Simultaneously Achieving Group Exposure Fairness and Within-Group Meritocracy in Stochastic Bandits

Subham Pokhriyal
Indian Institute of Technology Ropar
Rupnagar, India
subham.22csz0002@iitrpr.ac.in

Shweta Jain
Indian Institute of Technology Ropar
Rupnagar, India
shwetajain@iitrpr.ac.in

Ganesh Ghalme
Indian Institute of Technology Hyderabad
Hyderabad, India
ganeshghalme@ai.iith.ac.in

Swapnil Dhamal
Indian Institute of Technology Ropar
Rupnagar, India
swapnil.dhamal@iitrpr.ac.in

Sujit Gujar
International Institute of Information Technology, Hyderabad
Hyderabad, India
sujit.gujar@iiit.ac.in

ABSTRACT

Existing approaches to fairness in stochastic multi-armed bandits (MAB) primarily focus on exposure guarantee to individual arms. When arms are naturally grouped by certain attribute(s), we propose `Bi-Level Fairness`, which considers two levels of fairness. At the first level, `Bi-Level Fairness` guarantees a certain minimum exposure to each group. To address the unbalanced allocation of pulls to individual arms within a group, we consider meritocratic fairness at the second level, which ensures that each arm is pulled according to its merit within the group. Our work shows that we can adapt a UCB-based algorithm to achieve a `Bi-Level Fairness` by providing (i) anytime `Group Exposure Fairness` guarantees and (ii) ensuring individual-level `Meritocratic Fairness` within each group. We first show that one can decompose regret bounds into two components: (a) regret due to anytime group exposure fairness and (b) regret due to meritocratic fairness within each group. Our proposed algorithm BF-UCB balances these two regrets optimally to achieve the upper bound of $O(\sqrt{T})$ on regret; T being the stopping time. With the help of simulated experiments, we further show that BF-UCB achieves sub-linear regret; provides better group and individual exposure guarantees compared to existing algorithms; and does not result in a significant drop in reward with respect to UCB algorithm, which does not impose any fairness constraint.

KEYWORDS

Multi-Armed Bandit; Group Fairness; Individual Fairness

ACM Reference Format:
Subham Pokhriyal, Shweta Jain, Ganesh Ghalme, Swapnil Dhamal, and Sujit Gujar. 2024. Simultaneously Achieving Group Exposure Fairness and Within-Group Meritocracy in Stochastic Bandits. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 9 pages.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

The conventional stochastic multi-armed bandit (MAB) problem considers the problem of a learner (or bandit) having a collection of arms, where each arm is associated with an unknown probability distribution governing the rewards. The objective is to devise an arm-selection strategy that optimizes the cumulative expected reward over a series of arm selections. Stochastic MABs find its use in a wide range of applications like sponsored search auctions [28, 37], crowdsourcing [7, 21, 23–25, 39], resource allocation [8, 9, 12, 22, 42], question and answer (Q&A) forums [15] and many more.

This paper considers the problem of fair selection of arms in the stochastic multi-armed bandit problem. The fairness in stochastic MAB becomes important in applications where resources or opportunities are allocated over time among heterogeneous agents. In this context, each agent represents an arm, and pulling the arm corresponds to assigning a resource/opportunity to the selected agent. An optimal policy, in this case, would end up providing the tasks to the most rewarding agents, leaving other arms with significantly less access to resources/opportunities. Therefore, it is crucial to devise a policy that ensures sufficient exposure to each agent.

Current approaches towards fairness in stochastic MAB provide individual fairness guarantees to each arm by either offering *minimum exposure* to each arm [33] or assuring *meritocratic fairness*, i.e., ensuring that each arm is pulled in accordance with its merit (function of reward it generates) [40]. In many real-world applications, the number of arms is prohibitively large to guarantee exposure fairness at the level of individual arms. In such settings, the arms aka individual agents could be grouped based on certain attributes (e.g., gender, ethnicity, etc.) which makes aggregate group-level fairness a more natural notion [13]. However, just ensuring group-level exposure fairness may lead to selecting only the best arm within each group. In summary, there is a need for an apt fairness notion.

This paper introduces `Bi-Level Fairness` (BF) in the Multi-Armed Bandit (MAB) problem. The first level of fairness guarantees minimum exposure to each group of arms. We call this notion `Group Exposure Fairness`, which stipulates that, at the end of each round of decision-making, an arm from each group must be selected or "pulled" for a minimum pre-defined fraction of times. Group fairness is particularly relevant in settings such as crowdsourcing, job

screening, and college admissions, where each protected group is desired to be equitably represented [1]. For instance, in a crowdsourcing setting where tasks need to be assigned to workers available on the platforms, the workers are naturally grouped into different groups, possibly based on gender or ethnicity. A crowdsourcing platform may be considered discriminatory if marginalized groups receive a much lesser number of tasks as opposed to the other groups. The group fairness notion, ensuring each group receives a minimum fraction of tasks, helps to mitigate this disparity.

A group fair policy, though fair at the group level, may still allocate resources/opportunities to within group individuals/arms in a skewed manner. That is, even within a group, it may disproportionately favor one arm and hence may not give enough opportunity to the arms within the group. In this paper, we consider the notion of `Meritocratic Fairness` first proposed by Wang et al. [40] to address the problem of within-group allocation guarantee to individual arms. Meritocratic fairness ensures that each arm within each group is pulled in proportion to its merit. For example, in the credit scoring problem[4], a financial institute aims to determine the creditworthiness of potential borrowers. Here, each borrower acts as an arm and can be categorized into various groups based on a sensitive attribute (gender/marital-status/age). Each borrower's returns follow a probability distribution, which needs to be learned over time. A financial institute would like to diversify its lending amount across the different groups of borrowers, i.e., `Group Exposure Fairness`, while simultaneously ensuring that the amount is distributed in proportion to their financial capability, i.e., `Meritocratic Fairness` within the group.

One way to achieve anytime `Group Exposure Fairness` guarantees while ensuring `Meritocratic Fairness` is to combine the algorithms in [33] and [40]. Both the above works proposed upper confidence bound (UCB) based algorithms. The algorithm presented in [33] considers the minimum exposure guarantees to individual arms. One can extend this algorithm to ensure minimum exposure guarantees by applying the constraints to each group instead of each arm. The algorithm enforcing the minimum exposure constraint on each group will output a group to be pulled at each time. Once a group is selected, one can apply the algorithm presented in [40] to ensure meritocratic fairness within each group. Our proposed algorithm, Bi-Level Fair UCB, or BF-UCB in short, is primarily motivated by the above approach. The main novelty of our work lies in providing the regret guarantees for BF-UCB.

The *regret* of any online algorithm is defined as the difference in the reward obtained with the optimal algorithm and that with the online algorithm. The existing techniques from [33, 40] cannot be used to provide regret guarantees as the regret term becomes convoluted in terms of two fairness guarantees. We show that regret can be split into two terms, namely, regret due to the extra number of times a sub-optimal group is pulled and regret due to the learned fair policy within a group. Even after decomposing regret into two terms, there are two further challenges that need to be addressed to obtain the regret guarantee. First, the optimal policy in [33] is defined with respect to the best individual arm; however, here, we have optimality with respect to the best group aka collection of arms. Therefore, existing regret proof techniques in [33] that use UCB regret [3] techniques will not work here. Since we are tackling group-level fairness, our regret requires bounding the number of pulls of the sub-optimal group as opposed to a single arm. Thus, our setting requires extending the regret in [33] to combinatorial bandits setting [10]. Second, the algorithm in [40] assumes that the time horizon T is known. However, since we provide meritocratic fairness within a group, this constraint of known time horizon would mean that the algorithm would know the number of times each group is pulled before the algorithm begins. This is not possible because the number of times a group will be pulled would depend on how learning progresses and what group fairness constraints were fed to the algorithm. We overcome these challenges and prove that the proposed algorithm BF-UCB provides sub-linear regret guarantees $O(\sqrt{T})$, T being the arbitrary stopping time.

