Fairness of Exposure in Stochastic Multiple-play Multi-armed Bandits

Youngmi Jin [1] **Dongdeok Kim** [2] **Young-Joo Suh** [3]

Abstract

We study a stochastic multiple-play multi-armed bandit (MAB) problem under semi-bandit feedback, where a decision maker selects K arms from the set of M arms under the fairness constraints requiring that each arm should be selected at least a predefined fraction of time. The objective is to maximize cumulative expected rewards while satisfying the fairness constraints. Under mild conditions, we characterize an optimal policy of the fair multiple-play MAB problem and propose a class of algorithms, called Fair-MMAB(K), based on this characterization. We show that Fair-MMAB(K) satisfies the fairness constraints at each time step, regardless of any choice of UCB index, and achieves an $O(1)$ fairness-aware regret when instantiated with UCB1 or KL-UCB. Numerical experiments validate our theoretical findings and demonstrate that Fair-MMAB(K) outperforms existing fair multiple-play MAB algorithms.

1. Introduction

As machine learning algorithms are increasingly deployed in our daily lives, fairness has attracted significant attention in the machine learning research community. Initially studied in the context of fair classification, fairness considerations have expanded to more complex settings, including fair recommendation and ranking systems.

Recommendation systems are typically trained on user interaction data, such as clicks or purchases, to predict user behavior patterns. However, as users tend to follow the recommendations provided by these systems, the systems in turn influence user behavior, thereby creating a feedback loop between users and the system. Such feedback loops can exacerbate existing biases when the recommendation system is unfair, potentially leading to discriminatory outcomes. Indeed, discrimination in online advertising was reported as early as in 2013 (Sweeney, 2013) and more recent studies have identified gender bias in music recommendations (Ferraro et al., 2021) and Facebook job advertisements (Imana et al., 2021). These findings highlight the importance of ensuring fair item exposure in recommendation systems.

This paper investigates fair item exposure in recommendation systems under a multi-armed bandit (MAB) framework. In the standard MAB setting without fairness constraints, a decision maker sequentially selects K arms from a set of M arms based on past observations, with the goal of maximizing the cumulative rewards without prior knowledge of the reward distributions. When $K = 1$, the MAB problem is referred to as a single-play MAB (SMAB) (Lai & Robbins, 1985; Auer et al., 2002). When $K \geq 2$, it is known as a multiple-play or combinatorial MAB (Anantharam et al., 1987; Gai et al., 2012; Combes et al., 2015). MAB problems are widely applied in various domains, including online advertising, recommendation systems, news feeds, clinical trials, routing, and cognitive radio networks (Bubeck & Cesa-Bianchi, 2012; Cesa-Bianchi & Lugosi, 2006; Li et al., 2019).

In this paper, we study the problem of designing multiple-play MAB algorithms that satisfy fairness constraints while maximizing cumulative rewards. The fairness constraints considered in the paper require that each arm be selected at least a predetermined fraction of time as in (Li et al., 2019; Patil et al., 2020; Liu et al., 2022). Our work extends the fair single-play MAB framework, FAIR-LEARN, (Patil et al., 2020), which achieves both constant regret and uniform fairness (i.e., fairness constraints are satisfied at every time step).

However, extending this framework from the single-play setting to the multiple-play setting is fundamentally non-trivial. The main challenge arises from the interaction between fairness constraints and the multiple-play setting. In the single-play case ($K = 1$), only one arm is selected in each time step, and the evolution of selection counts across arms is weakly coupled. In contrast, when $K \geq 2$, multiple arms are selected simultaneously, which induces strong

[1]Institute of Information and Electronics, KAIST, Daejeon, Republic of Korea [2]Samsung Electronics, Suwon, Republic of Korea [3]Department of Computer Science and Engineering, POSTECH, Pohang, Republic of Korea. Correspondence to: Youngmi Jin <youngmijin@kaist.ac.kr>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning, Seoul, Republic of Korea, 2026. Copyright 2026 by the authors.

coupling among arm selection counts. Since fairness constraints are imposed on each arm individually, this coupling significantly complicates the problem, requiring the decision maker to coordinate arm selections across all arms to satisfy the fairness constraints and maximize cumulative rewards.

In a standard MAB setting without constraints, the optimal policy has a simple structure: selecting the top-K arms with highest expected rewards. However, once fairness constraints are imposed, this top-K arm structure no longer holds. Under fairness constraints, the problem remains relatively simple when $K = 1$ because the decision maker selects a single arm and only needs to identify the best arm. In contrast, when $K \geq 2$, the decision maker must consider combinations of arms that jointly satisfy fairness requirements while maintaining high rewards. This requirement substantially increases the combinatorial complexity of the problem, given the $\binom{M}{K}$ possible selections of K arms and fundamentally complicates the design of optimal policies.

Our analysis further reveals that fairness constraints induce a refined structural decomposition of arms (e.g., into three subsets) which does not arise in the single-play setting. In particular, our proposed fair multiple-play MAB framework partitions the set of M arms into three subsets whereas FAIR-LEARN in (Patil et al., 2020) partitions it into two subsets (the best arm and the rest). Moreover, standard single-play MABs and the multiple-play MAB without fairness constraints also induce a two-set partition (top-K arms and the rest). This demonstrates that the fair multiple-play MAB setting exhibits a strictly richer structure and greater sensitivity to the relative rankings, while still not a complete ranking. This distinction reveals a fundamental structural gap between our proposed fair multiple-play MABs and FAIR-LEARN.

We briefly summarize our contributions as follows:

(i) We propose two classes of fair multiple-play MAB algorithms, Fair-MMAB(K) and Fair-MMAB(K)-MF, depending on the strength of fair exposure requirements. Both are simple and easy to implement.

(ii) We prove that Fair-MMAB(K) satisfies uniform fairness under any UCB index and achieves an $O(1)$ regret bound under commonly used UCB indices, UCB1 and KL-UCB. We also show that Fair-MMAB(K)-MF satisfies uniform fairness under stronger fairness requirements.

(iii) Through numerical experiments, we demonstrate that our algorithms outperform several state-of-the art fair multiple-play MABs including LFG, UCB-LP, and UCB-PLLP (Li et al., 2019; Liu et al., 2022) in terms of fairness and regret.

2. Problem Formulation

There is a set of arms $[M] = \{1, 2, \cdots, M\}$. Let K be a positive integer strictly smaller than M. A set of K arms, $\{a_1, a_2, , ..., a_K \mid a_i \in [M],\ a_i \neq a_j \text{ for } i \neq j\}$, is called by a super-arm. Since the set $\{a_1, a_2, , ..., a_K \mid a_i \in [M],\ a_i \neq a_j \text{ for } i \neq j\}$ can be uniquely expressed by $\boldsymbol{a} = (a_1, \cdots, a_K)$ with $a_1 < a_2 < \cdots < a_K$, a super-arm is represented by $\boldsymbol{a} = (a_1, \cdots, a_K)$ in the whole paper instead of $\{a_1, \cdots, a_K\}$ with the convention that $(a_1, \cdots, a_K)$ implies $a_1 < a_2 < \cdots < a_K$. Let $\mathcal{I}$ be the set of all possible super-arms, $\mathcal{I} = \{(a_1, a_2, ..., a_K) \mid a_i \in [M]\}$, and $\mathcal{I}_k$ the set of all super-arms containing the specific arm k, $\mathcal{I}_k = \{(a_1, a_2, ..., a_K) \in \mathcal{I} \mid a_j = k \text{ for some } j \in [K]\}$. For given $\boldsymbol{a} = (a_1, a_2, \cdots, a_K)$, we denote $i \in \boldsymbol{a}$ if $i = a_k$ for some $1 \leq k \leq K$. For example, the set of arms $\{4, 1, 5\}$ is equivalent to $\boldsymbol{a} = (1, 4, 5)$ and $4 \in \boldsymbol{a}$.

At each time step t, the decision maker selects a super-arm, denoted by $Z(t)$, which we also refer to as a recommendation list at time t. We assume semi-bandit feedback: when a super-arm $\boldsymbol{a}$ is selected at time t, each arm $i \in \boldsymbol{a}$ generates a random reward $X_i(t) \in [0, 1]$ drawn from an unknown distribution with mean θ_i. We assume that for each $i \in [M]$, the sequence $\{X_i(t)\}_{t \geq 1}$ is independently and identically distributed with mean θ_i, and the reward processes of different arms are mutually independent. We denote by $m_{\boldsymbol{a},t}$ the number of times that super-arm $\boldsymbol{a} = (a_1, a_2, \cdots, a_K)$ is selected up to time t (including t) and by $n_{i,t}$ the number of times that *single* arm i is selected up to time t. Note that $n_{i,t} = \sum_{\boldsymbol{a} \in \mathcal{I}_i} m_{\boldsymbol{a},t}$. Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_M)$. Without loss of generality, it is assumed that $1 \geq \theta_1 > \theta_2 > \cdots > \theta_M \geq 0$.

A policy, $\mathcal{A}$, is a rule that decides $Z(t)$ at each time t based on the history of selected arms and observed rewards. We use $m^{\mathcal{A}}_{\boldsymbol{a},t}$, $n^{\mathcal{A}}_{i,t}$, or $Z^{\mathcal{A}}(t)$ to explicitly mention policy $\mathcal{A}$ if necessary.

Without any information on $\boldsymbol{\theta}$, the decision maker wants to design a policy, $\mathcal{A}$, that maximizes the expected accumulated rewards over the given time horizon T, while simultaneously satisfying that each arm $i \in [M]$ should be selected at least $\lfloor c_i T \rfloor$ with $c_i > 0$ for all i over the time horizon T. Recall that $\lfloor y \rfloor$ is the biggest integer that is no more than y. The fairness conditions are specified by a vector $\boldsymbol{c} = (c_1, ..., c_M)$ and require that every arm i should be selected at least $\lfloor c_i T \rfloor$ times. Note that the value c_i specifies the minimum fraction of times that arm i should be exposed or selected at least $\lfloor c_i T \rfloor$ times over T. We assume that c_i is strictly positive for all $i \in [M]$ and denote the fairness constraints by $\boldsymbol{c} = (c_1, \cdots, c_M)$.

If $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ is known, then the optimal policy selecting a super arm can be found by solving the following

maximization problem

$$\begin{aligned} \text{Maximize} \sum_{\boldsymbol{a}\in\mathcal{I}} m_{\boldsymbol{a},T}(\theta_{a_1}+\cdots+\theta_{a_K}) \quad (1)\\ \text{subject to} \sum_{\boldsymbol{a}\in\mathcal{I}_k} m_{\boldsymbol{a},T} \geq \lfloor c_k T\rfloor \text{ for } k\in[M]\\ 0\leq m_{\boldsymbol{a},T}\leq T \text{ for } \boldsymbol{a}\in\mathcal{I},\\ \sum_{\boldsymbol{a}\in\mathcal{I}} m_{\boldsymbol{a},T} = T. \end{aligned}$$

Let $m^*_{\boldsymbol{a},T}$ be the optimal solution of (1).

Since $\boldsymbol{\theta}$ is unknown to the decision maker, she wants to find a policy $\mathcal{A}$ that minimizes the regret over time horizon T which is defined as

$$\mathcal{R}^{\mathcal{A}}_{\boldsymbol{c}}(T) = \sum_{\boldsymbol{a}\in\mathcal{I}} m^*_{\boldsymbol{a},T}\sum_{k\in\boldsymbol{a}}\theta_k - \sum_{\boldsymbol{a}\in\mathcal{I}}\mathsf{E}[m^{\mathcal{A}}_{\boldsymbol{a},T}]\sum_{k\in\boldsymbol{a}}\theta_k \quad (2)$$

while satisfying the fairness constraints $n^{\mathcal{A}}_{k,T}\geq\lfloor c_k T\rfloor$ for all $k\in[M]$. We call $\mathcal{R}^{\mathcal{A}}_{\boldsymbol{c}}(T)$ by the $\boldsymbol{c}$-fairness aware regret which is the regret for our fair multiple-play MAB setting. Note that $\mathcal{R}^{\mathcal{A}}_{\boldsymbol{c}}(T)$ is different from traditional regret of MABs without fairness constraints, $\sum_{i=1}^T\sum_{i=1}^K\theta_i - \sum_{\boldsymbol{a}\in\mathcal{I}}\mathsf{E}[m^{\mathcal{A}}_{\boldsymbol{a},T}]\sum_{k\in\boldsymbol{a}}\theta_k$ since the optimal policy always selects the K arms with the highest θ_i values when there is no fairness constraint. We use the regret instead of the $\boldsymbol{c}$-fairness aware regret for brevity in the remaining of the paper.

Note that at the stage of problem formulation, the fairness constraints do not require uniform fairness, i.e., $n^{\mathcal{A}}_{k,t}\geq\lfloor c_k t\rfloor$ for *all* $t\in[T]$ and all $k\in[M]$ but only requires $n^{\mathcal{A}}_{k,T}\geq\lfloor c_k T\rfloor$ for any $k\in[M]$. Nevertheless, our proposed algorithms satisfy the stronger uniform fairness constraints, as will be shown in subsequent sections.

3. FAIR-MMAB(K)

3.1. Optimal Policy and Fairness Aware Regret

Finding a closed-form solution to the optimization problem in (1), $m^*_{\boldsymbol{a},T}$, is very challenging since the problem is an integer program with $\binom{M}{K}$ variables. In general, solving an integer program is NP-hard. A common approach is to relax the integrality constraints and consider a real-valued optimization problem. Specifically, one first solves the relaxed problem and then converts the resulting solution into an integer-valued one. To this end , we define $x_{\boldsymbol{a},T}=\frac{1}{T}m_{\boldsymbol{a},T}$. Then (1) becomes a linear optimization (3) with constraints (4)-(6):

$$\begin{aligned} \text{Maximize} \sum_{\boldsymbol{a}\in\mathcal{I}} x_{\boldsymbol{a},T}(\theta_{a_1}+\cdots+\theta_{a_K}) \quad (3)\\ \text{subject to} \sum_{\boldsymbol{a}\in\mathcal{I}_k} x_{\boldsymbol{a},T}\geq c_k \quad \text{for } k\in[M], \quad (4)\\ 0\leq x_{\boldsymbol{a},T}\leq 1 \text{ for } \boldsymbol{a}\in\mathcal{I}, \quad (5)\\ \sum_{\boldsymbol{a}\in\mathcal{I}} x_{\boldsymbol{a},T} = 1. \quad (6) \end{aligned}$$

The linear optimization problem (3) with constraints (4) - (6) has a unique optimal solution $\boldsymbol{x}^*_{\boldsymbol{a},T}$ ((Boyd & Vandenberghe, 2004)). However, finding the closed-form expression for the optimal solution remains still challenging due to the large number of optimization variables, $\binom{M}{K}$, especilly when M is large and $K\geq 2$.

Under mild and practical conditions, the optimal solution to (3) is characterized in Theorem 3.1. All proofs are provided in Appendices.

Theorem 3.1. *Let $\mathbf{c}=(c_1,c_2,\cdots,c_M)$ with $c_i>0$ for all i. If $\sum_{j=K}^M c_j<1$, then the optimal solution of (3) is*

$$x^*_{\boldsymbol{a},T}=\begin{cases} 1-\sum_{j=K+1}^M c_j & \text{if } \boldsymbol{a}=\boldsymbol{a}^*_K,\\ c_j & \text{if } \boldsymbol{a}=\boldsymbol{a}^*_{K-1,j},\\ & K+1\leq j\leq M,\\ 0 & \text{otherwise} \end{cases}$$

where $\boldsymbol{a}^*_K=(1,2,...,K-1,K)$ and $\boldsymbol{a}^*_{K-1,j}=(1,2,...,K-1,j)$ for $K+1\leq j\leq M$.

We convert the real-valued $x^*_{\boldsymbol{a},t}$ to the integer-valued $m^*_{\boldsymbol{a},t}$.

