Prof. Krishna Gummadi for pointing us to the Facebook lawsuit. Ganesh Ghalme gratefully acknowledges the travel support by Tata Trusts. Finally, we thank the anonymous reviewers for their helpful comments.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

Agrawal, S., and Goyal, N. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, 39–1.

Amin, K.; Kearns, M.; Key, P.; and Schwaighofer, A. 2012. Budget optimization for sponsored search: Censored learning in mdps. *arXiv preprint arXiv:1210.4847*.

Badanidiyuru, A.; Kleinberg, R.; and Slivkins, A. 2013. Bandits with knapsacks. In *Annual Symposium on Foundations of Computer Science*, 207–216.

Barocas, S.; Hardt, M.; and Narayanan, A. 2018. *Fairness and Machine Learning*. fairmlbook.org.

Bechavod, Y.; Ligett, K.; Roth, A.; Waggoner, B.; and Wu, Z. S. 2019. Equal opportunity in online classification with partial feedback. *arXiv preprint arXiv:1902.02242*.

Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Wei, L.; Wu, Y.; Heldt, L.; Zhao, Z.; Hong, L.; Chi, E. H.; et al. 2019. Fairness in recommendation ranking through pairwise comparisons. *arXiv preprint arXiv:1903.00780*.

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

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

Bubeck, S., and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5(1):1–122.

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

Celis, L. E.; Straszak, D.; and Vishnoi, N. K. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*.

Chatterjee, A.; Ghalme, G.; Jain, S.; Vaish, R.; and Narahari, Y. 2017. Analysis of thompson sampling for stochastic sleeping bandits. In *Uncertainty in Artificial Intelligence*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Theoretical Computer Science Conference*, 214–226.

Freeman, S. 2019. Original position. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Gillen, S.; Jung, C.; Kearns, M.; and Roth, A. 2018. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*.

Heidari, H.; Ferrari, C.; Gummadi, K.; and Krause, A. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*. 1265–1276.

Immorlica, N.; Sankararaman, K. A.; Schapire, R.; and Slivkins, A. 2018. Adversarial bandits with knapsacks. *arXiv preprint arXiv:1811.11881*.

Joseph, M.; Kearns, M.; Morgenstern, J. H.; and Roth, A. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*.

Kleinberg, R.; Niculescu-Mizil, A.; and Sharma, Y. 2010. Regret bounds for sleeping experts and bandits. *Machine learning* 80(2-3):245–272.

Lattimore, T., and Szepesvári, C. 2018. Bandit algorithms. *preprint*.

Lattimore, T.; Crammer, K.; and Szepesvári, C. 2014. Optimal resource allocation with semi-bandit feedback. *arXiv preprint arXiv:1406.3840*.

Lattimore, T.; Crammer, K.; and Szepesvári, C. 2015. Linear multi-resource allocation with semi-bandit feedback. In *Advances in Neural Information Processing Systems*.

Li, F.; Liu, J.; and Ji, B. 2019. Combinatorial sleeping bandits with fairness constraints. In *IEEE Conference on Computer Communications*.

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

Narasimhan, H. 2018. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, 1646–1654.

Rawls, J. 1971. *A theory of justice*. Harvard university press.

Rezaei, A.; Fathony, R.; Memarrast, O.; and Ziebart, B. D. 2019. Fair logistic regression: An adversarial perspective. *CoRR* abs/1903.03910.

Singh, A., and Joachims, T. 2018. Fairness of exposure in rankings. In *International Conference on Knowledge Discovery & Data Mining*, 2219–2228.

Singh, A., and Joachims, T. 2019. Policy learning for fairness in ranking. *arXiv preprint arXiv:1902.04056*.

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

Talebi, M. S., and Proutiere, A. 2018. Learning proportionally fair allocations with low regret. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 2(2):36.

Tran-Thanh, L.; Stavrogiannis, L.; Naroditskiy, V.; Robu, V.; Jennings, N. R.; and Key, P. 2014. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions.

Xia, Y.; Li, H.; Qin, T.; Yu, N.; and Liu, T.-Y. 2015. Thompson sampling for budgeted multi-armed bandits. In *International Joint Conference on Artificial Intelligence*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, 1171–1180.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017b. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, 962–970.

Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; and Baeza-Yates, R. 2017. Fa*ir: A fair top-k ranking algorithm. In *ACM Conference on Information and Knowledge Management*, 1569–1578.