In addition to theoretical analysis, the paper includes an empirical assessment of BF-UCB against conventional bandit algorithms and their fair variants. As baseline approaches, we consider the UCB algorithm without any fairness constraint, a group exposure fair algorithm by extending the algorithm in [33] to groups, and the meritocratic fair algorithm in [40]. In particular, we show that BF-UCB achieves sub-linear regret, and that a simple extension of [33] to group fairness may lead to biases within a group, while simple meritocratic fairness [40] does not provide enough exposure to the groups. Our contributions can be summarized as follows.

Contributions.

(1) We, for the first time, introduce the notion of `Group Exposure Fairness` in stochastic MABs.
(2) We provide `Bi-Level Fairness` notion in multi-armed bandits, which ensures not only group fairness but also meritocratic fairness within a group.
(3) Inspired from UCB-based algorithms in [33] and [40], we meld them perspicaciously to build BF-UCB. BF-UCB ensures `Bi-Level Fairness`, i.e., it satisfies anytime group fairness constraint and learns meritocratic fair policy within groups.
(4) We show that regret in our setting can be decomposed into two parts, allowing BF-UCB to achieve a regret of $O(\sqrt{T})$, where T is the total number of rounds.
(5) We finally validate our results via extensive experiments.

2 RELATED WORK

In the realm of multi-armed bandits (MAB), fairness has emerged as a significant concern. Joseph et al. [27] introduced the concept of meritocratic fairness, ensuring that arms with higher rewards have a higher probability of being selected. Liu et al. [32] emphasized calibrated fairness, where arms are selected in proportion to their probability of being the best candidate, rather than based solely on average quality. Gillen et al. [16] explored individual fairness, advocating for similar arms to be treated similarly in terms of selection probabilities. Patil et al. [33] considered external constraints, designing algorithms that minimize regret while ensuring each arm is pulled a minimum fraction of rounds. Wang et al. [40] proposed Fair UCB and Fair Thompson Sampling algorithms, defining fairness regret based on the minimum merit of arms and the bounded Lipschitz constant of the merit function. All the above works ensure arm-level fairness, i.e., some exposure guarantees to each arm. So far, no works have tackled the issue of group fairness in a multi-armed bandit setting.

There have been some works in the setting known as group bandits, which categorize the arms into several groups. For example, Jedor et al. [26] considered partial ordering over the groups and analyzed dominance among categories. Wang and Scarlett [41] introduced the idea of identifying groups with the highest mean reward for the worst arm. Gabillon et al. [14] focused on the quality identification of arms within each bandit under a fixed budget constraint. Scarlett et al. [35] tackled best-arm identification in overlapping groups. While all the papers above focus only on pulling the optimal group in some sense, Schumann et al. [36] addressed the potential biases in arm selection. In this context, fairness extends beyond the individual arm to group dynamics. This paper considers the case where pulling an arm from a particular group may inherently possess biases. It assumes that, in general, the rewards of the groups are equal and tries to mitigate the bias by learning the biases in each group. The paper does not consider any constraints required to pull from each group. Further, in real-world, the assumption of rewards coming from the same distribution for two groups may not hold. In addition to these works, contextual multi-armed bandits and clustering in multi-armed bandits have been widely explored with fairness considerations by Chen et al. [11], Grazzi et al. [17]. Closer to our work is [17], where authors propose to provide exposure fairness according to the relative ordering of the arm, which is dependent on the group it belongs to. However, the paper does not consider any group fair exposure constraints.

When it comes to multi-agent, multi-armed bandit settings, agent-side fairness is emerging as an alternative perspective, where the goal is not merely to identify the best arm but to distribute the arms fairly among multiple agents [31]. Concepts such as Nash welfare solutions [5, 20] have been developed to ensure fairness amongst agents. Our setting works in a single-agent, multi-armed bandit setting, and hence, we primarily focus on arm-side fairness.

3 MODEL AND PRELIMINARIES

A traditional stochastic multi-armed bandit (MAB) problem has a set of n arms denoted as $\mathcal{A}$, where each arm i, when pulled, yields a reward following an unknown distribution with a mean reward of μ_i. Initially, these mean rewards are concealed from the designer, and the primary objective is to learn these reward values within a specified time horizon denoted by T. In standard MAB algorithms, the central aim is to identify the optimal arm that generates the highest mean reward. In our setting, the set of arms $\mathcal{A}$ is partitioned into m groups, with $m << n$. We denote the set of groups by G. The policy employed by the algorithm is denoted by $\pi = \{\pi^t\}_{t=1}^T$, where $\pi^t(i)$ denotes the probability of pulling an arm i at time t by the algorithm. Let $I_t \in g_t$ be the arm that the learner pulls at round t where $g_t \in G$ be the group pulled at round t. Let us denote the number of pulls for each arm i till time t as $N_{i,t}$ and for the group $g \in G$ as $N_{g,t} = \sum_{i \in g} N_{i,t}$.

3.1 Group Exposure Fairness

Minimum pull guarantee for each arm was first introduced by Li et al. [30] with asymptotic guarantees, which was later extended to anytime fairness guarantee by Patil et al. [33]. In this work, the individual fairness constraints are exogenously specified by a predefined vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ such that $\sum_{i \in [n]} \alpha_i < 1$, with α_i denoting the minimum fraction of times arm i needs to be pulled by the algorithm. This leads to the following definition:

Definition 1 (Patil et al. [33]). *Given a fairness constraint vector $\alpha = (\alpha_i)_{i \in [n]}$, we call a strategy π fair if $\mathbb{E}_\pi[N_{i,t}] \geq \lfloor \alpha_i t \rfloor \quad \forall i \in [n] \ \forall t \geq 1$.*

We next extend the notion of fairness in Definition 1 to the group setting in the below definition.

Definition 2 (β–Group Exposure Fairness). *Let a given fairness constraint vector be $\beta = (\beta_g)_{g \in [m]}$ such that $\beta_g \in (0, \frac{1}{m}]$ for all $g \in G$ and $\sum_{g \in [m]} \beta_g < 1$. A policy π is said to satisfy β-Group Exposure Fairness (β-GEF) if $\mathbb{E}_\pi[N_{g,t}] \geq \lfloor \beta_g t \rfloor \ \forall g \in G \ \ \forall t \geq 1$.*

As standard in the literature, we also assume, $\beta_g \leq 1/m \, \forall g$. Note that, this extension to group-level fairness is inspired by a large body of work in the literature [1, 13, 18] that focuses on equitable fairness across groups of individuals. This aggregate guarantee is motivated by social justice and legal norms that require several protected groups to have sufficient access or exposure to opportunities and resources. Satisfying only group fairness may still lead to individual-level biases within a group, for example, by always pulling a single arm whenever a group is selected. To address this, we need to introduce equity fairness within the groups. To this, we now explain Meritocratic Fairness within groups.