Corollary 3.2. *If $c_i>0$ for all i and $\sum_{j=K}^M c_j<1$,*

$$m^*_{\boldsymbol{a},T}=\begin{cases} T-\sum_{j=K+1}^M\lfloor c_j T\rfloor & \text{if } \boldsymbol{a}=\boldsymbol{a}^*_K,\\ \lfloor c_j T\rfloor & \text{if } \boldsymbol{a}=\boldsymbol{a}^*_{K-1,j},\\ & K+1\leq j\leq M,\\ 0 & \text{otherwise.} \end{cases}$$

The condition $c_i>0$ for all i is mild. The primary role of the fairness constraints is to ensure sufficient exposure for arms with low expected rewards, namely those in $\{K+1,\cdots,M\}$. Thus, requiring $c_i>0$ for $i\in\{K+1,\cdots,M\}$ is essential. In contrast, arms belonged to the set $\in\{1,\cdots,K\}$ are naturally selected frequently as the decision maker learns their high expected rewards over time. Therefore, imposing $c_i>0$ does not significantly restrict the problem. Moreover, when no prior information on $\boldsymbol{\theta}$ is available, it is common to choose $c_i>0$ for all i to guarantee fairness in exposure.

The condition $\sum_{j=K}^M c_j<1$ is not strong. It is satisfied by several simple and practically relevant choices of c_i. For example, it holds if i) $\sum_{i=1}^M c_i<1$, ii) $c_i\leq\frac{1}{M}$ for all

Algorithm 1 Fair-MMAB(K): $c_i < \frac{1}{M-K+1}$ for all i

Input: $[M]$, $(c_i)_{i\in[M]}$

Initialize: $\hat{\theta}_{i,0} = n_{i,0} = u_i(0) = 0$ for all $i \in [M]$.
for $t = 1$ **to** T **do**
  $C(t) = \Big\{$ arms with $K-1$ highest UCB indices $u_i(t-1)\Big\}$.
  $F(t) = \{i \notin C(t) \mid f_i(t-1) > 0\}$ where $f_i(s) = c_i s - n_{i,s}$.
  if $F(t) \neq \emptyset$ **then**
    $\gamma(t) = \arg\max_{i\in F(t)} f_i(t-1)$.
  else
    $\gamma(t) = \arg\max_{i\notin C(t)} u_i(t-1)$.
  end if
  $Z(t) = C(t) \bigcup \{\gamma(t)\}$.
  Update $n_{i,t}$ and $\hat{\theta}_{i,n_{i,t}}$.
end for

$i \in [M]$, or iii) $c_i = c < \frac{1}{M-K+1}$ for all $i \in [M]$. Such choices are natural when there is no prior information on $\boldsymbol{\theta}$. In particular, since M is the total number of arms, the condition $c_i \leq \frac{1}{M}$ for all i is often satisfied in practice, which in turn implies $\sum_{j=K}^{M} c_j < 1$. When $c_i = c$ for all i, the condition $\sum_{j=K}^{M} c_j < 1$ is equivalent to $c < \frac{1}{M-K+1}$.

Theorem 3.3. *If $c_i > 0$ for all i and $\sum_{j=K}^{M} c_j < 1$, then it holds that*

$$\mathcal{R}_{\boldsymbol{c}}^{\mathcal{A}}(T) = \sum_{i=1}^{K}(\theta_i - \theta_K)(T - \mathsf{E}[n_{i,T}]) + \sum_{j=K+1}^{M}(\theta_K - \theta_j)(\mathsf{E}[n_{j,T}] - \lfloor c_j T \rfloor). \quad (7)$$

Note that the first term $\sum_{i=1}^{K}(\theta_i - \theta_K)(T - \mathsf{E}[n_{i,T}])$ is the regret incurred by selecting arm K instead of an arm i with $i < K$. The second term $\sum_{j=K+1}^{M}(\theta_K - \theta_j)(\mathsf{E}[n_{j,T}] - \lfloor c_j T \rfloor)$ represents the regret incurred by selecting arm $j > K$ instead of arm K.

3.2. Fair-MMAB(K) Algorithm

We propose a class of algorithms, Fair-MMAB(K), in Algorithm 1, based on the optimal solution $m^*_{\boldsymbol{a},T}$.

Fair-MMAB(K) is based on $u_i(t)$ and $f_i(t)$, the UCB index and unfairness index of arm i. The unfairness index, $f_i(t)$, of arm i at time t is defined as $f_i(t) = c_i t - n_{i,t}$. The UCB index, $u_i(t)$, of arm i at time t is generally defined using the empirical mean reward, $\hat{\theta}_{i,n_{i,t}}$, obtained from arm i up to time t, which is given by $\hat{\theta}_{i,n_{i,t}} = \frac{1}{n_{i,t}} \sum_{s=1}^{n_{i,t}} X_{i,s}$ where $X_{i,n}$ denotes the reward of arm i at n^{th} selection.

The main characteristics of Fair-MMAB(K) are as follows: (i) it selects $K-1$ arms using the K-1 highest UCB indices and at most a single arm using unfairness indices and (ii) the selection of $K-1$ arms with the $K-1$ highest UCB indices precedes the selection of the remaining arm using unfairness indices, thereby maximizing the total accumulated reward. Note that if $F(t) = \emptyset$, then $Z(t)$ consists of the top-K arms with the K highest UCB indices.

Fair-MMAB(K) has several desirable properties. First, Fair-MMAB(K) satisfies the uniform fairness property, i.e., $n_{i,t} \geq \lfloor c_i t \rfloor$ for all i and all t, regardless of the UCB index used. Second, Fair-MMAB(K) achieves an $O(1)$ regret upper bound, when either UCB1 index or KL-UCB index is used. Third, Fair-MMAB(K) is simple to implement in that it uses only $u_i(t)$ and $f_i(t)$. The simplicity allows any single-play MAB to be readily extended to a multiple-play MAB satisfying the uniform fairness property.

Theorem 3.4 states the uniform fairness property of Fair-MMAB(K).

Theorem 3.4. *For given fairness requirements $\boldsymbol{c} = (c_1, c_2, \cdots, c_M)$, if $c_i \in [0, \frac{1}{M-K+1})$ for all i, then Fair-MMAB(K) satisfies $\lfloor c_i t \rfloor \leq n_{i,t}$ for all $t \in [T]$ and all $i \in [M]$.*

Sketch of Proof: The condition $\lfloor c_k t \rfloor \leq n_{k,t}$ is equivalent to the condition $c_k t - n_{k,t} < 1$. Therefore, it is enough to show that $c_k t - n_{k,t} < 1$ holds for any k and t. For this objective, at each time step t, we define

$$W_{0,t} = \{k \in [M] \mid c_k t - n_{k,t} < 0\},$$
$$W_{j,t} = \{k \in [M] \mid q_{j-1} \leq c_k t - n_{k,t} < q_j\},$$

for $1 \leq j \leq M-K+1$ where $q_j = \frac{j}{M-K+1}$ for $0 \leq j \leq M-K+1$, and show that (i) $V_{0,t} = [M]$ and (ii) $|V_{j,t}| \leq M-K+2-j$ for all $2 \leq j < M-K+1$ and $V_{j,t} = \cup_{l=j}^{M-K+1} W_{l,t}$. □

Theorem 3.4 holds for any choice of UCB index, which implies that using Fair-MMAB(K), any UCB-based single-play MAB algorithm can be easily extended to a multiple-play MAB satisfying $n_{i,t} \geq \lfloor c_i t \rfloor$ for all $i \in [M]$ and all $t \in [T]$.

Consider the condition that $0 < c_i < \frac{1}{M-K+1}$ for all i. This condition satisfies the assumptions of both Theorem 3.3 and Theorem 3.4. We analyze the regret of Fair-MMAB(K) by combining the closed-form expression of the regret $\mathcal{R}_{\boldsymbol{c}}(T)$ in Theorem 3.3 and the uniform fairness property in Theorem 3.4.

For the regret analysis of Fair-MMAB(K), we consider two well-known UCB indices from the UCB1 and KL-UCB algorithms (Auer et al., 2002; Garivier & Cappé, 2011). We will refer to the corresponding Fair-MMAB(K) algorithms as Fair-MMAB(K)-UCB1 and Fair-MMAB(K)-KL-UCB,

respectively. The UCB1 index of arm i is given by

$$u_i(t) = \hat{\theta}_{i,n_{i,t}} + \sqrt{\frac{2\ln t}{n_{i,t}}} \tag{8}$$

for time $t \geq 1$ and $i \in [M]$ (Auer et al., 2002). The KL-UCB index is the UCB index used in KL-UCB algorithm proposed by (Garivier & Cappé, 2011), which is defined as

$$u_i(t) = \max\left\{ q > \hat{\theta}_{i,n_{i,t}} \;\middle|\; d(\hat{\theta}_{i,n_{i,t}}, q) \;\leq\; \frac{\ln(t\ln^a t)}{n_{i,t}} \right\}$$

for $t \geq 1$ and some $a \geq 0$ where $d(p,q)$ is the Kullback-Leibler divergence. [1] When $t = 0$, we define $u_i(0) = 0$ for all i. For any $p,\ q \in [0,1]$, the Kullback-Leibler divergence $d(p,q)$ is given by $d(p,q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$ with the convention, $0\ln 0 = 0,\ \ln\frac{0}{0} = 0$, and $x\ln\frac{x}{0} = \infty$ for any $x > 0$. We let $u_i(0) = 0$ for all i in either case UCB1 or KL-UCB index is used.

Theorem 3.5. *If $c_i \in (0, \frac{1}{M-K+1})$ and c_i is independent of T for all $i \in [M]$, then Fair-MMAB(K)-UCB1 achieves an $O(1)$ regret upper bound for sufficiently large T and satisfies the fairness constraints $n_{i,t} \geq \lfloor c_i t \rfloor$ for any arm $i \in [M]$ and any time step $t \in [T]$.*

Sketch of Proof: The basic ideas of the proof are: i) for any $T_0 < T$, we can write $\mathcal{R}_{\mathbf{c}}(T) = \mathcal{R}_{\mathbf{c}}(T_0) + r(T_0, T)$, ii) if we properly select T_0, then we can show that $r(T_0, T) < M_0$ where M_0 is independent of T and depends only on T_0. Such $T_0 (< T)$ can be found using $\lim_{t\to\infty} \frac{\ln t}{t} = 0$. We can interpret T_0 as the learning time to discern good arms from bad arms. The finiteness of $r(T_0, T)$ comes from the uniform fairness property and the term $-\lfloor c_j T \rfloor$ for $K+1 \leq j \leq M$ in (7). □

Theorem 3.6. *If $c_i \in (0, \frac{1}{M-K+1})$ and c_i is independent of T for all $i \in [M]$, then Fair-MMAB(K)-KL-UCB achieves an $O(1)$ regret upper bound for sufficiently large T and satisfies the fairness constraints $n_{i,t} \geq \lfloor c_i t \rfloor$ for any arm $i \in [M]$ and any time step $t \in [T]$.*

Sketch of Proof: The basic idea of the proof of Theorem 3.6 is similar to that of Theorem 3.5. The value of T_0 that makes $r(T_0, T)$ bounded can be found using $\lim_{t\to\infty} \frac{\ln(t\ln^a t)}{t} = 0$ for $a = 2$. □

Typical cases that c_i is independent of T include (i) when T is unknown and (ii) when c_i is a constant independent of T.

The work of (Patil et al., 2020) studies a fair single MAB where the fairness constraints are $n_{i,t} \geq \lfloor c_i t \rfloor - \alpha$ with $\alpha \geq 0$ for all i and t, where α controls the degree of the unfairness tolerance. In this paper, we focus on the case $\alpha = 0$, which is consistent with several existing works on fair MABs (Li et al., 2019; Liu et al., 2022). Under this formulation, the fairness requirement is directly governed by the parameters c_i. Extending our results to the case $\alpha > 0$ is straightforward.

When $K = 1$, Fair-MMAB(K) reduces to FAIR-LEARN (with $\alpha = 0$) in (Patil et al., 2020). Specifically, FAIR-LEARN selects a single arm with largest positive unfairness index if any arm satisfies $f_i(t-1) > 0$, and otherwise, selects the arm with the highest UCB index. For $K \geq 2$, there are many possible extensions of FAIR-LEARN to the multiple-play setting. Such extensions can be characterized by choosing n_U and n_F such that $n_U + n_F = K$ where n_U is the number of arms selected based on UCB indices and n_F is the maximum number of arms selected based on unfairness indices. Moreover, the order in which these selections are performed leads to different algorithms. Among these possibilities, our proposed Fair-MMAB(K) adopts $n_U = K-1$ and $n_F = 1$, selecting $K-1$ arms using the highest UCB indices first, followed by selecting the one arm using unfairness indices. This design is motivated by Theorem 3.1 and Corollary 3.2. Intuitively, prioritizing the top $K-1$ arms maximize cumulative rewards, while the remaining selection ensures fairness.

Fair-MMAB(K) partitions the set of M arms into three subsets, $\{1, 2, \cdots, K-1\}$, $\{K\}$, and $\{K+1, \cdots, M\}$, as characterized in Theorem 3.3. A standard multiple-play MAB without fairness constraints partitions the arms into two subsets: the top-K arms $\{1, 2, \cdots, K\}$ and the remaining arms $\{K+1, \cdots, M\}$. Although the multiple-play MAB setting considered in this paper does not depend on the ordering of arms within a recommendation list, the fairness constraints implicitly induce a partial ordering among the top-K arms: arms in $\{1, \cdots, K-1\}$ are selected more frequently after sufficient learning, whereas arm K is selected less frequently to meet the fairness requirements of the bad arms $\{K+1, \cdots, M\}$. In contrast, FAIR-LEARN partitions the arms into two subsets, $\{1\}$ and $\{2, \cdots, M\}$, which coincides with the structure of standard single-play MABs without fairness constraints. This difference between Fair-MMAB(K) and FAIR-LEARN indicates that fairness requirements interact more intricately with the multiple-play setting than with the single-play setting, resulting in a richer structure and increased complexity.

4. Fair-MMAB(K)-MF

This section considers how to select K arms when there exists at least one arm i such that $c_i \geq \frac{1}{M-K+1}$. This regime is more stringent than the case $0 < c_i < \frac{1}{M-K+1}$ for all i, as studied earlier. We propose Fair-MMAB(K)-

[1] Note that $\ln(t\ln^a t) = \ln(t + a\ln\ln t)$. In (Garivier & Cappé, 2011), the authors use $a = 3$ for the proof of regret bound of single-play KL-UCB algorithm and recommend to use $a = 0$ for numerical experiments. We have used $a = 2$ in our proof and $a = 0$ in our numerical experiments following their recommendation.

Algorithm 2 Fair-MMAB(K)-MF: $\frac{L-1}{M-K+L-1} \leq c_{\max} < \frac{L}{M-K+L}$

Input: $[M]$, $(c_i)_{i\in[M]}$
Initialize:
$\hat{\theta}_{i,0} = 0$ for all $i \in [M]$,
$n_{i,0} = 0$ for all $i \in [M]$,
$u_i(0) = 0$ for all $i \in [M]$.
for $t = 1$ **to** T **do**
  $C(t) = \{$ arms with $K - L$ highest UCB indices $u_i(t-1)\}$.
  $F(t) = \{i \notin C(t) \mid f_i(t-1) > 0\}$
    where $f_i(s) = c_i s - n_{i,s}$.
  $\eta_F = \min\{|F(t)|, L\}$.
  $\Phi_F(t) = \{i \in F(t) \mid$ arms with η_F highest unfairness indices $f_i(t-1)\}$.
  $\Phi_C(t) = \{i \in F(t)\backslash\Phi_F(t) \mid$ arms with $L-\eta_F$ highest UCB indices $u_i(t-1)\}$.
  $Z(t) = C(t) \bigcup \Phi_F(t) \bigcup \Phi_C(t)$.
  Update $n_{i,t}$ and $\hat{\theta}_{i,n_{i,t}}$.
end for

MF (Multple unFairness indices), a class of generalized algorithms for selecting K arms under these stricter fairness constraints and analyze its performance.