3.2 Meritocratic Fairness within Groups

While GEF ensures that each group of arms gets enough exposure, fair algorithms may still lead to a skewed distribution of opportunities within groups in favor of high-performing arms. We address this problem by imposing an additional constraint of Meritocratic Fairness (MF) within each group. MF ensures that each arm is pulled proportionately to its merit, defined by a merit function, and depends on the mean rewards. To define this fairness, we assume that there is a global merit function f that maps true means to the merit values. This merit function is considered to be the same for all the arms and is provided as an input to the algorithm. Before we define MF, we first state the following assumption of Lipschitz continuity of f [40].

Assumption 1. We assume that the merit function f is Lipschitz continuous, i.e., $|f(\mu_a) - f(\mu'_a) \leq L|\mu_a - \mu'_a| \ \forall \mu_a, \mu'_a$.

Assumption 2 (Minimum merit assumption)**.** There exists $0 < \gamma_1 < \gamma_2 < \infty$ such that $0 < \gamma_1 \leq f(\mu) \leq \gamma_2$ for all feasible expected rewards μ.

We can then define the Meritocratic Fairness within the group as follows.

Definition 3 (Meritocratic Fairness). *A policy $\pi_g^t(i)$ is said to satisfy Meritocratic Fairness iff $\frac{\pi_g^t(i)}{\pi_g^t(j)} = \frac{f(\mu_i)}{f(\mu_j)} \ \forall i, j \in g$. Here, $\pi_g^t(i)$ represents the probability of pulling an arm i conditioned on the event that group g is selected.*

The above definition is an extension of the definition in [40] to the individual groups. Let π_g^* represent a fair optimal policy, then it can be shown [40] that $\pi_g^*(i) = \frac{f(\mu_i)}{\sum_{j \in g} f(\mu_j)}$. Hence, if the μ's are known, the algorithm will follow π_g^* for all rounds $t \in \{1, \ldots, T\}$. However, since the μ's are not known, the goal is to learn a policy π_g^t which eventually converges to π_g^* over a period of time.

3.3 Bi-Level Fairness

We now introduce Bi-Level Fairness, which guarantees the fairness of exposure to arms as groups, and within a group, meritocratic fairness.

Definition 4 (β–Bi-Level Fairness). *Given a fairness constraint vector $\beta = (\beta_g)_{g\in[m]}$, we say that a policy π is said to satisfy β–Bi-Level Fairness iff*

(1) π satisfies β–Group Exposure Fairness, i.e., $\mathbb{E}_\pi[N_{g,t}] \geq \lfloor \beta_g t \rfloor \ \forall g \in G, \forall t \geq 1$, and

(2) π_g^t converges to π_g^* for each group g, i.e., $\lim_{N_{g,T}\to\infty} \frac{1}{N_{g,T}} \sum_{t:g_t=g} \sum_{i\in g} |\pi_g^t(i) - \pi_g^*(i)| = 0 \ \forall g \in G$.

Bi-Level Fairness notion essentially ensures β–Group Exposure Fairness at group level and ensures that the group level policy converges to fair optimal group level policy. Let us now see how an optimal policy with the knowledge of μ's, maximizing the total reward while satisfying β–Bi-Level Fairness, looks like. Since the probability of choosing an arm i within a group g is given by $\pi_g^*(i)$, the optimal group g^* will be the one with the maximum expected reward, i.e., $g^* = argmax_{g\in G} \left\{ \sum_{i\in g} \frac{f(\mu_i)}{\sum_{j\in g} f(\mu_j)} \mu_i \right\}$. We begin by observing that in any optimal fair policy, a sub-optimal group gets precisely $\lfloor \beta_g T \rfloor$ pulls, whereas the optimal group is pulled the remaining number of times; this leads to the following simple proposition.

Observation 1. *A policy π^* satisfying β-Bi-Level Fairness is said to be optimal iff it satisfies the following conditions at all time instances t:*

(1) For all $g \neq g^*$ such that $\beta_g = 0$, we have $N_{g,t} = 0$. That is, $N_{i,t} = 0$ for all $i \in g$.

(2) For all $g \neq g^*$ such that $\beta_g > 0$, we have $N_{g,t} = \lfloor \beta_g t \rfloor$ and $\pi_g^*(i) = \frac{f(\mu_i)}{\sum_{j\in g} f(\mu_j)}$.

(3) $N_{g^*,t} = t - \sum_{g\neq g^*} \lfloor \beta_g t \rfloor$ and $\pi_{g^*}^*(i) = \frac{f(\mu_i)}{\sum_{j\in g^*} f(\mu_j)}$.

The performance of any online policy is measured by its *regret* – the difference in the reward obtained by the optimal policy and that by the online policy. In order to find the regret, let us first find the reward by the optimal policy π^* which is given as:

$$R_\beta^*(T) = \sum_{g\in G} \lfloor \beta_g T \rfloor \left(\sum_{i\in g} \frac{f(\mu_i)}{\sum_{j\in g} f(\mu_j)} \mu_i \right) + \left(T - \sum_{g\in G} \lfloor \beta_g T \rfloor \right) \left(\sum_{i\in g^*} \frac{f(\mu_i)}{\sum_{j\in g^*} f(\mu_j)} \mu_i \right) \tag{1}$$

We will assume that g^* is unique for ease of explanation. However, this is not a necessary assumption for the regret guarantees to hold. We now define regret for a policy satisfying Bi-Level Fairness.

Definition 5. *Given a fairness constraint β_g, for all $g \in G$, the regret of a policy π satisfying Bi-Level Fairness is defined as:*

$$\mathfrak{R}_\pi^\beta(T) = R_\beta^*(T) - \sum_{g\in G} \sum_{i\in g} \mathbb{E}_\pi[N_{i,T}]\mu_i \tag{2}$$

In Section 5, we will show that the regret can be decomposed into two parts: (i) regret due to extra pull of a non-optimal group and (ii) regret due to suboptimal learning of policy within each group. We now propose BF-UCB in the next section, an upper confidence bound (UCB) based algorithm, satisfying Bi-Level Fairness.

4 BF-UCB: PROPOSED ALGORITHM

In this section, we propose our algorithm that ensures group exposure fairness (GEF) guarantees while maintaining meritocratic fairness (MF) within a group. The detailed algorithm is presented in Algorithm 1. As a standard practice in any MAB algorithm, our algorithm starts by pulling each arm once to get some estimates of μ_i's $\forall i \in N$. Note that a simple round-robin arm-pulling strategy breaks GEF if some of the groups have a lot more arms than other groups. In order to prevent this, we use the fact that each $\beta_g \leq 1/m$, and therefore, we select the groups in round-robin fashion until each arm in each group is pulled at least once. This is depicted in line numbers 3 to 13 of Algorithm 1. If, for a group, all the arms are completely exhausted, we start pulling the arms based on maintaining exposure fairness (line number 10 of Algorithm 1).

Once all the arms are pulled at least once, the algorithm (i) first selects a group from which arm is to be pulled (Group Selection Strategy) and then (ii) chooses the arm to pull within the group (Arm within Group Selection Strategy).

Group Selection Strategy

Motivated from [33], we propose an algorithm that provides anytime GEF guarantees. As described above, *Initialization Phase* does not violate GEF. For the remaining rounds, we use a similar approach as used by Patil et al. [33], but on the groups instead of arms. At each time t, the algorithm maintains a set $UFG\_Set(t)$, which denotes the set of groups that are on the verge of violating GEF (Line 16). If at all there exists a group in $UFG\_Set(t)$, Algorithm 1 selects a group $g \in argmax\ \beta_g(t-1) - N_{g,t-1}$ to ensure group fairness in the next round. If $UFG\_Set(t)$ is empty, the idea is to select the group with the maximum expected reward. Since the maximum expected reward is unknown beforehand, the function $Learn(\cdot)$ returns the group which helps in learning these estimates better. One could use the $Learn(\cdot)$ function based on Upper Confidence Bound (UCB) based algorithm [2, 29] or Thompson sampling-based algorithms [38]. For completeness, we have given a UCB-based algorithm in Subroutine 3. Theorem 2 in Section 5 shall show that the algorithm satisfies Group Exposure Fairness.

Arm within Group Selection Strategy

Once the group is chosen, the algorithm's arm selection strategy basically selects the arm based on Group Exposure Fairness. Our Exposure subroutine, given in Subroutine 2, looks similar to the algorithm provided by Wang et al. [40] with one key distinction. The algorithm in [40] assumes that T is known to the algorithm. Since we aim to ensure exposure fairness within each group, each group g is not chosen T number of times but is chosen $N_{g,T}$ number of rounds, which is a random variable. Therefore, we must design an algorithm without information about how often a group is selected. To tackle this challenge, we replace parameter w_0 with the value $\sqrt{2\ln(4N_{g,t}k_g/\delta)}$ instead of $\sqrt{2\ln(4TK/\delta)}$ in [40]. Here, $k_g = |g|$ denotes the number of arms in group g. In the next section, we prove that this change still provides sub-linear regret guarantees

Algorithm 1 BF-UCB

Initialize: $[n]$, Group Partition $G = \{g_1, \ldots, g_m\}$, Fairness parameter $\{\beta_g\}_{g\in G}$, Learning Function $Learn(.)$

Initialization Phase
$N_{g,0} = 0 \ \ \forall g \in G$, where $N_{g,t}$ is the number of times a group is chosen till time t
$S_{i,0} = 0 \ \ \forall i \in [n]$, where $S_{i,t}$ denotes the reward of arm i till time t
$max_{size} = \text{argmax}_{j\in[1,\ldots,m]} |g_j|$
for $k \in [1, ..., max_{size}]$ **do**
  for $g \in [1, ..., m]$ **do**
    if $\exists i \in g$ such that $N_{i,t} = 0$ **then**
      Pull arm $I_t = i$
    else
      $I_t = Exposure(g, t, f, \{N_{i,t}\}_{i\in g}, \{S_{i,t}\}_{i\in g})$
    end if
    Update $N_{I_t,t} = N_{I_t,t} + 1$, $N_{g,t} = N_{g,t} + 1$, and update $S_{I_t,t}$ based on reward
  end for
end for
$t_{init} = m \cdot max_{size}$
for $t \in [t_{init} + 1, ..., T]$ **do**
  $UFG\_Set(t) = \{g \mid \beta_g(t-1) - N_{g,t-1} > 0\}$
  if $UFG\_Set(t) \neq \phi$ **then**
    $g = \text{argmax}_{k\in UFG\_Set(t)}\{\beta_k(t-1) - N_{k,t-1}\}$
    $I_t = Exposure(g, t, f, \{N_{i,t}\}_{i\in g}, \{S_{i,t}\}_{i\in g})$
  else
    $g = Learn(G, t, f, \{N_{i,t}\}_{i\in[n]}, \{S_{i,t}\}_{i\in[n]})$
    $I_t = Exposure(g, t, f, \{N_{i,t}\}_{i\in g}, \{S_{i,t}\}_{i\in g})$
  end if
  Update $N_{I_t,t} = N_{I_t,t} + 1$, $N_{g,t} = N_{g,t} + 1$ and update $S_{I_t,t}$ based on reward
end for

without knowledge of T. Theorem 3 in Section 5 shall show that the algorithm satisfies `Meritocratic Fairness`.

Subroutine 2 Exposure$(g, t, f, \{N_{i,t}\}_{i\in g}, \{S_{i,t}\}_{i\in g})$

$\hat{\mu}_{i,t} = \frac{S_{i,t}}{N_{i,t}}, \forall\ i \in g$
$w_t^g = \sqrt{2\ln(4N_{g,t}k_g/\delta)}$
$w_{i,t} = \frac{w_t^g}{\sqrt{N_{i,t}}}, \forall\ i \in g$
$CR_t = (\mu : \forall\ i \in g,\ \mu_i \in [\hat{\mu} - w_{i,t}, \hat{\mu} + w_{i,t}])$
$\tilde{\mu}_t^g = \text{argmax}_{\mu\in CR_t} \sum_{i\in g} \frac{f(\mu_i)}{\sum_{i'\in g} f(\mu_{i'})}\mu_i$
$\pi_{i,t} = \frac{f(\tilde{\mu}_{i,t})}{\sum_{i'\in g} f(\tilde{\mu}_{i',t})}, \forall\ i \in g$
$I_t \sim \pi_t$
Return I_t

5 THEORETICAL RESULTS

This section presents three main results of the paper, namely,

(1) `Bi-Level Fairness` *Guarantees of BF-UCB*: The policy output by Algorithm 1 satisfies β–`Bi-Level Fairness` (Definition 4).

Subroutine 3 Learn$(G, t, f, \{N_{i,t}\}_{i\in[n]}, \{S_{i,t}\}_{i\in[n]})$

$\hat{\mu}_{i,t} = \frac{S_{i,t}}{N_{i,t}}, \forall\ i \in [n]$
$w_t^g = \sqrt{2\ln(4N_{g,t}k_g/\delta)}\ , \forall\ g \in G$
$w_{i,t} = \frac{w_t^g}{\sqrt{N_{i,t}}}, \forall\ i \in g, \forall\ g \in G$
$CR_t^g = (\mu : \forall\ i \in g,\ \mu_i \in [\hat{\mu}_i - w_{i,t}, \hat{\mu}_i + w_{i,t}]), \forall\ g \in G$
$\tilde{\mu}_t^g = \text{argmax}_{\mu\in CR_t^g} \sum_{i\in g} \frac{f(\mu_i)}{\sum_{i'\in g} f(\mu_{i'})}\mu_i, \forall\ g \in G$
$j = \text{argmax}_{g\in G} \sum_{i\in g} \frac{f(\tilde{\mu}_{i,t}^g)}{\sum_{i'\in g} f(\tilde{\mu}_{i',t}^g)}\tilde{\mu}_{i,t}^g$
Return j

(2) *Regret Decomposition Result:* The regret in Definition 5 can be decomposed into two parts, namely, `Group Exposure Fairness` regret and `Meritocratic Fairness` regret.
(3) *Sub-linear Regret:* The regret achieved by our algorithm is $O(\sqrt{NT})$.

5.1 `Bi-Level Fairness` Guarantees of BF-UCB

We show that BF-UCB satisfies `Bi-Level Fairness`, in two parts. First, it satisfies `GEF`, and second, it satisfies `MF`.