Let $c_{\max} = \max_{1\leq i\leq M} c_i$. Since $\frac{k-1}{M-K+k-1} < \frac{k}{M-K+k}$ for any positive integer k, there exists a unique integer L such that

$$\frac{L-1}{M-K+L-1} \leq c_{\max} < \frac{L}{M-K+L}.$$

With this unique L, we have $0 < c_i < \frac{L}{M-K+L}$ for all $i \in [M]$. Note that the case of $L = 1$ corresponds to the setting of Fair-MMAB(K) in the previous section. Fair-MMAB(K)-MF generalizes Fair-MMAB(K) and Fair-MMAB(K) can be viewed as a special case of Fair-MMAB(K)-MF with $L = 1$. The pseudo-code of Fair-MMAB(K)-MF is given in Algorithm 2.

Fair-MMAB(K)-MF selects K arms as follows;
1) Select $K - L$ arms with the highest UCB indices.
2) Let $\eta_F = \min\{|F(t)|, L\}$. Select η_F arms with the highest unfairness indices among those with positive unfairness indices.
3) Select $L - \eta_F$ arms with highest UCB indices among the arms which are not selected during steps 1) and 2).

If there are at most L arms with positive unfairness indices, then Fair-MMAB(K)-MF selects all of them.

Recall the definition of n_F and n_U in the discussion of Section 3.2: n_F denotes the maximum number of arms selected using unfairness indices and n_U denotes the number of arms selected using high UCB indices. In Fair-MMAB(K)-MF, $n_F = L$ and $n_U = K - L$ whereas in Fair-MMAB(K), $n_F = 1$ and $n_U = K - 1$. It implies that Fair-MMAB(K)-MF selects more arms based on unfairness indices than Fair-MMAB(K) in order to satisfy stringent fairness requirements. Moreover, since the value of n_U in Fair-MMAB(K)-MF is smaller than in Fair-MMAB(K), this suggests that satisfying fairness requirements comes at the cost of reduced cumulative rewards.

The set $\Phi_F(t) \cup \Phi_C(t)$ in Fair-MMAB(K)-MF plays a role analogous to $\{\gamma(t)\}$ in Fair-MMAB(K). Similar to Fair-MMAB(K), Fair-MMAB(K)-MF partitions the set of M arms into three subsets. However, the partition of the top-K arms differs: under Fair-MMAB(K)-MF, the top-K arms are partitioned into $\{1, 2, \cdots, K - L\}$ and $\{K - L + 1, \cdots, K\}$, whereas under Fair-MMAB(K), they are divided into $\{1, 2, \cdots, K - 1\}$, $\{K\}$, after sufficient learning.

Fair-MMAB(K)-MF, like Fair-MMAB(K), satisfies the uniform fairness property as stated in the following theorem.

Theorem 4.1. *Let $c_i \in [0, \frac{K}{M})$ for all $i \in [M]$ and L be the unique integer such that $\frac{L-1}{M-K+L-1} \leq c_{\max} < \frac{L}{M-K+L}$. Fair-MMAB(K)-MF satisfies $\lfloor c_i t \rfloor < n_{i,t}$ for all i and all t.*

Sketch of Proof: The proof is similar to that of Theorem 3.4. At each time step t, we define $W_{j,t} = \{k \in [M] \mid q_{j-1} \leq c_k t - n_{k,t} < q_j\}$ for $0 \leq j \leq M - K + L$ and show that (i) $V_{0,t} = [M]$ and (ii) $|V_{j+1,t}| \leq M - K + L - j$ for $L \leq j \leq M - K + L$ where $q_j = \frac{1}{M-K+L}$ and $V_{j,t} = \cup_{l=j}^{M-K+L} W_{l,t}$. Note that q_j and the property of (ii) are now different. □

5. Numerical Experiments

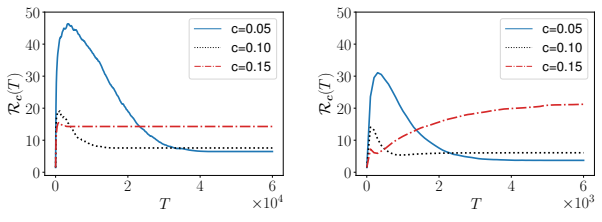
This section presents the results of numerical experiments for Fair-MMAB(K) and Fair-MMAB(K)-MF. In all the experiments, we set $M = 8$ and $K = 3$. For the reward generation of arm i, we use Bernoulli distribution with mean θ_i. Recall that we refer to the arms $\{1, \cdots, K\}$ as the good arms and the arms $\{K + 1, \cdots, M\}$ as the bad arms. To examine the effect of $\boldsymbol{\theta}$, we consider two sets of θ_i values. The entries corresponding to the good arms are highlighted in bold:

$$\boldsymbol{\theta}_1 = (\mathbf{0.9}, \mathbf{0.8}, \mathbf{0.7}, 0.6, 0.5, 0.4, 0.3, 0.2), \text{ and}$$

$$\boldsymbol{\theta}_2 = (\mathbf{0.9}, \mathbf{0.85}, \mathbf{0.8}, 0.5, 0.45, 0.4, 0.35, 0.3).$$

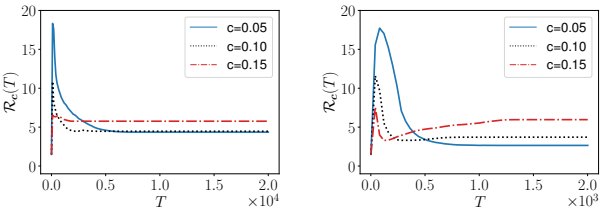
Let $\Delta_{i,j}(\boldsymbol{\theta}) = \theta_i - \theta_j$ for $i < j$. In $\boldsymbol{\theta}_1$, $\Delta_{i,i+1}(\boldsymbol{\theta}_1) = 0.1$ for all $i \in [M - 1]$. In $\boldsymbol{\theta}_2$, $\Delta_{i,i+1}(\boldsymbol{\theta}_2) = 0.05$ for $i \neq K$ and $\Delta_{K,K+1}(\boldsymbol{\theta}_2) = \Delta_{3,4}(\boldsymbol{\theta}_2) = 0.3$.

This setting implies the following:
i) Under $\boldsymbol{\theta}_1$, distinguishing the good arms, $\{1, 2, 3\}$, from the bad arms is more difficult than under $\boldsymbol{\theta}_2$, since



(a) when $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ *(b)* when $\boldsymbol{\theta} = \boldsymbol{\theta}_2$

Figure 1. $\mathcal{R}_{\boldsymbol{c}}(t)$ for Fair-MMAB(K)-UCB1



(a) when $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ *(b)* when $\boldsymbol{\theta} = \boldsymbol{\theta}_2$

Figure 2. $\mathcal{R}_{\boldsymbol{c}}(T)$ for Fair-MMAB(K)-KL-UCB

$\Delta_{3,4}(\boldsymbol{\theta}_1) = 0.1 < \Delta_{3,4}(\boldsymbol{\theta}_2) = 0.3$.
ii) Under $\boldsymbol{\theta}_1$, identifying arm K among the good arms is easier than under $\boldsymbol{\theta}_2$, since $\Delta_{2,3}(\boldsymbol{\theta}_1) = 0.1 > \Delta_{2,3}(\boldsymbol{\theta}_2) = 0.05$.

For performance comparision, we use existing fair multiple-play MAB algorithms, LFG in (Li et al., 2019), UCP-LP and UCB-PLLB in (Liu et al., 2022). The existing algorithms, LFG, UCB-LP and UCB-PLLP are designed to seek asymptotic fairness satisfying $\liminf_{t\to\infty} \frac{\mathsf{E}[n_{i,t}]}{t} \geq c_i$ for all i. Appendix G provides a brief summary of LFG, UCB-LP, UCB-PLLP and parameter settings for the plots.

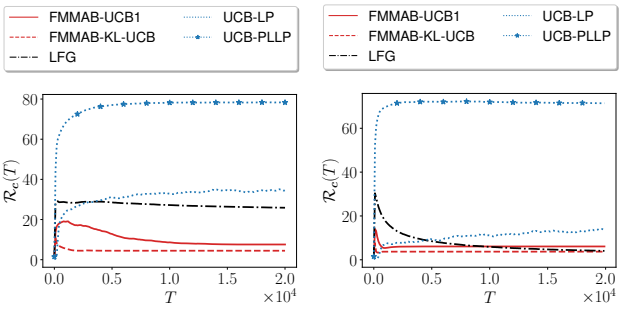
5.1. Experiments for Fair-MMAB(K)

For Fair-MMAB(K), the fairness requirements are $c_i = c$ for all $i \in [M]$ with the choice of $c \in \{0.05,\ 0.10,\ 0.15\}$. The value of $c \in \{0.05, 0.10, 0.15\}$ satisfies the conditions of Theorems 3.4- 3.6 , $c_i = c < \frac{1}{M-K+1} = \frac{1}{6}$ for all i. Additional experiments for non-constant $\boldsymbol{c}$ for Fair-MMAB(K) can be found in Appendix G.

5.1.1. Regret of Fair-MMAB(K)

We first investigate the regret of Fair-MMAB(K) and then compare it with that of existing fair multiple-play MAB algorithms.

Figures 1 and 2 exhibit the regret, $\mathcal{R}_{\boldsymbol{c}}(T)$, of Fair-MMAB(K)-UCB1 and Fair-MMAB(K)-KL-UCB for $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$. From Figures 1 and 2, we have the following observations: (i) The graphs of $\mathcal{R}_{\boldsymbol{c}}(T)$ are bounded for all choices of c and $\boldsymbol{\theta}$, which verifies Theorems 3.5 and 3.6.
(ii) A larger value of c leads to higher $\mathcal{R}_{\boldsymbol{c}}(T)$ for sufficiently



(a) when $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ *(b)* $\boldsymbol{\theta} = \boldsymbol{\theta}_2$

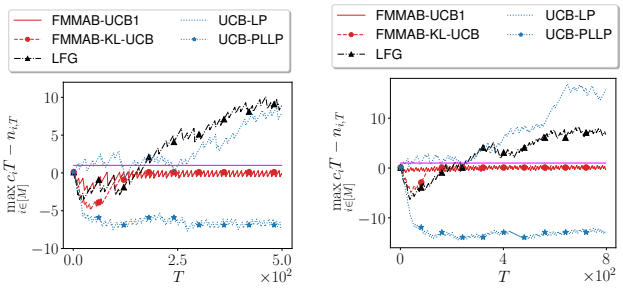
Figure 3. Comparison of $\mathcal{R}_{\boldsymbol{c}}(T)$ for Fair-MMAB(K) ($c = 0.1$)

large T when $c_i = c$ for all i. This observation suggests the trade-off between fairness and performance (regret) for large T.
(iii) When t is large, we observe that $\mathcal{R}_{\boldsymbol{c}}^{\boldsymbol{\theta}_1}(t) > \mathcal{R}_{\boldsymbol{c}}^{\boldsymbol{\theta}_2}(t)$ for small c, whereas $\mathcal{R}_{\boldsymbol{c}}^{\boldsymbol{\theta}_1}(t) < \mathcal{R}_{\boldsymbol{c}}^{\boldsymbol{\theta}_2}(t)$ for large c. For small c, identifying the good arms contributes more significantly to the regret than identifying arm K. For high c, identifying arm K contributes more significantly to the regret in the following reasons. As c increases, the bad arm must be selected more frequently to satisfy stricter fairness constraints, suggesting the decrease in the selections of the good arms, especially arm K which implies that identifying arm K from the good arms gets difficult. Moreover under $\boldsymbol{\theta}_2$, identifying arm K^{th} is more difficult than under $\boldsymbol{\theta}_1$.
(iv) Fair-MMAB(K)-KL-UCB has a lower regret than Fair-MMAB(K)-UCB1, which is consistent with the well-known fact that KL-UCB has a lower regret than UCB1 when there is no fairness constraint (Garivier & Cappé, 2011).

We compare the regret of our proposed Fair-MMAB(K) with that of the existing fair multiple-play MAB algorithms. Figure 3 shows their regret graphs for $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ and $c_i = c = 0.1$ for all i.

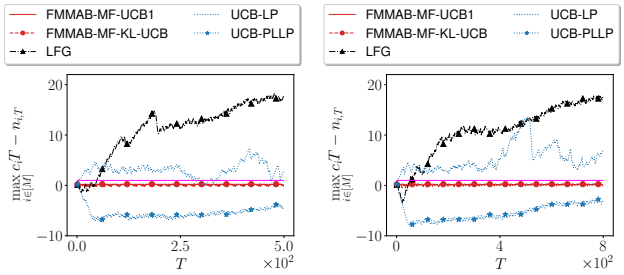
We observe that our Fair-MMAB(K)-UCB1 and Fair-MMAB(K)-KL-UCB outperform the existing fair multiple-play MAB algorithms. UCB-PLLP incurs the largest regret among the considered algorithms, implying that UCB-PLLP selects the bad arms frequently. In fact, this frequent selection of the bad arms is observed in Figure 4 which is to be discussed in the subsequent Section 5.1.2.

Based on theoretical analysis, our algorithms and UCB-LP achieve $O(1)$ regret upper bound, whereas LFG and UCB-PLLP achieve regret upper bounds of $O(\sqrt{KMT\ln T})$ and $O(K\sqrt{T\ln T})$, respectively. Although UCB-LP also achieves an $O(1)$ regret upper bound like our algorithms, it fails to satisfy the uniform fairness property, whereas our algorithms do, which is to be discussed in the subsequent Section 5.1.2.



(a) when $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ *(b)* when $\boldsymbol{\theta} = \boldsymbol{\theta}_2$

Figure 4. Uniform fairness for Fair-MMAB(K) for $c_i = 0.1$ for all i (pink line: $y = 1$): uniform fairness holds if $\max_{i\in[M]} c_iT - n_{i,T} < 1$ for all T.



(a) when $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ *(b)* when $\boldsymbol{\theta} = \boldsymbol{\theta}_2$

Figure 5. Uniform fairness of Fair-MMAB(K)-MF for $\boldsymbol{c} = (0.05, 0.10, 0.20, 0.15, 0.10, 0.15, 0.20, 0.25)$ (pink line: $y = 1$): uniform fairness holds if $\max_{i\in[M]} c_iT - n_{i,T} < 1$ for all T.

5.1.2. Fairness of Exposure of Fair-MMAB(K)

We examine the uniform fairness property, $n_{i,t} \geq \lfloor c_i t \rfloor$ for all t and all i, of Fair-MMAB(K) and the existing fair multiple-play MABs. Since the uniform fairness property is not represented as an averaged quantity but as a property for an instantiation, we do not use $\mathsf{E}[n_{i,t}]$ but use $n_{i,t}$ for a specific instantiation.

Since $n_{i,t} \geq \lfloor c_i t \rfloor$ is equivalent to $c_i t - n_{i,t} < 1$ (see Lemma C.1 in Appendix C), it is enough to investigate whether $\max_{i\in[M]} c_i t - n_{i,t} < 1$ holds for all t in order to check the uniform fairness property of Fair-MMAB(K).

Figure 4 plots the graphs of $\max_{i\in[M]} c_i t - n_{i,t}$ for $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ when $c = 0.1$ (i.e., $c_i = 1$ for all i). The pink lines in the plots represent $y = 1$. The graphs of $\max_{i\in[M]} c_i t - n_{i,t}$ of Fair-MMAB(K)-UCB1 always lies below the pink line of $y = 1$; Figure 4 verifies the uniform fairness property of Fair-MMAB(K). The graph of UCB-PLLB also lies below the pink line for all t. But the graphs of LFG and UCB-LP lie above the pink line, $y = 1$,for large t, which implies they do not have the uniform fairness property.