Theorem 2. *Algorithm 1 satisfies anytime* `GEF` *guarantees, i.e.,* $\lfloor \beta_g t \rfloor \leq N_{g,t}$ *for all* $t \geq 1$ *and for all groups* $g \in G$. *We have* $\beta_g > 0$ *and for any* β–`Bi-Level Fairness` *algorithm* $\beta_g \in (0, \frac{1}{m}]$ *for all* $g \in [m]$ *and* $\sum_{g\in m} \beta_g < 1$.

Proof. Let $t_{init} = m \cdot max_{size}$, and up to this round, each group is pulled in a round-robin fashion. Consequently, for all groups $g \in G$, the number of times group g is pulled, denoted as $N_{g,t}$, satisfies the inequality $N_{g,t} \geq \lfloor t/m \rfloor \geq \beta_g t$. The last inequality is derived from the fact that for all groups $g \in G,\ \beta_g \leq 1/m$.
For all $t \geq t_{init}$, the correctness proof follows analogous steps as outlined in [33] by establishing a mapping between each group in our setting and an arm in their setting. □

Theorem 3. *Algorithm 1 satisfies* `MF`*, i.e.,*

$$\lim_{N_{g,T}\to\infty} \frac{1}{N_{g,T}} \sum_{t:g_t=g} \sum_{i\in g} |\pi_g^t(i) - \pi_g^*(i)| = 0\ \forall g \in G.$$

Proof. Let us consider a set $T_g = \{t_1, t_2, \ldots, t_{N_{g,T}}\}$ which denotes the time steps when the group g is pulled. It is easy to see from Hoeffding's inequality [19] that,

$$\mathbb{P}(\mu_i \in CR_t) \geq 1 - \frac{\delta^2}{8N_{g,t}^2 k_g}\ \forall t > t_{k_g}, i \in [g] \tag{3}$$

We also know that the sequence $\sqrt{1/N_{i_t,t}} - \mathbb{E}_{i\in\pi_g^t}\sqrt{1/N_{i,t}}$ is a martingale difference sequence $\forall t > t_{k_g}$. Thus, we have $\left|\sqrt{1/N_{i_t,t}} - \mathbb{E}_{i\in\pi_g^t}\sqrt{1/N_{i,t}}\right| \leq 1$. We can apply the Azuma-Hoeffding's inequality to get that with probability at least $1 - \delta/2$,

$$\left|\sum_{t\in T_g} \mathbb{E}_{i\in\pi_g^t}\sqrt{1/N_{i,t}} - \sum_{t\in T_g}\sqrt{1/N_{i_t,t}}\right| \leq \sqrt{2N_{g,t}\ln(4/\delta)} \tag{4}$$

Thus, for any group g, we have:

$$\sum_{t\in T_g}\sum_{i\in g}|\pi_g^*(i)-\pi_g^t(i)| \le \sum_{t\in T_g}\frac{2\sum_{i\in g}\frac{f(\tilde{\mu}_{i,t})}{f(\tilde{\mu}_{i,t})}|f(\tilde{\mu}_{i,t})-f(\mu_i)|}{\sum_{j\in g}f(\tilde{\mu}_{i,t})}$$

(By following steps of proof of Theorem 3.2.1 from [40])

$$\le \sum_{t\in T_g}\frac{2L\sum_{i\sim\pi_g^t}|\tilde{\mu}_{i,t}-\mu_i|}{\gamma_1} \quad \text{(From Assumptions 1 and 2)}$$

$$\le \sum_{t\in T_g}\frac{2L\sum_{i\sim\pi_g^t}w_{i,t}}{\gamma_1}, \quad \text{w.p. } \left(1-\frac{\delta^2}{8N_{g,t}^2k_g}\right) \text{ (From Equation (3))}$$

$$\le \sum_{t\in T_g}\frac{2L\sum_{i\sim\pi_g^t}\sqrt{\frac{2\ln(4k_gN_{g,t}/\delta)}{N_{i,t}}}}{\gamma_1}, \quad \text{w.p. } \left(1-\frac{\delta^2}{8N_{g,t}^2k_g}\right)$$

$$\le \frac{2L\sqrt{2\ln(4k_gN_{g,T}/\delta)}}{\gamma_1}\sum_{t\in T_g}\mathbb{E}_{i\in\pi_g^t}\sqrt{\frac{1}{N_{i,t}}} \quad (\because N_{g,t}\le N_{g,T})$$

$$\le \frac{2L\sqrt{2\ln(4k_gN_{g,T}/\delta)}}{\gamma_1}\left(\sqrt{2N_{g,T}\ln(4/\delta)}+\sum_{t\in T_g,t\ge t_{k_g}}\sqrt{\frac{1}{N_{i_t,t}}}\right)$$

$$\le \frac{2L\sqrt{2\ln(4k_gN_{g,T}/\delta)}}{\gamma_1}\left(\sqrt{2N_{g,T}\ln(4/\delta)}+2\sqrt{N_{g,T}k_g}\right)$$

The last inequality follows from AM-GM inequality. The fact that $\sum_{t:g_t=g}\sum_{i\in g}|\pi_g^t(i)-\pi_g^*(i)|$ is sub-linear in $N_{g,T}$ completes the proof. □

It is to be noted that $\sum_{t\in T_g}\sum_{i\in g}|\pi_g^*(i)-\pi_g^t(i)|$ is also referred to as fairness regret FR_T in [40]. Theorem 3 says that fairness regret due to `Meritocratic Fairness` is sublinear, i.e., $O(\sqrt{T})$. The fairness regret due to `Group Exposure Fairness` will be zero since we provide anytime `Group Exposure Fairness` guarantees.

5.2 Regret Decomposition Theorem

Our next result shows that the regret of any algorithm satisfying `Bi-Level Fairness` can be decomposed into `GEF` regret and `MF` regret. Let $R_g^*=\sum_{i\in g}\pi_g^*(i)\mu_i$ denote the optimal expected reward of group g. Further, define $R_g^t=\sum_{i\in g}\pi_g^t(i)\mu_i$ to be the expected reward generated from policy π_g^t. Also, $\Delta_g=R_{g^*}^*-R_g^*$ and $\Delta_g^t=R_g^*-R_g^t$. Then, we have the following theorem.

Theorem 4 (Regret decomposition Theorem)**.** *The reward regret, $\mathfrak{R}_\pi^\beta(T)$, can be decomposed into two parts, namely, the regret due to extra pull of non-optimal group and the regret due to suboptimal learning of policy within each group, i.e.,*

$$\mathfrak{R}_\pi^\beta(T)=\sum_{g\in G}\left(\mathbb{E}_\pi[N_{g,T}]-\lfloor\beta_gT\rfloor\right)\Delta_g+\sum_{t=1}^{T}\sum_{g\in G}\mathbb{1}\left(g_t=g\right)\Delta_g^t. \quad (5)$$

Here, g_t denotes the group that is selected by the algorithm at time t.