We take a closer look at UCB-PLLP. As observed, UCB-PLLP satisfies the uniform fairness property. In Figure 4, the value $\max_{i\in[M]} c_i t - n_{i,t}$ for UCB-PLLP remains consistently below -5 for large t, suggesting that the bad arms are frequently selected. Consequently, UCB-PLLP incurs the highest regret among all considered algorithms, as confirmed in Figure 3.

In contrast to UCB-PLLP, the graphs of Fair-MMAB(K)-UCB1 and Fair-MMAB(K)-KL-UCB in Figure 4 stay below 1 but close to 1 as t increase, indicating that the proposed algorithms judiciously control the selection of the bad arms while satisfying the fairness requirements. Such controlled selection of the bad arms yields low regret as illustrated in Figure 3.

5.2. Experiments for Fair-MMAB(K)-MF

We examine the uniform fairness property of Fair-MMAB(K)-MF and the existing fair multiple-play MABs. Recall that Fair-MMAB(K)-MF generalizes Fair-MMAB(K) to handle stricter fairness requirements. The fairness requirements are given by $\boldsymbol{c} = (\mathbf{0.05}, \mathbf{0.10}, \mathbf{0.20}, 0.15, 0.10, 0.15, 0.20, 0.25)$. Note that the bad arms have a stricter fairness requirement than that of the good arms. In this case, $c_{\max} = 0.25$ and $\frac{L-1}{M-K+L-1} \leq c_{\max} < \frac{L}{M-K+L}$ holds with $L = 2$, i.e., $\frac{1}{6} \leq c_{\max} < \frac{2}{7}$.

Figure 5 plots the values of $\max_{i\in[M]} c_i t - n_{i,t}$ over time. The graphs exhibit behavior similar to those shown in Figure 4. Our algorithms and UCB-PLLP satisfy the uniform fairness property whereas LFG and UCB-LP do not. The graph of UCB-PLLP in Figure 5 takes lower values than the corresponding graph in Figure 4 to meet the more stringent fairness requirements.

6. Related Work

Fairness in MABs have been studied under various definitions of fairness. In (Joseph et al., 2016), a variant of the (single-play) UCB algorithm is proposed to ensure that an arm with worse performance is never favored over an arm with better performance with high probability, called meritocratic fairness. The work of (Wang et al., 2021) also studies merit-based fairness of exposure and proposes single-play MAB algorithms that ensure the exposure of each arm is proportional to its expected reward.

Another widely studied notion of fairness requires each arm be selected at least a predetermined fraction of time. This requirement has been considered in (Li et al., 2019; Patil et al., 2020; Liu et al., 2022) and is also adopted in this paper. The works of (Li et al., 2019) and (Liu et al., 2022) study a combinatorial sleeping MAB (selecting at most K arms) problems under such fairness constraints. In (Li et al., 2019), Li

et al. propose the LFG algorithm by integrating UCB-based learning and virtual queue techniques, achieving a regret bound of $O\big(\sqrt{KMT\ln T}\big)$. In (Liu et al., 2022), Liu et al. propose UCB-LP and UCB-PLLP which construct randomized policies based on UCB estimates. UCB-LP achieves an $O(1)$ regret bound, while UCB-PLLP, a low complexity version of UCB-PLLP, achieves an $O(K\sqrt{T\ln T})$ regret upper bound. In contrast to these works, we consider a setting in which exactly K arms are selected every time step and the fairness constraints are satisfied at every time step (uniform fairness). This is in contrast to (Li et al., 2019) and (Liu et al., 2022) which allows fewer than K arms to be selected and only require fairness to hold asymptotically.

Our work is close related to (Patil et al., 2020), which proposes FAIR-LEARN, a fair single-play MAB framework achieving uniform fairness and $O(1)$ regret bound. We extend this framework from the single-play setting ($K = 1$) to the multiple-play setting ($K \geq 2$). This extension is fundamentally non-trivial due to the strong coupling among arms induced by selecting multiple arms simultaneously. Moreover the interaction between multiple-play and fairness constraints leads to a substantial increase in combinatorial complexity, making the problem significantly more challenging. Our analysis reveals a qualitative structural difference: our proposed algorithms partition the set of M arms into three subsets, whereas FAIR-LEARN and and standard single-play and multiple-play MAB algorithms without fairness constraints induce only a two-set partition, the set of good arms and the rest.

7. Summary

We propose Fair-MMAB(K) and its extension Fair-MMAB(K)-MF for stronger fairness requirements. We prove that Fair-MMAB(K) satisfies the fairness constraints $n_{i,t} \geq \lfloor c_i t \rfloor$ for all i at every time step t , regardless of any choice of UCB index. Furthermore, we show that it achieves a finite regret upper bound when instantiated with UCB1 and KL-UCB indices. We also prove that Fair-MMAB(K)-MF satisfies the uniform fairness property. Experimental results demonstrate that the proposed algorithms, instantiated with UCB and KL-UCB indices, outperform existing fair multiple-play MAB algorithms in terms of both fairness and regret, while ensuring uniform fairness and finite regret bounds. Our framework assumes semi-bandit feedback, where the decision maker observes the individual arms' rewards whenever selected.

An interesting research direction for future work is to design fair multiple-play MAB algorithms that achieve the uniform fairness property and a finite regret bound under full-bandit feedback where only the aggregate reward of selected arms is observed.

Acknowledgements

This research was supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2022-NR075369, RS-2022-NR070870).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Anantharam, V., Varaiya, P., and Walrand, J. Asymtotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d rewards. *IEEE Transaction on Automatic Control*, 32(11):968–976, 1987.

Auer, P., Cesa-Bianch, N., and Fischer, P. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, New York, 2004.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends in Machine Learning, 2012.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Leanring, and Games*. Cambridge University Press, New Yor, 2006.

Combes, R., Talebi, M. S., Proutiere, A., and Lelarge, M. Combinatorial bandits revisited. In *Proc. of the 28th Neural Information Processing Systems (NeurIPS 2015)*, pp. 2116–2124, Montreal, 2015. MIT Press.

Ferraro, A., Serra, X., and Bauer, C. Break the loop: gender imbalance in music remmenders. In *Proc. of the 2021 Conference on Human Information Interaction and Retrieval (CHIR 2021)*, pp. 249–254, Canberrra, 2021. ACM.

Gai, Y., Krishnamachari, B., and Jain, R. Combinatorial network optimization with unknown variables: multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5): 1466–1478, 2012.

Garivier, A. and Cappé, O. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proc. of the 24th Annual Conference on Learning Theory (COLT 2011)*, pp. 359–376, Budapest, 2011. PMLR.

Imana, B., Korolova, A., and Heidemann, J. Auditing for discrimination in algorithms delivering job ads. In *Proc. of the Web Conference (WWW 2021)*, pp. 3767–3778, Ljubljana Slovenia, 2021. ACM.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Proc. of the 29th Neural Information Processing Systems (NeurIPS 2016)*, pp. 325–323, Barcelona, 2016. NeurIPS.

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

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

Li, F., Liu, J., and Ji, B. Combinatorial sleeping bandits with fairness constraints. In *Proc. of INFOCOM 2019*, pp. 1799–1813, Paris, 2019. IEEE.

Liu, Q., Xu, W., Wang, S., and Fang, Z. Combinatorial bandits with linear constraints: beyond knapsacks and fairness. In *Proc. of the 36th Conference on Neural Information Processing System (NeurIPS 2022)*, pp. 2997–3010, Long Beach, 2022. NeurIPS.

Neely, M. J. *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Synthesis Lectures Communication Networks. Springer Cham, New York, 2010.

Patil, V., Ghalme, G., Nair, V., and Narahari, Y. Achieving fairness in the stochastic multi-armed bandit problem. In *Proc. of the 34th National Conference on American Association for Artificial Intelligence (AAAI 2020)*, pp. 5379–5386, New York, 2020. AAAI.

Sweeney, L. Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, 11(3):10–29, 2013.

Wang, L., Bai, Y., Sun, W., and Joachims, T. Fairness of exposure in stochastic bandits. In *Proc. of the 38th International Conference on Machine Learning (ICML 2021)*, pp. 139:10686–10696. PMLR, 2021.

A. Proof of Theorem 3.1

Recall that $\boldsymbol{a} = (a_1, a_2, \cdots, a_K) \in \mathcal{I}$ and $a_k \in \boldsymbol{a}$ for $1 \leq k \leq K$. The unique optimal solution $\boldsymbol{x}^*$ should meet the KKT conditions (Boyd & Vandenberghe, 2004), which are the constraints of (4)-(6) and the following conditions

$$\lambda_k(c_k - \sum_{\boldsymbol{a} \in \mathcal{I}_k} x_{\boldsymbol{a},T}) = 0 \quad \text{for } 1 \leq k \leq M \tag{9}$$

$$-\mu_{\boldsymbol{a}} x_{\boldsymbol{a},T} = 0 \qquad \text{for } \boldsymbol{a} \in \mathcal{I}\,, \tag{10}$$

$$\beta_{\boldsymbol{a}}(x_{\boldsymbol{a},T} - 1) = 0 \qquad \text{for } \boldsymbol{a} \in \mathcal{I}\,, \tag{11}$$

$$\gamma(\sum_{\boldsymbol{a} \in \mathcal{I}} x_{\boldsymbol{a},T} - 1) = 0 \qquad \text{for } \boldsymbol{a} \in \mathcal{I}\,, \tag{12}$$

$$-\sum_{j=1}^{K} \theta_{a_j} - \sum_{j=1}^{K} \lambda_{a_j} - \mu_{\boldsymbol{a}} + \beta_{\boldsymbol{a}} + \gamma = 0 \qquad \text{for } \boldsymbol{a} \in \mathcal{I}\,, \tag{13}$$

with $\gamma \in \mathbb{R}$, $\lambda_l \geq 0$ for all $l \in [M]$, $\mu_{\boldsymbol{a}} \geq 0$ for all $\boldsymbol{a} \in \mathcal{I}$, and $\beta_{\boldsymbol{a}} \geq 0$ for all $\boldsymbol{a} \in \mathcal{I}$.

By the constraint (4), $\sum_{\boldsymbol{a} \in \mathcal{I}_k} x_{\boldsymbol{a},T} \geq c_k$ for all k, it must hold that $x_{\boldsymbol{a},T} < 1$ for all $\boldsymbol{a}$. Thus, by (11), $\beta_{\boldsymbol{a}} = 0$ for all $\boldsymbol{a} \in \mathcal{I}$. Therefore, (13) reduces

$$\gamma = \mu_{\boldsymbol{a}} + \sum_{j=1}^{K} \theta_{a_j} + \sum_{j=1}^{K} \lambda_{a_j} \qquad \text{for all } \boldsymbol{a} \in \mathcal{I}. \tag{14}$$

We guess that the super-arm $(1, 2, \cdots, K)$ will be very frequently selected to maximize the total sum of rewards. The frequent selection of $(1, 2, \cdots, K)$ will yield large $x_{\boldsymbol{a}^*_K,T}$ so that both $x_{\boldsymbol{a}^*_K,T} > c_K$ and $\sum_{\boldsymbol{a} \in \mathcal{I}_i} x_{\boldsymbol{a},t} \geq c_i$ for $1 \leq i \leq K-1$ will be satisfied. From this observation, we first assume that $x_{\boldsymbol{a}^*_K,T} > c_K$ and $\sum_{\boldsymbol{a} \in \mathcal{I}_i} x_{\boldsymbol{a},t} \geq c_i$ for $1 \leq i \leq K-1$, and then, find the optimal solution $x_{\boldsymbol{a},T}$ and dual variables under the assumptions. After finding them, we check whether or not the assumptions and all of the KKT conditions are satisfied by the values of $x_{\boldsymbol{a},T}$ and dual variables we have found.

Assume that $x_{\boldsymbol{a}^*_K,T} > c_K$. By (10), $\mu_{\boldsymbol{a}^*_K} = 0$ holds. Moreover, by (9), $\lambda_k = 0$ for $k \in [K]$ since $\sum_{\boldsymbol{a} \in \mathcal{I}_k} x_{\boldsymbol{a},T} \geq x_{\boldsymbol{a}^*_K,T} > c_K$.

Consider a^*_K. If both $\mu_{\boldsymbol{a}^*_K} = 0$ and $\lambda_k = 0$ for $k \in [K]$, then $\gamma = \sum_{k=1}^{K} \theta_k$ holds by (14).

Consider $\boldsymbol{a}^*_{K-1,l} = (1, 2 \cdots, K-1, l)$ with $l \geq K+1$. We have $\mu_{\boldsymbol{a}^*_{K-1,l}} = \theta_K - \theta_l - \lambda_l$ by (14), $\gamma = \sum_{k=1}^{K} \theta_k$ and $\lambda_k = 0$ for $k \leq K$. Following the requirement $\mu_{\boldsymbol{a}} \geq 0$, we let $\mu_{\boldsymbol{a}^*_{K-1,l}} = 0$ for $l \geq K+1$. Then $\lambda_l = \theta_K - \theta_l$ holds for $K+1 \leq l \leq M$.

Consider $\boldsymbol{a} \in \{\boldsymbol{a}^*_K, \boldsymbol{a}^*_{K-1,l} \text{ with } l \geq K+1\}$. There exists a unique integer $m \leq K-2$ such that $a_m \leq K$ and $a_{m+1} \geq K+1$. Using $\gamma = \sum_{k=1}^{K} \theta_k$ and (14), we have

$$\begin{aligned}
\mu_{\boldsymbol{a}} &= \sum_{k=1}^{K} \theta_k - \Big(\sum_{l=1}^{K} \theta_{a_l} + \sum_{l=1}^{K} \lambda_{a_l}\Big) && (15)\\
&= \sum_{k=1}^{K} \theta_k - \sum_{l=1}^{m} \theta_{a_l} - \sum_{l=m+1}^{K} (\theta_{a_l} + \lambda_{a_l}) && (\because \lambda_k = 0 \text{ for } 1 \leq k \leq K)\\
&= \sum_{1 \leq k \leq K,\ k \notin \boldsymbol{a}} \theta_k - \sum_{l=m+1}^{K} (\theta_{a_l} + \lambda_{a_l})\\
&\geq \sum_{l=m+1}^{K} (\theta_K - \theta_{a_l} - \lambda_{a_l})\\
&= \sum_{l=m+1}^{K} \mu_{\boldsymbol{a}^*_{K-1,l}} && (\text{by (14)})\\
&= 0. && (16)
\end{aligned}$$

Since $\mu_{\boldsymbol{a}} > 0$, $x_{\boldsymbol{a},T} = 0$ holds by (10) for $\boldsymbol{a} \notin \{\boldsymbol{a}^*_K, \boldsymbol{a}^*_{K-1,l} \text{ with } l \geq K+1\}$.

Now we are ready to find $x_{a^*_K,T}$ and $x_{a^*_{K-1,l},T}$ with $l \geq K+1$. Before finding them, we summarize the values found so far:
(i) $x_{\boldsymbol{a},T} = 0$ for $\boldsymbol{a} \notin \{\boldsymbol{a}^*_K, \boldsymbol{a}^*_{K-1,l} \text{ with } l \geq K+1\}$
(ii) $\gamma = \sum_{k=1}^M \theta_k$,
(iii) $\beta_{\boldsymbol{a}} = 0$ for all $\boldsymbol{a}$,
(iv) $\lambda_l = \theta_K - \theta_l$ for $l \geq K+1$, otherwise, $\lambda_k = 0$.
(v) $\mu_{\boldsymbol{a}} = 0$ for $\boldsymbol{a} \in \{\boldsymbol{a}^*_K, \boldsymbol{a}^*_{K-1,l} \text{ with } l \geq K+1\}$, otherwise, $\mu_{\boldsymbol{a}}$ is given by (15).

Consider (9), i.e., $\lambda_l(c_l - \sum_{\boldsymbol{a}\in\mathcal{I}_l} x_{\boldsymbol{a},T}) = 0$ for $l \geq K+1$. Since $x_{\boldsymbol{a},T} = 0$ for $\boldsymbol{a} \notin \{\boldsymbol{a}^*_K, \boldsymbol{a}^*_{K-1,l} \text{ with } l \geq K+1\}$, it holds that $\sum_{\boldsymbol{a}\in\mathcal{I}_l} x_{\boldsymbol{a},T} = x_{\boldsymbol{a}^*_{K-1,l},T}$. Hence (9) reduces to $\lambda_l(c_l - x_{\boldsymbol{a}^*_{K-1,l},T}) = 0$. Since $\lambda_l = \theta_K - \theta_l > 0$ for $l \geq K+1$, we have $x_{\boldsymbol{a}^*_{K-1,l},T} = c_l$ for $K+1 \leq l \leq M$.

Consider (6), i.e., $\sum_{\boldsymbol{a}\in\mathcal{I}} x_{\boldsymbol{a},T} = 1$. It holds that $x_{\boldsymbol{a}^*_K,T} = 1 - \sum_{l=K+1}^M x_{\boldsymbol{a}^*_{K-1,l},T} = 1 - \sum_{l=K+1}^M c_l$ by using $x_{\boldsymbol{a}^*_{K-1,l},T} = c_l$ for $l \geq K+1$.

Finally we check whether the assumptions, $x_{\boldsymbol{a}^*_K,T} > c_K$ and $c_i < \sum_{\boldsymbol{a}\in\mathcal{I}_i} x_{\boldsymbol{a},T}$ for $1 \leq i \leq K-1$, indeed hold. For $1 \leq i \leq K-1$, note that $\sum_{\boldsymbol{a}\in\mathcal{I}_i} x_{\boldsymbol{a},T} = 1$; the assumption $c_i < 1 = \sum_{\boldsymbol{a}\in\mathcal{I}_i} x_{\boldsymbol{a},T}$ holds. For $k = K$, we have $\sum_{\boldsymbol{a}\in\mathcal{I}_K} x_{\boldsymbol{a},T} = x_{\boldsymbol{a}^*_K,T} = 1 - \sum_{l=K+1}^M c_l > c_K$ by the assumption $\sum_{l=K}^M c_l < 1$.

Now we have

$$
\begin{aligned}
x_{\boldsymbol{a},T} &= \begin{cases} 1 - \sum_{l=K+1}^M c_l & \text{if } \boldsymbol{a} = \boldsymbol{a}^*_K, \\ c_l & \text{if } \boldsymbol{a} = \boldsymbol{a}^*_{K-1,l} \text{ with } K+1 \leq l \leq M, \\ 0 & \text{otherwise.} \end{cases} \\
\beta_{\boldsymbol{a}} &= 0 \text{ for } \boldsymbol{a} \in \mathcal{I}, \\
\gamma &= \sum_{k=1}^K \theta_k, \\
\lambda_k &= \begin{cases} 0 & \text{for } 1 \leq k \leq K, \\ \theta_K - \theta_k & \text{for } K+1 \leq k \leq M, \end{cases} \\
\mu_{\boldsymbol{a}} &= \begin{cases} 0 & \text{if } \boldsymbol{a} = \boldsymbol{a}^*_K \text{ or } \boldsymbol{a} = \boldsymbol{a}^*_{K-1,l} \text{ with } K+1 \leq l \leq M, \\ \sum_{k=1}^K \theta_k - \sum_{k\in\boldsymbol{a}} (\theta_k + \lambda_k) & \text{otherwise.} \end{cases}
\end{aligned}
$$

It can be easily checked the values of $x_{\boldsymbol{a},t},\ \beta_{\boldsymbol{a}}, \lambda_l, \mu_{\boldsymbol{a}}$ meet the KKT conditions.

B. Proof of Theorem 3.3

We derive (7). By definition of $\mathcal{R}^{\mathcal{A}}_{\boldsymbol{c}}(T)$, we have

$$
\begin{aligned}
\mathcal{R}^{\mathcal{A}}_{\boldsymbol{c}}(T) &= W^* - \sum_{\boldsymbol{a}\in\mathcal{I}} \mathsf{E}[m^{\mathcal{A}}_{\boldsymbol{a},T}] \cdot (\theta_{a_1} + \cdots + \theta_{a_K}) \\
&= \sum_{i=1}^{K-1} \theta_i (T - \mathsf{E}[n_{i,T}]) + \theta_K (T - \sum_{l=K+1}^M \lfloor c_l T \rfloor - \mathsf{E}[n_{K,T}]) + \sum_{l=K+1}^M \theta_l (\lfloor c_l T \rfloor - \mathsf{E}[n_{l,T}]) \\
&= \sum_{i=1}^{K} \theta_i (T - \mathsf{E}[n_{i,T}]) - \sum_{l=K+1}^M \{(\theta_K - \theta_l)\lfloor c_l T \rfloor + \theta_l \mathsf{E}[n_{l,T}]\}. \qquad (17)
\end{aligned}
$$

We rewrite $\sum_{l=K+1}^M \{(\theta_K - \theta_l)\lfloor c_l T \rfloor + \theta_l \mathsf{E}[n_{l,T}]\}$ by adding and subtracting $\theta_K \mathsf{E}[n_{l,T}]$ in the summand of the term;

$$
\begin{aligned}
\sum_{l=K+1}^M \{(\theta_K - \theta_l)\lfloor c_l T \rfloor + \theta_l \mathsf{E}[n_{l,T}]\} &= \sum_{l=K+1}^M \{(\theta_K - \theta_l)\lfloor c_l T \rfloor + \theta_l \mathsf{E}[n_{l,T}] - \theta_K \mathsf{E}[n_{l,T}] + \theta_K \mathsf{E}[n_{l,T}]\} \\
&= -\sum_{l=K+1}^M (\theta_K - \theta_l)(\mathsf{E}[n_{l,T}] - \lfloor c_l T \rfloor) + \theta_K \sum_{l=K+1}^M \mathsf{E}[n_{l,T}]. \qquad (18)
\end{aligned}
$$

Since we select K arms at each time step, the expected number of times the entire arms are selected up to time T is KT. In other words, $\sum_{i=1}^{M} \mathsf{E}[n_{i,T}] = KT$ holds. Therefore, $\sum_{l=K+1}^{M} \mathsf{E}[E_{l,T}]$ is given by

$$\begin{aligned}\sum_{l=K+1}^{M} \mathsf{E}[n_{l,T}] &= \sum_{i=1}^{M} \mathsf{E}[n_{i,T}] - \sum_{i=1}^{K} \mathsf{E}[n_{i,T}] \\ &= KT - \sum_{i=1}^{K} \mathsf{E}[n_{i,T}] \\ &= \sum_{i=1}^{K} (T - \mathsf{E}[n_{i,T}]). \end{aligned} \tag{19}$$

Combining (17), (18), and (19), $\mathcal{R}_{\boldsymbol{c}}^{\mathcal{A}}(T)$ can be written as follows.

$$\begin{aligned}\mathcal{R}_{\boldsymbol{c}}^{\mathcal{A}}(T) &= \sum_{i=1}^{K} \theta_i (T - \mathsf{E}[n_{i,T}]) + \sum_{l=K+1}^{M} (\theta_K - \theta_l)(\mathsf{E}[n_{l,T}] - \lfloor c_l T \rfloor) - \theta_K \sum_{i=1}^{K} (T - \mathsf{E}[n_{i,T}]) \\ &= \sum_{i=1}^{K} (\theta_i - \theta_K)(T - \mathsf{E}[n_{i,T}]) + \sum_{l=K+1}^{M} (\theta_K - \theta_l)(\mathsf{E}[n_{l,T}] - \lfloor c_l T \rfloor). \end{aligned} \tag{20}$$

C. Proof of Theorem 3.4

Lemma C.1. *The condition $\lfloor c_k t \rfloor \le n_{k,t}$ is equivalent to the condition $c_k t - n_{k,t} < 1$.*

Proof. Let $c_k t = \lfloor c_k t \rfloor + h_t$ for some $0 \le h_t < 1$. If $\lfloor c_k t \rfloor - n_{k,t} \le 0$, then obviously $c_k t - n_{k,t} < 1$ holds since $c_k t - n_{k,t} = \lfloor c_k t \rfloor + h - n_{k,t} < h < 1$. Conversely, we prove that if $c_k t - n_{k,t} < 1$ then $\lfloor c_k t \rfloor - n_{k,t} \le 0$. If $c_k t - n_{k,t} < 1$, then $\lfloor c_k t \rfloor - n_{k,t} = c_k t - h_t - n_{k,t} < 1 - h_t$. Since $\lfloor c_k t \rfloor - n_{k,t}$ is an integer, the inequality $\lfloor c_k t \rfloor - n_{k,t} < 1 - h_t$ implies that $\lfloor c_k t \rfloor - n_{k,t} \le 0$. We have proved Lemma C.1. □

Let $q_j = \frac{j}{M-K+1}$ for $0 \le j \le M-K+1$. At each round t, we define S_t and $Q_{j,t}$ for $1 \le j \le M-K+1$ as below;

$$\begin{aligned} W_{0,t} &= \{k \in [M] \mid c_k t - n_{k,t} < 0\}, \\ W_{j,t} &= \{k \in [M] \mid q_{j-1} \le c_k t - n_{k,t} < q_j\}. \end{aligned}$$

By Lemma C.1, if we show that $c_k t - n_{k,t} < 1$ holds for any k and t, then Theorem 3.4 is proved. To prove that $c_k t - n_{k,t} < 1$ for all $k \in [M]$ and t, it is enough to show that $\left(\dot{\cup}_{j=0}^{M-K+1} W_{j,t}\right) = [M]$ for any t. We will prove Lemma C.2 using the mathematical induction.

Lemma C.2. *Let $V_{j,t} = \dot{\cup}_{l=j}^{M-K+1} W_{l,t}$ for $0 \le j \le M-K+1$. For $t \ge 1$, it holds that*

(a) $V_{0,t} = [M]$

(b) $|V_{j,t}| \le M - K + 2 - j$ *for* $2 \le j \le M-K+1$.

To prove Lemma C.2, we need Lemma C.3 summarizing how $W_{j,t}$ changes after Fair-MMAB(K) is exected at time t. For simple notation, we will use Z_t instead of $Z(t)$ in Fair-MMAB(K) in this proof. Recall that $A \dot{\cup} B$ means a disjoint union of A and B ($A \cap B = \emptyset$).

Lemma C.3. *Let Z_{t+1} be the set of arms selected at time step $t+1$ by Fair-MMAB(K).*

(a) If $k \in Z_{t+1} \cap W_{0,t}$, then $k \in W_{0,t+1}$.

(b) If $k \in Z_{t+1} \cap W_{j,t}$, then $k \in \begin{cases} W_{0,t+1} & \text{for } 1 \le j \le M-K, \\ W_{0,t} \cup W_{1,t+1} & \text{for } j = M-K+1. \end{cases}$

(c) If $k \in W_{0,t} \setminus Z_{t+1}$, then $k \in W_{0,t+1} \dot{\cup} W_{1,t+1}$.

(d) If $k \in W_{j,t} \setminus Z_{t+1}$ for $1 \leq j \leq M-K$, then $k \in W_{j,t+1} \dot{\cup} W_{j+1,t+1}$.

Proof. Recall that $c_i < \frac{1}{M-K+1}$ for all $i \in [M]$.

(a) Let $k \in Z_{t+1} \cap W_{0,t}$. Since $k \in Z_{t+1}$, we have $n_{k,t+1} = n_{k,t} + 1$. Then $c_k(t+1) - n_{k,t+1} = c_k t - n_{k,t} + c_k - 1$. Since $k \in W_{0,t}$, it holds that $c_k t - n_{k,t} < 0$. Therefore $c_k t - n_{k,t} + c_k - 1 < 0$, which means that $k \in W_{0,t+1}$.

(b) Since $k \in Z_{t+1}$, we have $n_{k,t+1} = n_{k,t} + 1$ and $c_k(t+1) - n_{k,t+1} = c_k t - n_{k,t} + c_k - 1$.

Consider the case that $1 \leq j \leq M-K$. The assumption $k \in W_{j,t}$ implies that we have $q_{j-1} \leq c_k t - n_{k,t} < q_j$. Hence,

$$c_k(t+1) - n_{k,t+1} = c_k t - n_{k,t} + c_k - 1 < q_j + c_k - 1 < \frac{j - (M-K+1)}{M-K+1} + c_k \leq 0,$$

since $j \leq M-K$ and $c_k < \frac{1}{M-K+1}$. Therefore, $k \in W_{0,t+1}$.

Consider the case that $j = M-K+1$. The assumption $k \in W_{M-K+1,t}$ implies that $q_{M-K} \leq c_k t - n_{k,t} < 1$. Therefore $c_k t - n_{k,t} + c_k - 1 < c_k < \frac{1}{M-K+1}$, which means that $k \in W_{0,t+1} \cup W_{1,t+1}$.

(c) Since $k \notin Z_{t+1}$, we have $n_{k,t+1} = n_{k,t}$. Therefore, $c_k(t+1) - n_{k,t+1} = c_k t - n_{k,t} + c_k$. Since $k \in W_{0,t}$, it holds that $c_k t - n_{k,t} < 0$, which means $c_k t - n_{k,t} + c_k < c_k < \frac{1}{M-K+1}$. Hence $k \in W_{0,t+1} \cup W_{1,t+1}$.

(d) Since $k \notin Z_{t+1}$, we have $n_{k,t+1} = n_{k,t}$ and $c_k(t+1) - n_{k,t+1} = c_k - n_{k,t} + c_k$. Since $k \in W_{j,t}$ for $1 \leq j \leq M-K$, it holds that $q_{j-1} \leq c_k t - n_{k,t} < q_j$. Hence, we have

$$q_{j-1} < q_{j-1} + c_k \leq c_k(t+1) - n_{k,t+1} = c_k - n_{k,t} + c_k < q_j + c_k < q_{j+1},$$

which implies that $k \in W_{j,t} \cup W_{j+1,t}$.

We have proved Lemma C.3. □

Now, we use the mathematical induction to prove Lemma C.2.

At time $t = 1$: We should prove that Lemma C.2 holds.
Recall that $c_i < \frac{1}{M-K+1}$ for all $i \in [M]$.
(i) For $k \in Z_1$ (arm k is selected), we have $c_k t - n_{k,t} = c_k \cdot 1 - 1 < 0$, which implies that $k \in W_{0,1}$. Hence $Z_1 \subset W_{0,1}$.
(ii) For $k \notin Z_1$ (arm k is not selected), we have $c_k \cdot 1 - n_{k,1} = c_k < \frac{1}{M-K+1}$, which implies that $k \in W_{1,1}$. Therefore $[M] - Z_1 \subset W_{1,1}$.

From (i) and (ii), it holds that $[M] = W_{0,1} \dot{\cup} W_{1,1}$. Moreover, (i) implies $|W_{0,1}| \geq |Z_1| = K$ and (ii) implies $|W_{1,1}| \geq M-K$. Since $|W_{0,1} \cup W_{1,1}| = M$ and $W_{0,1} \cap W_{1,1} = \emptyset$, it must hold that $|W_{0,1}| = |Z_1| = K$ and $|W_{1,1}| = M-K$, implying $|W_{j,1}| = 0$ for $j \geq 2$. Therefore $|V_{1,1}| = |\cup_{l=1}^{M-K+1} W_{l,1}| = M-K$ and $|V_{j,1}| = |\cup_{l=j}^{M-K+1} W_{l,1}| = 0$ for $j \geq 2$. We have proved that Lemma C.2 holds when $t = 1$.

At time t: We assume that Lemma C.2 holds at time t.

At time $t+1$: We have to show that Lemma C.2 still holds at time $t+1$.