Proof.

$$\mathfrak{R}_\pi^\beta(T)=\sum_{g\in G}\lfloor\beta_gT\rfloor R_g^*+\left(T-\sum_{g\in G}\lfloor\beta_gT\rfloor\right)R_{g^*}^*-\sum_{g\in G}\sum_{i\in g}\mathbb{E}_\pi[N_{i,T}]\mu_i$$

(From Equation (1))

$$=\sum_{g\in G}\lfloor\beta_gT\rfloor R_g^*+\left(T-\sum_{g\in G}\lfloor\beta_gT\rfloor\right)R_{g^*}^*-\sum_{t\in T}\sum_{g\in G}\mathbb{1}(g_t=g)R_g^t$$

(By the definition of R_g^t)

$$=TR_{g^*}^*-\sum_{g\in G}\lfloor\beta_gT\rfloor\left(R_{g^*}^*-R_g^*\right)-\sum_{t\in T}\sum_{g\in G}\mathbb{1}(g_t=g)R_g^t$$

(Rearranging terms)

$$=TR_{g^*}^*-\sum_{g\in G}\lfloor\beta_gT\rfloor\Delta_g-\sum_{t\in T}\sum_{g\in G}\mathbb{1}(g_t=g)R_g^t$$

From the definition of Δ_g and Δ_g^t, we have, $R_g^t=R_g^*-\Delta_g^t=R_{g^*}-\Delta_g-\Delta_g^t$. Substituting the same in the last term of regret, we get:

$$\sum_{t\in T}\sum_{g\in G}\mathbb{1}(g_t=g)R_g^t=\sum_{t\in T}\sum_{g\in G}\mathbb{1}(g_t=g)(R_{g^*}^*-\Delta_g-\Delta_g^t)$$
$$=TR_{g^*}^*-\sum_{g\in G}\mathbb{E}_\pi[N_{g,T}]\Delta_g-\sum_{t\in T}\sum_{g\in G}\mathbb{1}(g_t=g)\Delta_g^t$$

Substituting the same in the regret, we get: $\mathfrak{R}_\pi^\beta(T)=\sum_{g\in G}(\mathbb{E}_\pi[N_{g,T}]-\lfloor\beta_gT\rfloor)\Delta_g+\sum_t\sum_{g\in G}\mathbb{1}(g_t=g)\Delta_g^t$. □

The first term in Equation (5), i.e., $\sum_{g\in G}(\mathbb{E}_\pi[N_{g,T}]-\lfloor\beta_gT\rfloor)\Delta_g$, represents the cumulative regret due to extra number of times suboptimal group is pulled above the minimum guaranteed pulls $\lfloor\beta_gT\rfloor$ required to satisfy group-fairness constraints. The second term, $\sum_t\sum_{g\in G}\mathbb{1}(g_t=g)\Delta_g^t$, represents the regret due to choosing a suboptimal policy for arm pulls within the group. For a group g, the optimal policy gives the expected reward of R_g^*, whereas choosing a policy π^t, gives the reward of R_g^t. We call this difference the regret due to choosing a non-optimal policy.

5.3 Regret of BF-UCB

The regret of BF-UCB can be bounded by bounding each term separately. We now provide these bounds here (proofs provided in the extended version of the paper [34]).

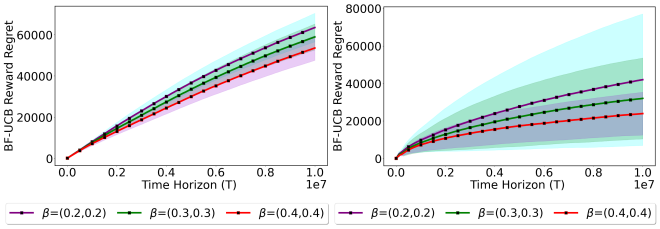
Bounding Regret due to Sub-optimal group selection. In order to bound this, we show that if we have pulled a sub-optimal group enough number of rounds, we will be able to distinguish the sub-optimal group from the optimal group with high probability and therefore, we will never select that group further. This leads to the following lemma.

Lemma 5. *Under Assumption 2,*

$$\sum_{g\in G}\mathbb{E}_\pi([N_{g,T}]-\lfloor\beta_gT\rfloor)\Delta_g\le\left(1+\frac{\pi^2}{3}\right)\sum_{g\in G}\Delta_g$$
$$+\sum_{g\in G}\left(\frac{k_gf(\gamma_2)}{f(\gamma_1)}\left(\frac{8L_1^2}{(\Delta_{min})^2}\ln\left(\frac{4N_{g,T}k_g}{\delta}\right)+\sqrt{\frac{N_{g,T}\ln(k_g/\delta)}{2}}\right)-\beta_gT\right)\Delta_g$$

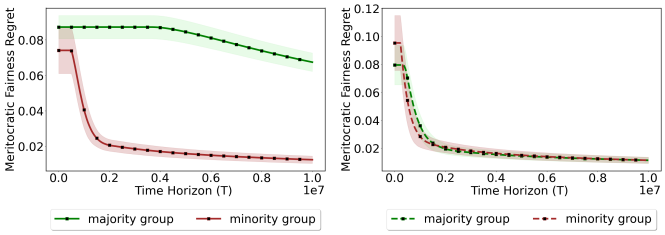
Here, $\Delta_{min}=\min_{g\neq g^*}R_{g^*}^*-R_g^*$ denotes the minimum difference between expected reward between the optimal and sub-optimal group with known rewards. Here, L_1 is a Lipschitz's constant that satisfies $|R_g(\mu)-R_g(\mu')|\le L_1|\mu-\mu'|$. Lipschitz continuity on reward function follows from Lipschitz continuity of merit function $f(\cdot)$.

We provide the proof in the extended version [34] which essentially follows similar steps to that of UCB by making use of a



(a) For high number of arms (b) For low number of arms

Figure 1: For the BF-UCB algorithm: Comparison of Reward Regret over time for different values of β



(a) For high number of arms (b) For low number of arms

Figure 2: For the BF-UCB algorithm: Comparison of Meritocratic Fairness Regret over time across the different groups

few additional results such as Lipschitz continuity on the reward function, minimum number of pulling an arm when a group is selected. Once we have these results, we can prove that if a group is pulled sufficient number of times, then each arm in that group is also pulled sufficient number of times due to meritocratic fairness. Since the reward function is Lipschitz continuous, this leads to a distinction of sub-optimal group from the optimal group.

Bounding regret due to fairness of exposure within each group. In order to bound the second term of the regret, the difference in policy is considered for the time periods when a group g is selected. The proof follows similar steps as that in [40] after replacing T with $N_{g,t}$ (number of times a group g is pulled till time t in the confidence region). Thus, the second part of the regret is given by the following lemma.

Lemma 6. *The second part of the regret is* $O\left(\sum_{g\in G}\sqrt{N_{g,T}k_g}\right)$ *with probability at least* $1-\delta$.

Thus, combining the results above leads to the following bound on the reward regret.

Theorem 7. *The reward regret (Group-Merit RR) of BF-UCB is given as:*

$$\Re_\pi^\beta(T) = \left(1+\frac{\pi^2}{3}\right)\sum_{g\in G}\Delta_g + \sum_{g\in G}\sqrt{N_{g,T}k_g}(1-\delta)+\delta T$$

$$+\sum_{g\in G}\left(\frac{k_g f(\gamma_2)}{f(\gamma_1)}\left(\frac{8L_1^2}{(\Delta_{min})^2}\ln\left(\frac{4N_{g,T}k_g}{\delta}\right)+\sqrt{\frac{N_{g,T}\ln(k_g/\delta)}{2}}\right)-\beta_g T\right)\Delta_g$$

Substituting δ to be $\Omega(1/\sqrt{T})$, we get the regret of $O(\sqrt{Tn})$.