We first show that $V_{0,t+1} = [M]$. From the induction hypothesis at time t, we have $[M] = \cup_{l=0}^{M-K+l} W_{l,t}$. We denote $[M]$ as $[M] = W_{0,t} \bigcup \left(\cup_{l=1}^{M-K} W_{l,t} \right) \bigcup V_{M-K+1,t}$ (recall that $V_{k,t} = \cup_{l=k}^{M-K+L} W_{l,t}$). The followings are hold

i) $W_{0,t} \subset \cup_{l=0}^{1} W_{l,t+1}$ by (a) and (c) of Lemma C.2,

ii) $\cup_{l=1}^{M-K} W_{l,t} \cap Z_{t+1} \subset W_{0,t+1}$ by (b) of Lemma C.2,

iii) $\cup_{l=1}^{M-K} W_{l,t} \setminus Z_{t+1} \subset \cup_{l=1}^{M-K+1} W_{l,t+1}$ by (d) of Lemma C.2, and

iv) $W_{M-K+1,t} \subset Z_{t+1}$ and $W_{M-K+1,t} \subset \cup_{j=0}^{1} W_{j,t+1}$.

We prove item iv) holds. By induction hypothesis, $|W_{M-K+1,t}| = |V_{M-K+1,t}| \leq 1$ holds. Therefore, $W_{M-K+1,t}$ has at most a single element. Recall that $Z_{t+1} = C(t+1) \cup \{\gamma(t+1)\}$. If $W_{M-K+1,t} = \emptyset$, then item iv) holds. Suppose that $W_{M-K+1,t} \neq \emptyset$. Then $W_{M-K+1,t}$ has a single element, namely l^*. This l^* has the highest positive unfairness index becase $|W_{M-K+1,t}| = 1$ and $[M] = V_{0,t} = \cup_{l=0} W_{l,t}$. If $l^* \in C(t+1)$, then $l^* \in Z_{t+1}$. If $l^* \notin C(t+1)$, then $\gamma(t+1) = l^*$ by the definition of $\gamma(t+1)$. In either case, $l^* \in Z_{t+1}$. Thus $W_{M-K+1,t} \subset Z_{t+1}$. By (b) of Lemma C.2, $W_{M-K+1,t} \subset \cup_{j=0}^{1} W_{j,t+1}$. Item iv) holds.

By induction hypothesis, we have $[M] = \cup_{l=0}^{M-K+1} W_{l,t}$. Items i)-iv) imply that $\cup_{l=0}^{M-K+1} W_{l,t} \subset \cup_{l=0}^{M-K+1} W_{l,t+1}$. Hence $[M] = \cup_{l=0}^{M-K+1} W_{l,t+1} = V_{0,t+1}$.

We show (b), i.e, that $|V_{j+1,t}| \leq M - K + 1 - j$ for $0 \leq j \leq M - K$.

From Lemma C.2, the set $W_{j,t}$ is partitioned to $W_{j,t} = W_{j,t}^{play} \cup W_{j,t}^{stay} \cup W_{j,t}^{move}$ for $0 \leq j \leq M - K + 1$, where $W_{j,t}^{play} = W_{j,t} \cap Z_{t+1}$, $W_{j,t}^{stay} = W_{j,t} \cap W_{j,t+1}$, and $W_{j,t}^{move} = W_{j,t} - (W_{j,t}^{play} \cup W_{j,t}^{stay})$.

Using Lemma C.2 and these notation, for $j \geq 2$, we can represent $W_{j,t+1}$ as $W_{j,t+1} = W_{j-1,t}^{move} \cup W_{j,t}^{stay}$. Using $W_{j,t+1} = W_{j-1,t}^{move} \cup W_{j,t}^{stay}$, for $j \geq 2$, $V_{j,t+1}$ can be expressed as follows: for $j \geq 2$,

$$\begin{aligned} V_{j,t+1} &= \cup_{l=j}^{M-K+1} W_{l,t+1} = \cup_{l=j}^{M-K+1} \left(W_{l-1,t}^{move} \cup W_{l,t}^{stay} \right) \\ &= W_{j-1,t}^{move} \cup \left(W_{j,t}^{stay} \cup W_{j,t}^{move} \right) \cup \cdots \cup \left(W_{M-K,t}^{stay} \cup W_{M-K,t}^{move} \right) \cup W_{M-K+1,t}^{stay} \\ &\subset V_{j-1,t} - Z_{t+1} \\ &= V_{j-1,t} - (Z_{t+1} \cap V_{j-1,t}). \end{aligned} \tag{21}$$

Let $j_{\max} = \max\{j \in [M] \mid W_{j,t} \neq \emptyset\}$. We set $j_{\max} = 0$ if the set is empty. If $j_{\max} \neq 0$, then $W_{j_{\max},t} \neq \emptyset$. We will show that $W_{j_{\max},t} \neq \emptyset$ induces $W_{j_{\max},t} \cap Z_{t+1} \neq \emptyset$ holds by the following reasoning. If $C(t+1) \cap W_{j_{\max},t} \neq \emptyset$, then $W_{j_{\max},t} \cap Z_{t+1} \neq \emptyset$. If $C(t+1) \cap W_{j_{\max},t} = \emptyset$, then $\gamma(t)$ must be the arm in $W_{j_{\max},t}t$ whose unfairness index is the highest among the arms $W_{j_{\max},t}t$, (i.e., $\gamma(t) = \arg\max f_i(t) \,|i \in W_{j_{\max},t}t$) by the definitions of $j_{\max}$ and $\gamma(t)$. Therefore $W_{j_{\max},t} \cap Z_{t+1} \neq \emptyset$:

Assume $j_{\max} \geq 1$. Then for $j \geq j_{\max} + 1$, then $W_{j,t} = \emptyset$ for $j \geq j_{\max} + 1$. Thus $V_{j,t} = \emptyset$ for $j \geq j_{\max} + 1$. Consider $j \leq j_{\max}$. Then $V_{j,t} \neq \emptyset$ since $W_{j_{\max},t} \subset V_{j,t}$.

We are ready to show (b). For $j \geq j_{\max} + 2$, by (21), $V_{j,t+1} \subset V_{j-1,t} - Z(t+1) = \emptyset$ since $V_{k,t} = \emptyset$ for $k \geq j_{\max} + 1$.

For $2 \leq j \leq j_{\max}+1$, by (21), it holds that $V_{j,t+1} \subset V_{j-1,t} - Z(t+1) = V_{j-1,t} - Z_{t+1} \cap V_{j-1,t}$. Note that $W_{j_{\max}} \subset V_{j-1,t}$ and $W_{j_{\max},t} \cap Z_{t+1} \neq \emptyset$, implying $|Z_{t+1} \cap V_{j-1,t}| \geq 1$. Hence $|V_{j,t+1}| \leq |V_{j-1,t}| - 1 \leq M - K + 1 - j$,

Consider the case $j_{\max} = 0$. i.e, the set $\{j \in [M] \mid W_{j,t} \neq \emptyset\}$ is empty. Then $W_{0,t} = [M]$, which implies that $V_{j,t} = \emptyset$ for $j \geq 1$. For $j \geq 2$, it holds that $V_{j,t+1} \subset V_{j-1,t} - Z_{t+1} = \emptyset$ by (21). Thus $|V_{2,t+1}| = 0 < M - K + 1$.

We have proved (b).

D. Proof of Theorem 3.5

For simple notation, we use $\mathcal{R}(T)$ for the regret of Fair-MMAB(K)-UCB1. By Theorem 3.4, it is enough to show that $\mathcal{R}(T)$ has $O(1)$ fairness aware regret bound. Recall that $Z(t)$ is the recommendation list, the set of arms selected at time t by Fair-MMAB(K)-UCB1. Let $B_k(t)$ be the set of arms whose UCB index is one of the the 1^{st}, $2^{nd}, \cdots, k^{th}$ highest ones. Hence $B_{K-1}(t-1) = C(t)$ for $C(t)$ in Algorithm 1. Using $T - \mathsf{E}[n_{i,T}] = \sum_{t=1}^{T} \mathsf{E}[\mathbb{1}\{i \notin Z(t)\}] = \sum_{t=1}^{T} \mathsf{P}[i \notin Z(t)]$

and $\mathsf{E}[n_{j,T}] = \sum_{t=1}^{T} \mathsf{E}[\mathbb{1}\{j \in Z(t)\}] = \sum_{t=1}^{T} \mathsf{P}[j \in Z(t)]$, we can rewrite fairness regret $\mathcal{R}_{\boldsymbol{c}}(t)$ given by (7) as follows;

$$\begin{aligned}
\mathcal{R}_{\boldsymbol{c}}(T) &= \sum_{i=1}^{K-1}(\theta_i - \theta_K)(T - \mathsf{E}[n_{i,T}]) + \sum_{j=K+1}^{M}(\theta_K - \theta_j)(\mathsf{E}[n_{j,T}] - \lfloor c_j T \rfloor) \\
&= \sum_{i=1}^{K-1}(\theta_i - \theta_K)\sum_{t=1}^{T}\mathsf{E}[\mathbb{1}\{i \notin Z(t)\}] + \sum_{j=K+1}^{M}(\theta_K - \theta_j)\Big(\sum_{t=1}^{T}\mathsf{E}[\mathbb{1}\{j \in Z(t)]\} - \lfloor c_j T \rfloor\Big) \\
&= \sum_{i=1}^{K-1}(\theta_i - \theta_K)\sum_{t=1}^{T}\mathsf{P}[i \notin Z(t)] + \sum_{j=K+1}^{M}(\theta_K - \theta_j)\Big(\sum_{t=1}^{T}\mathsf{P}[j \in Z(t)] - \lfloor c_j T \rfloor\Big).
\end{aligned}$$

Let $1 \le i \le K-1$ and $K+1 \le j \le M$.

Consider $\mathsf{P}[i \notin Z(t)]$. Since $i \in Z(t)$ implies that $i \in B_K(t-1)$ or i has the highest positive unfairness index, it holds that

$$\begin{aligned}
\mathsf{P}[i \notin Z(t)] &\le \mathsf{P}[i \notin B_K(t-1)] \\
&= \mathsf{P}[\exists\, j \ge K+1 \text{ s.t } u_j(t-1) > u_i(t-1)] \\
&\le \sum_{j=K+1}^{M}\mathsf{P}[u_j(t-1) > u_i(t-1)].
\end{aligned}$$

Hence for any positive integer T_0, it holds that

$$\begin{aligned}
\sum_{t=1}^{T}\mathsf{P}[i \notin Z(t)] &\le T_0 + \sum_{t=T_0+1}^{T}\mathsf{P}[i \notin Z(t)] \\
&\le T_0 + \sum_{j=K+1}^{M}\phi_{ij}(T_0)
\end{aligned}$$

where $\phi_{ij}(T_0) = \sum_{t=T_0+1}^{T}\mathsf{P}[u_j(t-1) \ge u_i(t-1)]$.

Consider $\mathsf{P}[j \in Z(t)]$.

$$\begin{aligned}
\mathsf{P}[j \in Z(t)] &\le \mathsf{P}[j \in B_K(t-1)] + \mathsf{P}[j \in \arg\max_{i \in F(t)} f_i(t-1)] \\
&\le \mathsf{P}[j \in B_K(t-1)] + \mathsf{P}[n_{j,t-1} < c_j(t-1)] \\
&= \mathsf{P}[j \in B_K(t-1)] + \mathsf{P}[n_{j,t-1} = \lfloor c_j(t-1) \rfloor] \qquad (\because n_{j,t-1} \ge \lfloor c_j(t-1) \rfloor \text{ by Theorem 3.4}) \\
&\le \sum_{i=1}^{K}\mathsf{P}[u_j(t-1) \ge u_i(t-1)] + \mathsf{P}[n_{j,t-1} = \lfloor c_j(t-1) \rfloor].
\end{aligned}$$

Hence we can write $\sum_{t=1}^{T}\mathsf{P}[j \in Z(t)]$ as

$$\begin{aligned}
\sum_{t=1}^{T}\mathsf{P}[j \in Z(t)] &\le T_0 + \sum_{t=T_0+1}^{T}\mathsf{P}[j \in Z(t)] \\
&= T_0 + \sum_{i=1}^{K}\phi_{ij}(T_0) + \psi_j(T_0).
\end{aligned}$$

where $\psi_j(T_0) = \sum_{t=T_0+1}^{T}\mathsf{P}[n_{j,t-1} = \lfloor c_j(t-1) \rfloor]$. For simple notations, we use ϕ_{ij} and ψ_j by dropping T_0 from now on.

Summarizing all these, we have

$$\mathcal{R}_{\boldsymbol{c}}(T) \le \sum_{i=1}^{K-1}(\theta_i - \theta_K)A_1 + \sum_{j=K+1}^{M}(\theta_K - \theta_j)A_2 \tag{22}$$

where $A_1 = T_0 + \sum_{j=K+1}^{M}\phi_{ij}$ and $A_2 = T_0 + \sum_{i=1}^{K}\phi_{ij} + \psi_j(T_0) - \lfloor c_j T \rfloor$. If we properly select T_0, then we can show that the followings hold

i) $\phi_{ij} \le \frac{\pi^2}{3}$ for $i \le K-1$, $j \ge K+1$,

ii) $\psi_j \le c_j(T-T_0)+1$ for $j \ge K+1$.

If i) and ii) hold, then $A_1 \le T_0 + \frac{(M-K)\pi^2}{3}$ and $A_2 < (1-c_j)T_0 + \frac{K\pi^2}{3} + 2$, hence $\mathcal{R}_{\boldsymbol{c}}(T) < \infty$, for $T > T_0$.

We choose a proper $T_0 < T$. Let $\varepsilon < \frac{\theta_K - \theta_{K+1}}{2}$ and $c_{\min} = \min_{1\le i\le M} c_i$. Since $\lim_{t\to\infty} \frac{\ln t}{t} = 0$, there exists a positive integer $T_0 \ge \frac{2}{c_{\min}}$ such that $\frac{\ln t}{t} \le \frac{c_{\min}\varepsilon^2}{32}$ for any $t \ge T_0$, which is the T_0 we want if $T_0 < T$. We check if $T_0 < T$ holds. Since $c_{\min}$ is independent of T, the value $\frac{2}{c_{\min}}$ (used in the condition $T_0 \ge \frac{2}{c_{\min}}$) is also independent of T . Hence $T_0 < T$ if T is sufficiently large. For this T_0, if $t \ge T_0$ and $s \ge \lfloor c_i t \rfloor$, then $s \ge \lfloor c_i t \rfloor \ge \lfloor c_{\min} t \rfloor \ge 2$ and

$$\sqrt{\frac{2\ln t}{s}} \le \sqrt{\frac{2\ln t}{\lfloor c_i t \rfloor}} = \sqrt{\frac{2\ln t}{t} \cdot \frac{t}{\lfloor c_i t \rfloor}} \le \frac{\varepsilon}{4}\sqrt{\frac{c_{\min} t}{\lfloor c_i t \rfloor}} \le \frac{\varepsilon}{2}. \tag{23}$$

With this choice of T_0, we prove that i) and ii) hold. Let $\hat{\theta}_{i,s_i} = \frac{1}{s_i}\sum_{k=1}^{s_i} X_{i,k}$ for $1 \le i \le M$. Recall that UCB index for Fair-MMAB(K)-UCB1 is $u_i(t) = \hat{\theta}_{i,n_{i,t}} + \sqrt{\frac{2\ln t}{n_{i,t}}}$ and that the unfairness index is $f_i(t) = c_i t - n_{i,t}$ for arm i.

Theorem D.1. *(Chernoff-Hoeffding bound) Let $X_1,\ X_2, \cdots, X_n$ be a sequence of independently identically distributed random variables with $a \le X_k \le b$ with $\mathsf{E}[X_k] = \theta$ for all k. For any $s > 0$, it holds that*

$$\mathsf{P}\Big[\frac{1}{n}\sum_{k=1}^{n} X_k \ge \theta + \varepsilon\Big] \le e^{-\frac{2n\varepsilon^2}{(b-a)^2}},$$
$$\mathsf{P}\Big[\frac{1}{n}\sum_{k=1}^{n} X_k \le \theta - \varepsilon\Big] \le e^{-\frac{2n\varepsilon^2}{(b-a)^2}}.$$

Proof of i):
We prove that $\phi_{ij} = \sum_{t=T_0+1}^{T} \mathsf{P}[u_j(t-1) \ge u_i(t-1)] \le \frac{\pi^2}{3}$ for $i \le K-1$ and $j \ge K+1$. Let $\bar{X}_{k,s_k} = \hat{\theta}_{k,s_k} + \sqrt{\frac{2\ln(t-1)}{s_k}}$ for $k \in [M]$. Since $i \le K-1$ and $j \ge K+1$, we have

$$\begin{aligned}
\phi_{ij} &= \sum_{t=T_0+1}^{T} \mathsf{P}[u_j(t-1) \ge u_i(t-1)] \\
&= \sum_{t=T_0+1}^{T} \mathsf{P}\Big[u_j(t-1) \ge u_i(t-1),\ n_{j,t-1} \ge \lfloor c_j(t-1)\rfloor,\ n_{i,t-1} \ge \lfloor c_i(t-1)\rfloor\Big] \quad \text{(by Theorem 3.4)} \\
&\le \sum_{t=T_0+1}^{T} \mathsf{P}[\max_{\lfloor c_j(t-1)\rfloor \le s_j < t} \bar{X}_{j,s_j} \ge \min_{\lfloor c_i(t-1)\rfloor \le s_i < t} \bar{X}_{i,s_i}] \\
&\le \sum_{t=T_0+1}^{T} \sum_{s_i=\lfloor c_i(t-1)\rfloor}^{t-1} \sum_{s_j=\lfloor c_j(t-1)\rfloor}^{t-1} \mathsf{P}[\bar{X}_{j,s_j} \ge \bar{X}_{i,s_i}]
\end{aligned} \tag{24}$$

The event $\bar{X}_{j,s_j} \ge \bar{X}_{i,s_i}$ implies that at least one of the following events should holds

*E*1: $\hat{\theta}_{i,s_i} \le \theta_i - \sqrt{\frac{2\ln(t-1)}{s_i}}$,

*E*2: $\hat{\theta}_{j,s_j} \ge \theta_j + \sqrt{\frac{2\ln(t-1)}{s_j}}$,

*E*3: $\theta_j + 2\sqrt{\frac{2\ln(t-1)}{s_j}} > \theta_i$.

Hence (24) becomes

$$\sum_{t=T_0+1}^{T} \mathsf{P}[u_j(t-1) > u_i(t-1)] \quad \leq \quad \sum_{t=T_0+1}^{T} \sum_{s_i=\lfloor c_i(t-1)\rfloor}^{t-1} \sum_{s_j=\lfloor c_j(t-1)\rfloor}^{t-1} \mathsf{P}[E1] + \mathsf{P}[E2] + \mathsf{P}[E3]$$

By Chernoff-Heoffding bound, $\mathsf{P}[E1] \leq (t-1)^{-4}$ and $\mathsf{P}[E2] \leq (t-1)^{-4}$. We will prove that $\mathsf{P}[E3] = 0$. Since $s_j \geq \lfloor c_j(t-1)\rfloor$ and $t \geq T_0 + 1$, we have

$$\theta_j + 2\sqrt{\frac{2\ln(t-1)}{s_j}} < \theta_j + 2\sqrt{\frac{2\ln(t-1)}{\lfloor c_j(t-1)\rfloor}} < \theta_j + \varepsilon < \theta_j + \frac{\theta_K - \theta_{K+1}}{2} < \theta_K. \tag{25}$$

The first inequality holds by (23) and the last inequality holds because $j \geq K+1$ (i.e., $\theta_j < \theta_K$). From (25), we have $\theta_j + 2\sqrt{\frac{2\ln(t-1)}{s_j}} < \theta_K < \theta_i$. Hence $\mathsf{P}[E3] = 0$. With $\mathsf{P}[E1] \leq (t-1)^{-4}$, $\mathsf{P}[E2] \leq (t-1)^{-4}$, $\mathsf{P}[E3] = 0$, we have

$$\begin{aligned}
\phi_{ij} = \sum_{t=T_0+1}^{T} \mathsf{P}[u_j(t-1) \geq u_i(t-1)] \quad &\leq \quad \sum_{t=T_0+1}^{T} \sum_{s_i=\lfloor c_i(t-1)\rfloor}^{t-1} \sum_{s_j=\lfloor c_j(t-1)\rfloor}^{t-1} 2(t-1)^{-4} \\
&\leq \quad \sum_{t=1}^{\infty} \sum_{s_i=\lfloor c_i(t-1)\rfloor}^{t-1} \sum_{s_j=\lfloor c_j(t-1)\rfloor}^{t-1} 2t^{-4} \\
&\leq \quad \frac{\pi^2}{3}. \qquad (\because \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6})
\end{aligned} \tag{26}$$

Proof of ii):
We should show that $\psi_j \leq c_j(T - T_0) + 1$ for $j \geq K+1$. Recall $\psi_j = \sum_{t=T_0+1}^{T} \mathsf{P}[n_{j,t-1} = \lfloor c_j(t-1)\rfloor]$. Let $E_f(s)$ be the event that arm j is selected at time s only when $n_{j,s-1} = \lfloor c_j(s-1)\rfloor$. Then $\psi_j = \sum_{t=T_0+1}^{T} \mathsf{E}[\mathbb{1}\{E_f(t)\}]$.

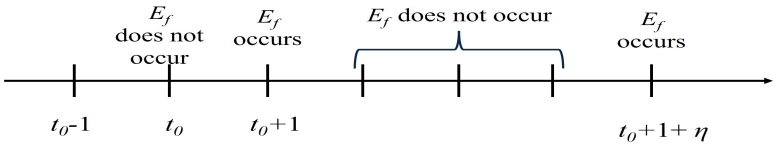


Figure 6. After $t_0 + 1$, E_f occurs for the first time at $t_0 + 1 + \eta$.

We will find an upper bound of $\sum_{t=T_0+1}^{T} \mathbb{1}\{E_f(t)\}$. More specifically we find the maximum of $\sum_{t=T_0+1}^{T} \mathbb{1}\{E_f(t)\}$. For this objective, we assume that $n_{j,t}$ increases by 1 only when $E_f(t)$ occurs and count how many times event $E_f(t)$ occurs during $T_0 + 1 \leq t \leq T$. We temporarily assume that $n_{j,t}$ does not increase when $j \in B_{K-1}(t)$ or $j \in B_K(t)$. If $n_{j,t}$ increases when $j \in B_{K-1}(t)$ or $j \in B_K(t)$, then the number that event E_f occurs gets reduced.

To find the maximum number that event E_f take places, we consider the following scenario for some given time t_0;

(a) At t_0, event E_f does not occur. Therefore, $n_{j,t_0-1} \geq \lfloor c_j(t_0-1)\rfloor + 1$ and $n_{j,t_0} = n_{j,t_0-1}$.

(b) At $t_0 + 1$, event E_f does occur. Therefore, $n_{j,t_0} = \lfloor c_j t_0\rfloor$ and $n_{j,t_0+1} = n_{j,t_0} + 1$.

(c) Let $s = t_0 + 1 + \eta$ be the earliest time when E_f occurs after $t_0 + 1$ (see Fig. 6); that is, s is the smallest positive integer such that $s > t_0 + 1$ and $n_{j,s-1} = \lfloor c_j(s-1)\rfloor$. This means that $n_{j,\tau} = n_{j,\tau-1} = n_{j,t_0+1}$ and $n_{j,\tau-1} \neq \lfloor c_j(\tau-1)\rfloor$ (i.e., $n_{j,\tau} \geq \lfloor c_j\tau\rfloor + 1$) for $t_0 + 2 \leq \tau \leq s-1$.

Let $\kappa = \lfloor c_j t_0\rfloor$. Then $c_j t_0 = \kappa + h_{t_0}$ for $0 \leq h_{t_0} < 1$. From (a) and (b), $\lfloor c_j(t_0-1)\rfloor + 1 \leq n_{j,t_0-1} = n_{j,t_0} = \lfloor c_j t_o\rfloor$ and $n_{j,t_0+1} = n_{j,t_0} + 1$. Summarizing this, we have $n_{j,t_0-1} = n_{j,t_0} = \kappa$ and $n_{j,t_0+1} = \kappa + 1$. From $\kappa = n_{j,t_0-1} = \lfloor c_j(t_0-1)\rfloor + 1$ and $n_{j,t_0-1} = \kappa$, it holds that $\kappa \geq \lfloor c_j(t_0-1)\rfloor + 1 = \lfloor \kappa + h_{t_0} - c_j\rfloor + 1$, which implies that $h_{t_0} < c_j$ holds.

From (c), it holds that $n_{j,s-1} = n_{j,t_0+1}$, which implies that $n_{j,s-1} = \kappa + 1$. At time s, event E_f occur, i.e., $\lfloor c_j(s-1) \rfloor = n_{j,s-1}$. Since $n_{j,s-1} = \kappa + 1$ and $\lfloor c_j(s-1) \rfloor = \lfloor c_j(t_0+\eta) \rfloor = \lfloor \kappa + h_{t_0} + c_j\eta \rfloor$, it should hold $\kappa + 1 = \lfloor \kappa + h_{t_0} + c_j\eta \rfloor$, which implies that $1 \le h_{t_0} + c_j\eta < 2$. Note η is the smallest integer such that $1 \le h_{t_0} + c_j\eta < 2$ since η is the smallest integer such that E_f occurs at $t_0 + 1 + \eta$., Now η should meet

$$\eta > \frac{1}{c_j} - \frac{h_{t_0}}{c_j} > \frac{1}{c_j} - 1 \qquad (\because h_{t_0} < c_j). \tag{27}$$

Since η is the smallest positive integer satisfying (27), it must hold that $\eta = \lceil \frac{1}{c_j} \rceil$. This implies that event E_f occurs every $\eta = \lceil \frac{1}{c_j} \rceil$ time steps. The maximum number that event E_f occurs during $T_0+1 \le t \le T$ at most $\frac{T-T_0}{\lceil \frac{1}{c_j} \rceil} + 1 < c_j(T-T_0)+1$. Therefore the following holds

$$\psi_j(T_0) = \sum_{t=T_0+1}^{T} \mathsf{P}[n_{j,t-1} = c_j(t-1)] < c_j(T-T_0) + 1. \tag{28}$$

E. Proof of Theorem 3.6

For simple notation, we use $\mathcal{R}(T)$ for the regret of Fair-MMAB(K)-KL-UCB. Since uniform fairness property holds by Theorem 3.4, it is enough to show that $\mathcal{R}(T)$ has $O(1)$ fairness aware regret bound. We need some lemmas on the properties of KL-divergence $d(p,q)$ with $p,\ q \in [0,1]$.

Lemma E.1. *(Lemma 10.2 of (Lattimore & Szepesvári, 2020)) Let $p, q, \delta \in [0,1]$.*

i) The functions $d(\cdot, q)$ and $d(p, \cdot)$ are convex.

ii) For fixed p, the function $d(p,q)$ is increasing with $q \in [p,1]$ and decreasing with $q \in (0,p]$.

iii) For fixed q, the function $d(p,q)$ is decreasing with $p \in (0,q]$ and increasing with $p \in [q,1]$.

iv) $d(p,q) \ge 2(p-q)^2$ (Pinsker's inequality).

v) If $p \le q - \delta \le q$, then $d(p, q-\delta) < d(p,q) - d(q-\delta, q) \le d(p,q) - 2\delta^2$.

Lemma E.2. *(Lemma 10.3, Corollary 10.4, and Notes 1 on page 118 of (Lattimore & Szepesvári, 2020)) Let $X_1, X_2, \cdots, X_n$ be a sequence of i. i. d. random variables with $E[X_1] = \theta$ and $X_1 \in [0,1]$ almost surely. Let $\hat{\theta} = \frac{1}{n}\sum_{t=1}^{n} X_t$ be the empirical mean. Then, the followings are hold;*

i) $\mathsf{P}[\hat{\theta} \ge \theta + \varepsilon) \le e^{-nd(\theta+\epsilon, \theta))}$ for $\varepsilon \in [0,\ 1-\theta]$.

ii) $\mathsf{P}[\hat{\theta} \le \theta - \varepsilon] \le e^{-nd(\theta-\varepsilon, \theta)}$ for $\varepsilon \in [0,\ \theta]$.

iii) $\mathsf{P}[d(\hat{\theta}, \theta) \ge a,\ \hat{\theta} \le \theta] \le e^{-na}$ for any $a \ge 0$.

iv) $\mathsf{P}[d(\hat{\theta}, \theta) \ge a,\ \hat{\theta} \ge \theta] \le e^{-na}$ for any $a \ge 0$.

Recall that KL-UCB upper bound index is defined as $u_i(t) = \max\{q > \hat{\theta}_{i,n_{i,t}} \mid d(\hat{\theta}_{i,n_{i,t}}, q) \le \frac{\ln f(t)}{n_{i,t}}\}$ where $f(t) = t\ln^2 t$ and that $\mathcal{R}(T)$ is given by (22). As we did in the proof of Theorem 3.5, we will find each upper bound of ϕ_{ij} and ψ_j for $i \le K-1$ and $K+1 \le j$ after properly selecting T_0 for Fair-MMAB(K)-KL-UCB. Since ψ_j is independent of the definition of $u_i(t)$, (28) is valid for Fair-MMAB(K)-KL-UCB. It is enough to find a finite upper bound of ϕ_{ij} for $i \le K-1$ and $j \ge K+1$. More specifically, with a proper choice of T_0, we will show that for $i \le K-1$ and $j \ge K+1$,

$$\phi_{ij}(T_0) \ = \ \sum_{t=T_0+1}^{T} \mathsf{P}[u_j(t-1) \ge u_i(t-1)] \ \le \ \frac{1}{4\varepsilon_1^2 c_j} + \frac{1}{2\varepsilon_1^2 \ln T_0}$$

where $\varepsilon_1 = \min_{i \lneq K \lneq j} \frac{\theta_i - \theta_j}{4}$ (note that $i \lneq K \lneq j$ is equivalent to that $i \leq K-1$ and $j \geq K+1$.) Then $\mathcal{R}_{\boldsymbol{c}}(T)$ is bounded as below

$$\mathcal{R}_{\boldsymbol{c}}(T) \leq \sum_{i=1}^{K-1} (\theta_i - \theta_K) B_1 + \sum_{j=K+1}^{M} (\theta_K - \theta_j) B_2.$$

where $B_1 = T_0 + (M-K)\frac{1}{4\varepsilon_1^2}\big(\frac{1}{c_j} + \frac{2}{\ln T_0}\big)$ and $B_2 = (1-c_j)T_0 + \frac{K}{2\varepsilon_1^2}\big(\frac{1}{c_j} + \frac{2}{\ln T_0}\big)$.

Note that $\theta_i - \varepsilon_1 > \theta_j + \varepsilon_1$ for $i \lneq K \lneq j$ by the choice of ε_1 (recall that $\theta_i > \theta_j$ for $i < j$.)