6 EXPERIMENTS

In this section, we analyze our algorithm for regret and fairness via simulated experiments. The goal is to study the effect of the number of arms on regret and fairness guarantees, and also, how (i) GEF and (ii) MF guarantees of BF-UCB compares with that of UCB [2], [33] and [40]. We first start by explaining these baselines, followed by our experimental setup and results.

6.1 Baselines

6.1.1 UCB. This baseline is a conventional UCB algorithm [29] that aims to maximize the total reward obtained by pulling any arm without any fairness constraints.

6.1.2 Meritocratic Fair Algorithm (MF). The MF algorithm ensures meritocratic fairness across all arms independent of the groups [40] in which they are present.

6.1.3 Group Exposure Fair Algorithm (GEF). This algorithm is an adaption from [33] to group exposure where when a group is chosen, the arm with the highest reward is preferred instead of ensuring meritocratic fairness within the group.

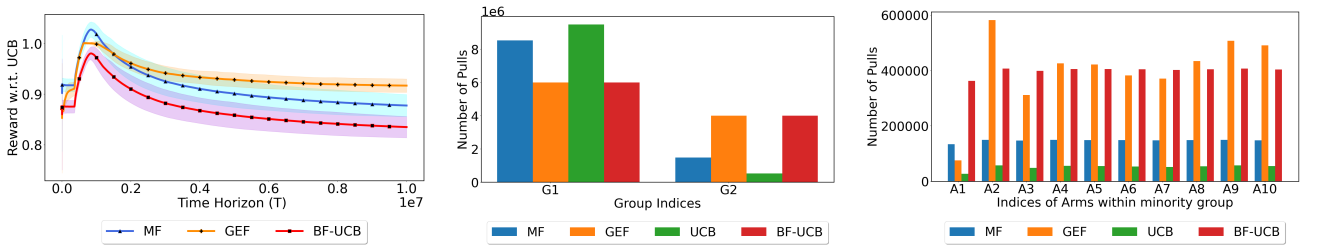
6.2 Experimental Setup

We have considered two groups, inline with majority and minority groups in the group fair literature. We ran 50 random runs of each of the experiments for a total time $T = 10^7$ to generate plots [1]. We have further considered two settings:

(1) *Low number of arms*: In this setting, we consider the number of arms in minority and majority groups to be five and ten, respectively. The mean rewards of the arms from both groups are generated uniformly from $[0.6, 0.85]$. In this setting, there is very little separability amongst the rewards of the arms, and thus, each run may lead to a different optimal group. The arm probabilities are generated afresh in each run.
(2) *High number of arms*: Here, the minority and majority groups contain ten and fifty arms, respectively. The mean rewards of arms from the majority and minority groups are generated uniformly from $[0.7, 1]$ and $[0.5, 0.8]$ respectively. This setting has clear separability amongst the optimal and sub-optimal group where majority group is optimal for all the rounds.

These threshold on number of arms is motivated by real-world examples such as the Adult dataset [6], where a typical ratio between two group values (sensitive attribute race) is typically 1:8 and in gender attributes, the typical ratio is 1:2. We consider merit function $f(\mu) = \mu$ and $\delta = .01$. The merit function is chosen thus as it can be shown that the maximum value of the reward function with the above merit function is always achieved at the highest value of μ when all μ_i's are greater than 0.5. The proof of this result is provided in the extended version of the paper [34]. Therefore, such a merit function allows us to directly use the upper confidence value of μ without explicitly computing the optimal value. It is also to be noted that the regrets will not be affected much by different merit functions. We now explain the results of BF-UCB on different performance measures in comparison with the baselines.

[1] The source code is available at: https://github.com/MultiFair-Bandits/Stochastic_Fair_Bandits/



(a) Total Reward (w.r.t. UCB) (b) Group Exposure (c) Exposure across Arms (within minority group)

Figure 3: Comparison of BF-UCB, GEF and MF on different performance measures for the setting involving high number of arms

6.3 Experimental Results

For all the comparisons, we consider $\beta = (0.4, 0.4)$ except for the comparison of regret, where we plot the total regret against all three different β values, namely, $(0.2, 0.2)$, $(0.3, 0.3)$, and $(0.4, 0.4)$.

6.3.1 Reward Regret. Figures 1a and 1b show the reward regret for the two settings, namely, high and low number of arms, respectively, for different values of β. As only BF-UCB maintains `Bi-Level Fairness`, the regret of only BF-UCB is plotted. It can be seen from both the figures that the regret is sub-linear. A higher value of β puts more constraint on the group exposure guarantee, leading to lower regret due to the sub-optimal group pull. For instance, when $\beta = (0.5, 0.5)$, both BF-UCB and the optimal algorithm will end up pulling both the groups in a round-robin fashion, thus leading to a regret of zero in the first term. The high variance for a lower number of arms setting is due to non-separability in rewards of the arms. This leads to a change in the optimal group over different runs, leading to high variance.

6.3.2 `Meritocratic Fairness` *Regret.* Figures 2a and 2b show the policy regret for the different groups, i.e., $|\pi_g^* - \pi_g^t|$ in the two settings, respectively. It can be seen from the figures that policy regret eventually converges to zero. It should be noted that though one would expect the policy regret of the majority group, which is optimal in almost all cases, in Figure 2a to converge faster, we do not see such a trend here. This is primarily due to the large number of arms in the majority group, which makes it difficult to converge faster. On the other hand, when we have low number of arms, we see this convergence much faster in Figure 2b.

6.3.3 Total Reward. Figure 3a compares the total reward of BF-UCB with different baselines for the higher number of arms setting. The rewards of different baselines are normalized with respect to the reward of UCB. As can be seen from the figure, the rewards of different algorithms initially increase with respect to UCB and then decrease gradually with time. The initial increase is due to the exploration phase of all the algorithms leading to similar rewards in the initial rounds. After a few rounds, UCB will start picking the arm with maximum reward, whereas other algorithms will have to satisfy the fairness constraint and hence, they will receive a lesser reward as compared to UCB. Since `GEF` still picks the best arm in the group whereas `MF` has to ensure exposure fairness across all the arms, the reward of `GEF` is higher than that of `MF`. It must be noted that the normalized rewards are not too far from 1 and the difference in the rewards across various baselines is not much. As expected, BF-UCB receives the least reward amongst all the algorithms as it needs to satisfy the strictest fairness notion.

6.3.4 Group Exposure. Figure 3b compares the number of times each group is pulled across different algorithms for the higher number of arms setting. As can be seen, BF-UCB and `GEF` give the most balanced exposure to the two groups. UCB gives the least exposure. On the other hand, since `MF` provides the exposure guarantees across all arms, it still ends up pulling the majority group a significantly larger number of times as compared to the minority group. This figure shows that just ensuring exposure fairness across individual arms does not guarantee enough exposure to the groups.

6.3.5 Exposure across Arms. Figure 3c plots the exposure of different arms only from the minority group for the higher number of arms setting. It shows that `MF` gives the least exposure to these arms, as there is a high number of arms in the majority group, thus leading to low exposure of arms in the minority group. The exposure to the arms is best when employing BF-UCB. `GEF` algorithm, though it seems to be giving good exposure, it should be noted that it has high variance because at each run, the optimal arm will be different and `GEF` aims to pull the optimal arm. UCB algorithm gives the least exposure to the arms present in the minority group. This figure shows that BF-UCB not only ensures group exposure but also ensures individual arm exposure within each group.

7 CONCLUSION

In summary, our novel fair Multi-Armed Bandit (MAB) framework, BF-UCB, ensures both `Meritocratic Fairness` and `Group Exposure Fairness`. Through rigorous regret decomposition analysis and from `Bi-Level Fairness` guarantee, we established its theoretical foundation. Our experimental results demonstrated competitiveness in achieving normalized rewards relative to UCB, in comparison to MF and GF. We also showcased its practical utility in achieving fair exposure to the arms within minority groups. In conclusion, our `Bi-Level Fairness` MAB algorithm, BF-UCB, is the first to give a robust solution for achieving `Bi-Level Fairness` with sublinear regret.

ACKNOWLEDGEMENT

Shweta Jain and Ganesh Ghalme gratefully acknowledge financial support from the Department of Science & Technology, India, with grant number CRG/2022/007927.

REFERENCES

[1] Mohsen Abbasi, Aditya Bhaskara, and Suresh Venkatasubramanian. 2021. Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 504–514.

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

[3] Peter Auer and Ronald Ortner. 2010. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica* 61, 1-2 (2010), 55–65.

[4] Golnoosh Babaei, Paolo Giudici, and Emanuela Raffinetti. 2023. Explainable fintech lending. *Journal of Economics and Business* 125 (2023), 106126.

[5] Siddharth Barman, Arindam Khan, Arnab Maiti, and Ayush Sawarni. 2023. Fairness and welfare quantification for regret in multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6762–6769.

[6] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository.

[7] Arpita Biswas, Shweta Jain, Debmalya Mandal, and Y Narahari. 2015. A Truthful Budget Feasible Multi-Armed Bandit Mechanism for Crowdsourcing Time Critical Tasks.. In *Proceedings of the 2015 international conference on Autonomous agents and multi-agent systems*. 1101–1109.

[8] Sanjay Chandlekar, Shweta Jain, and Sujit Gujar. 2023. A Novel Demand Response Model and Method for Peak Reduction in Smart Grids – PowerTAC. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. 3497–3504.

[9] Sanjay Chandlekar, Easwar Subramanian, and Sujit Gujar. 2023. Multi-armed Bandit Based Tariff Generation Strategy for Multi-agent Smart Grid Systems. In *International Workshop on Engineering Multi-Agent Systems*. Springer, 113–129.

[10] Wei Chen, Yajun Wang, and Yang Yuan. 2013. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*. PMLR, 151–159.

[11] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. 2020. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 181–190.

[12] Debojit Das, Shweta Jain, and Sujit Gujar. 2022. Budgeted Combinatorial Multi-Armed Bandits. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 345–353.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[14] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. 2012. Best arm identification: A unified approach to fixed budget and fixed confidence. In *In Proceedings of the 25th International Conference on Neural Information Processing Systems*. 3212–3220.

[15] Ganesh Ghalme, Swapnil Dhamal, Shweta Jain, Sujit Gujar, and Y Narahari. 2021. Ballooning multi-armed bandits. *Artificial Intelligence* 296 (2021), 103485.

[16] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. 2018. Online learning with an unknown fairness metric. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2605–2614.

[17] Riccardo Grazzi, Arya Akhavan, John IF Falk, Leonardo Cella, and Massimiliano Pontil. 2022. Group meritocratic fairness in linear contextual bandits. *Advances in Neural Information Processing Systems* 35 (2022), 24392–24404.

[18] Shivam Gupta, Ganesh Ghalme, Narayanan C Krishnan, and Shweta Jain. 2023. Efficient algorithms for fair clustering with a new notion of fairness. *Data Mining and Knowledge Discovery* 37 (2023), 1959–1997.

[19] Wassily Hoeffding. 1963. Probability Inequalities for Sums of Bounded Random Variables. *J. Amer. Statist. Assoc.* 58, 301 (1963), 13–30.

[20] Safwan Hossain, Evi Micha, and Nisarg Shah. 2021. Fair algorithms for multi-agent multi-armed bandits. *Advances in Neural Information Processing Systems* 34 (2021), 24005–24017.

[21] Shweta Jain, Ganesh Ghalme, Satyanath Bhat, Sujit Gujar, and Y Narahari. 2016. A deterministic mab mechanism for crowdsourcing with logarithmic regret and immediate payments. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. 86–94.

[22] Shweta Jain and Sujit Gujar. 2020. A multiarmed bandit based incentive mechanism for a subset selection of customers for demand response in smart grids. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2046–2053.

[23] Shweta Jain, Sujit Gujar, Satyanath Bhat, Onno Zoeter, and Yadati Narahari. 2018. A quality assuring, cost optimal multi-armed bandit mechanism for expertsourcing. *Artificial Intelligence* 254 (2018), 44–63.

[24] Shweta Jain, Sujit Gujar, Onno Zoeter, and Y Narahari. 2014. A quality assuring multi-armed bandit crowdsourcing mechanism with incentive compatible learning. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1609–1610.

[25] Shweta Jain, Balakrishnan Narayanaswamy, and Y Narahari. 2014. A multiarmed bandit incentive mechanism for crowdsourcing demand response in smart grids. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 721–727.

[26] Matthieu Jedor, Jonathan Louëdec, and Vianney Perchet. 2019. Categorized bandits. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 14422–14432.

[27] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in learning: classic and contextual bandits. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 325–333.

[28] Abhishek Kumar, Shweta Jain, and Sujit Gujar. 2020. Designing Truthful Contextual Multi-Armed Bandits based Sponsored Search Auctions. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 1732–1734.

[29] Tze Leung Lai, Herbert Robbins, et al. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6, 1 (1985), 4–22.

[30] Fengjiao Li, Jia Liu, and Bo Ji. 2019. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering* 7, 3 (2019), 1799–1813.

[31] Keqin Liu and Qing Zhao. 2010. Distributed learning in multi-armed bandit with multiple players. *IEEE transactions on signal processing* 58, 11 (2010), 5667–5681.

[32] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875* (2017).

[33] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. 2021. Achieving fairness in the stochastic multi-armed bandit problem. *The Journal of Machine Learning Research* 22, 1 (2021), 7885–7915.

[34] Subham Pokhriyal, Shweta Jain, Ganesh Ghalme, Swapnil Dhamal, and Sujit Gujar. 2024. Simultaneously Achieving Group Exposure Fairness and Within-Group Meritocracy in Stochastic Bandits. *arXiv preprint arXiv:2402.05575* (2024).

[35] Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. 2019. Overlapping multi-bandit best arm identification. In *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2544–2548.

[36] Candice Schumann, Zhi Lang, Nicholas Mattei, and John P Dickerson. 2022. Group Fairness in Bandits with Biased Feedback. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 1155–1163.

[37] Akash Das Sharma, Sujit Gujar, and Y Narahari. 2012. Truthful multi-armed bandit mechanisms for multi-slot sponsored search auctions. *Current Science* 103, 9 (2012), 1064–1077.

[38] William R Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3-4 (1933), 285–294.

[39] Long Tran-Thanh, Sebastian Stein, Alex Rogers, and Nicholas R Jennings. 2014. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence* 214 (2014), 89–111.

[40] Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. 2021. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*. PMLR, 10686–10696.

[41] Zhenlin Wang and Jonathan Scarlett. 2022. Max-min grouped bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 8603–8611.

[42] Jinhang Zuo and Carlee Joe-Wong. 2021. Combinatorial multi-armed bandits for resource allocation. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 1–4.