Let $0 < \varepsilon_0 < \min_{i \lneq K \lneq j} d(\theta_j + \varepsilon_1, \theta_i - \varepsilon_1)$. Since $\lim_{t\to\infty} \frac{\ln(t \ln^2 t)}{t} = 0$, there exists $T_0 \geq \frac{2}{c_{\min}}$ such that $\frac{\ln(t \ln^2 t)}{t} < \frac{\varepsilon_0 c_{\min}}{4}$ for any $t \geq T_0$. Note that it holds $T_0 < T$ if T is sufficiently large, since $\frac{2}{c_{\min}}$ is independent of T. Then for any $t \geq T_0$, $s \geq \lfloor c_i t \rfloor$ and $1 \leq i \leq M$, it holds that

$$\frac{\ln f(t)}{s} \leq \frac{\ln f(t)}{\lfloor c_i t \rfloor} = \frac{\ln f(t)}{t} \cdot \frac{t}{\lfloor c_i t \rfloor} \leq \frac{\varepsilon_0}{4} \frac{c_{\min} t}{\lfloor c_i t \rfloor} \leq \frac{\varepsilon_0}{4} \frac{c_{\min} t}{\lfloor c_{\min} t \rfloor} \leq \frac{\varepsilon_0}{2}. \tag{29}$$

We are ready to show that $\phi_{ij}(T_0) \leq \frac{1}{4\varepsilon_1^2 c_j} + \frac{1}{2\varepsilon_1^2 \ln T_0}$ for $i \leq K-1$ and $j \geq K+1$. It obviously holds that

$$\phi_{ij}(T_0) = \mathsf{P}[u_j(t-1) \geq u_i(t-1)] = G(t-1) + H(t-1)$$

where

$$\begin{aligned} G(t-1) &= \mathsf{P}[u_j(t-1) \geq u_i(t-1) \geq \theta_i - \varepsilon_1], \\ H(t-1) &= \mathsf{P}[u_j(t-1) \geq u_i(t-1), u_i(t-1) < \theta_i - \varepsilon_1]. \end{aligned}$$

Moreover, $G(t-1) = G_1(t-1) + G_2(t-1)$ where

$$\begin{aligned} G_1(t-1) &= \mathsf{P}[u_j(t-1) \geq u_i(t-1) \geq \theta_i - \varepsilon_1, \hat{\theta}_{j,n_{j,t-1}} < \theta_j + \varepsilon_1] \\ G_2(t-1) &= \mathsf{P}[u_j(t-1) \geq u_i(t-1) \geq \theta_i - \varepsilon_1, \hat{\theta}_{j,n_{j,t-1}} \geq \theta_j + \varepsilon_1]. \end{aligned}$$

Hence $\phi_{ij}(T_0) = \sum_{t=T_0+1}^{T} G_1(t-1) + G_2(t-1) + H(t-1)$.

We will prove that

i) $\sum_{t=T_0+1}^{T} G_1(t-1) = 0$,

ii) $\sum_{t=T_0+1}^{T} G_2(t-1) \leq \frac{1}{4\varepsilon_1^2 c_j}$,

iii) $\sum_{t=T_0+1}^{T} H(t-1) \leq \frac{1}{2\varepsilon_1^2 \ln T_0}$.

Proof of i):
Recall that $\theta_i - \varepsilon_1 > \theta_j + \varepsilon_1$. Hence it holds that

$$\begin{aligned} G_1(t-1) &= \mathsf{P}[u_j(t-1) \geq u_i(t-1) \geq \theta_i - \varepsilon_1 \geq \theta_j + \varepsilon_1 \geq \hat{\theta}_{j,n_{j,t-1}}] \\ &\leq \mathsf{P}[d(\hat{\theta}_{j,n_{j,t-1}}, \theta_j + \varepsilon_1) < d(\hat{\theta}_{j,n_{j,t-1}}, u_j(t-1)) - d(\theta_j + \varepsilon_1, u_j(t-1))] \quad \text{(by Lemma E.1)} \\ &\leq \mathsf{P}\Big[d(\hat{\theta}_{j,n_{j,t-1}}, \theta_j + \varepsilon_1) < \frac{\ln f(t-1)}{n_{j,t-1}} - d(\theta_j + \varepsilon_1, \theta_i - \varepsilon_1)\Big] \\ &\qquad \text{(by the definition of } u_j(t-1) \text{ and } d(p,q) \text{ is decreasing with } q \text{ for fixed } p \text{ and } q \in [p,1]) \\ &\leq \mathsf{P}\Big[d(\theta_j + \varepsilon_1, \theta_i - \varepsilon_1) < \frac{\ln f(t-1)}{n_{j,t-1}}\Big]. \end{aligned}$$

Now we have

$$\sum_{t=T_0+1}^{T} G_1(t-1) \;\leq\; \sum_{t=T_0+1}^{T} \mathsf{P}\Big[d(\theta_j+\varepsilon_1,\theta_i-\varepsilon_1) < \frac{\ln f(t-1)}{n_{j,t-1}}\Big] \;=\; 0$$

since $t \geq T_0+1$ and $\frac{\ln f(t)}{n_{j,t}} < \varepsilon_0$ for $t \geq T_0$, $\varepsilon_0 < \min_{i<j} d(\theta_j+\varepsilon_1,\theta_i-\varepsilon_1)$.

Proof of ii)

$$\begin{aligned}
G_2(t-1) &\leq \mathsf{P}[\hat{\theta}_j, n_{j,t} \geq \theta_j+\varepsilon_1] \\
&= \sum_{s=\lfloor c_j(t-1)\rfloor}^{t-1} \mathsf{P}[\hat{\theta}_j, n_{j,t} \geq \theta_j+\varepsilon_1, n_{j,t}=s] \quad \text{(By Chernoff-Hoeffding bound)} \\
&\leq \sum_{s=\lfloor c_j(t-1)\rfloor}^{t-1} \mathsf{P}[\hat{\theta}_{j,s} \geq \theta_j+\varepsilon_1].
\end{aligned}$$

$$\begin{aligned}
\sum_{t=T_0+1}^{T} G_2(t-1) &\leq \sum_{t=T_0}^{T-1}\sum_{s=\lfloor c_j t\rfloor}^{t} \mathsf{P}[\hat{\theta}_{j,s} \geq \theta_j+\varepsilon_1] \\
&\leq \sum_{t=T_0}^{T-1}\sum_{s=\lfloor c_j t\rfloor}^{t} e^{-2s\varepsilon_1^2} \\
&\leq \sum_{t=T_0}^{T-1} \frac{e^{-2\lfloor c_j t\rfloor\varepsilon_1^2}}{2\varepsilon_1^2} \\
&\leq \sum_{t=T_0}^{T-1} \frac{e^{-2c_j(t-1)\varepsilon_1^2}}{2\varepsilon_1^2} \qquad \text{(since } \lfloor c_j t\rfloor = c_j t - h_0 \text{ for } 0\leq h_0\leq 1\text{)} \\
&\leq \frac{e^{-2c_j(T_0-1)\varepsilon_1^2}}{4\varepsilon_1^2 c_j}.
\end{aligned}$$

Proof of iii):

$$\begin{aligned}
H(t-1) &\leq \mathsf{P}[u_i(t-1) < \theta_i-\varepsilon_1] \\
&= \mathsf{P}[u_i(t-1) < \theta_i-\varepsilon_1,\ \hat{\theta}_{i,n_{i,t-1}} < u_i(t-1)] \\
&\leq \mathsf{P}[d(\hat{\theta}_{i,n_{i,t-1}},\ \theta_i-\varepsilon_1) \geq \frac{\ln f(t-1)}{n_{i,t-1}},\ \hat{\theta}_{i,n_{i,t-1}} < \theta_i-\varepsilon_1] \quad \text{(by the definition of } u_i(t-1)\text{)} \\
&= \sum_{s=\lfloor c_j(t-1)\rfloor}^{t-1} \mathsf{P}[d(\hat{\theta}_{i,s},\theta_i-\varepsilon_1) \geq \frac{\ln f(t-1)}{s},\ \hat{\theta}_{i,s} < \theta_i-\varepsilon_1,\ n_{i,t-1}=s] \quad \text{(since } \lfloor c_i(t-1) \leq n_{i,t-1} \leq t-1\text{)} \\
&\leq \sum_{s=\lfloor c_j(t-1)\rfloor}^{t-1} \mathsf{P}[d(\hat{\theta}_{i,s},\theta_i-\varepsilon_1) \geq \frac{\ln f(t-1)}{s},\ \hat{\theta}_{i,s} < \theta_i-\varepsilon_1] \\
&\leq \sum_{s=\lfloor c_j(t-1)\rfloor}^{t-1} \mathsf{P}[d(\hat{\theta}_{i,s},\theta_i) \geq \frac{\ln f(t-1)}{s}+2\varepsilon_1^2, \hat{\theta}_{i,s} < \theta_i-\varepsilon_1^2]. \quad \text{(by Lemma E.1)}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}\sum_{t=T_0+1}^{T} H(t-1) &= \sum_{t=T_0}^{T-1}\sum_{s=\lfloor c_j t\rfloor}^{t} \mathsf{P}[d(\hat{\theta}_{i,s},\theta_i) \geq \frac{\ln f(t-1)}{s} + 2\varepsilon_1^2, \hat{\theta}_{i,s} < \theta_i - \varepsilon_1^2] \\ &\leq \sum_{t=T_0}^{T-1}\sum_{s=\lfloor c_j t\rfloor}^{t} \frac{-2s\varepsilon_1^2}{f(t)} \\ &\leq \frac{1}{2\varepsilon_1^2}\int_{t=T_0}^{T-1} \frac{1}{t\ln^2 t} dt \\ &\leq \frac{1}{2\varepsilon_1^2 \ln T_0}.\end{aligned}$$

F. Proof of Theorem 4.1

We prove Theorem 4.1 in a similiar way as do Theorem 3.4.

Recall that L is the unique integer such that $\frac{L-1}{M-K+L-1} \leq c_{\max} < \frac{L}{M-K+L}$, and that $1 \leq l \leq K \leq \frac{M}{2}$. Hence $0 < c_k < \frac{L}{M-K+L}$ for any $k \in [M]$. Then $0 < c_i < \frac{L}{M-K+L}$ for $i \in [M]$. Let $q_j = \frac{j}{M-K+L}$ for $0 \leq j \leq M-K+L$. At each round t, we define $W_{j,t}$ for $0 \leq j \leq M-K+L$ as below

$$\begin{aligned}W_{0,t} &= \{k \in [M] \mid c_k t - n_{k,t} < 0\}, \\ W_{j,t} &= \{k \in [M] \mid q_{j-1} \leq c_k t - n_{k,t} < q_j\}, \quad \text{for } 1 \leq j \leq M-K+L.\end{aligned}$$

We will prove Lemma F.1 using the mathematical induction.

Lemma F.1. *Let $V_{j,t} = \dot{\cup}_{l=j}^{M-K+L} W_{l,t}$ for $0 \leq j \leq M-K+L$. For $t \geq 1$, it holds that*

(a) $V_{0,t} = [M]$

(b) $|V_{j+1,t}| \leq M-K+L-j$ *for* $L \leq j \leq M-K$.

To prove Lemma F.1, we need Lemma F.2 summarizing what happens when Fair-MMAB(K)-MF is applied. For simple notation, we will use Z_t instead of $Z(t)$ in Fair-MMAB(K)-MF in this proof. Recall that $A \dot{\cup} B$ means a disjoint union of A and B ($A \cap B = \emptyset$).

Lemma F.2. *Let Z_{t+1} be the set of arms selected at time step $t+1$ by Fair-MMAB(K)-MF.*

(a) If $k \in Z_{t+1} \cap W_{0,t}$, then $k \in W_{0,t+1}$.

(b) If $k \in Z_{t+1} \cap W_{j,t}$, then $k \in \begin{cases} W_{0,t+1} & \text{for } 1 \leq j \leq M-K, \\ \dot{\cup}_{l=0}^{j-(M-K)} W_{0,t} \cup W_{l,t+1} & \text{for } j \geq M-K+1. \end{cases}$

(c) If $k \in W_{0,t} \setminus Z_{t+1}$, then $k \in \dot{\cup}_{l=0}^{L} W_{l,t+1}$.

(d) If $k \in W_{j,t} \setminus Z_{t+1}$ for $1 \leq j \leq M-K$, then $k \in \cup_{l=j}^{j+L} W_{l,t+1}$.

Proof. (a) Since $k \in Z_{t+1}$, it holds that $n_{k,t+1} = n_{k,t}+1$. Hence $c_k(t+1) - n_{k,t+1} = c_k t - n_{k,t} + c_k - 1 < q_0 + c_k - 1 < 0$ since $k \in W_{0,t}$ and $\frac{L}{M-K+L} < 1$. Hence $k \in W_{0,t+1}$.

(b) Since $k \in Z_{t+1}$, it holds that $c_k(t+1) - n_{k,t+1} = c_k t - n_{k,t} + c_k - 1$. From the assumption that $k \in W_{j,t}$, we have $q_{j-1} \leq c_k t - n_{k,t} < q_j$. Hence it holds that $q_{j-1} + c_k - 1 \leq c_k(t+1) - n_{k,t+1} < q_j + c_k - 1$, which implies that $q_{j-1} - 1 < c_k(t+1) - n_{k,t+1} < \frac{j+L}{M-K+L} - 1$. If $1 \leq j \leq M-K$, then $k \in W_{0,t+1}$. If $M-K+1 \leq j \leq M-K+L$, then $k \in \dot{\cup}_{l=0}^{j-(M-K)} W_{l,t+1}$.

(c) Note that $n_{k,t+1} = n_{k,t}$ since $k \notin Z_{t+1}$. Consider $c_k(t+1) - n_{k,t+1}$. It holds that $c_k(t+1) - n_{k,t+1} = c_k t - n_{k,t} + c_k < c_k$ since $k \in W_{0,t}$. Therefore, we have $k \in \dot{\cup}_{l=0}^{L} W_{l,t+1}$.

(d) Consider $c_k(t+1) - n_{k,t+1}$; $c_k(t+1) - n_{k,t+1} = c_k t - n_{k,t} + c_k$. Since $k \in W_{j,t}$, it holds that $q_{j-1} \leq c_k t - n_{k,t} < q_j$. Hence we have $q_{j-1} + c_k \leq c_k(t+1) - n_{k,t+1} < q_j + c_k$, which implies that $q_{j-1} < c_k(t+1) - n_{k,t+1} < q_{j+L}$. Therefore $k \in \dot{\cup}_{l=j}^{j+L} W_{l,t}$.

□

Proof of Lemma F.1: We use the mathematical induction to prove Lemma F.1.

At time $t = 1$**:** We should prove that (a) and (b) of Lemma F.1 hold. Recall that $0 < c_i < \frac{L}{M-K+L}$ for all $i \in [M]$.

i) For $k \in Z_1$ (arm k is selected), we have $c_k t - n_{k,t} = c_k \cdot 1 - 1 < 0$, which implies that $k \in W_{0,t}$. Hence $Z_1 \subset W_{0,1}$.

ii) For $k \notin Z_1$, we have $c_k \cdot 1 - n_{k,1} = c_k < q_L$, which implies that $k \in \dot{\cup}_{j=0}^{L} W_{j,1}$.

By the above results, i) and ii), we have $[M] = \dot{\cup}_{j=0}^{L} W_{j,1}$ and $W_{j,1} = \emptyset$ for $j \geq L+1$, which means that $|V_{j,1}| = |\cup_{l=j}^{M-K+L} W_{l,1}| = 0 < M - K + L - j + 1$ for $j \geq L+1$. We have proved that Lemma F.1 holds when $t = 1$.

At time t**:** We assume that Lemma F.1 holds at time t.

At time $t+1$**:** We have to show that Lemma F.1 still holds at time $t+1$. We first show that $V_{0,t+1} = [M]$. From the induction hypothesis at time t, we have $[M] = \cup_{l=0}^{M-K+l} W_{l,t}$. We denote $[M]$ as $[M] = W_{0,t} \bigcup \left(\cup_{l=1}^{M-K} W_{l,t} \right) \bigcup V_{M-K+1,t}$ (recall that $V_{k,t} = \cup_{l=k}^{M-K+L} W_{l,t}$). The followings are hold

(i) $W_{0,t} \subset \cup_{l=0}^{L} W_{l,t+1}$ by (a) and (c) of Lemma F.1,

(ii) $\cup_{l=1}^{M-K} W_{l,t} \cap Z_{t+1} \subset W_{0,t+1}$ by (b) of Lemma F.1,

(iii) $\cup_{l=1}^{M-K} W_{l,t} \setminus Z_{t+1} \subset \cup_{l=1}^{M-K+L} W_{l,t+1}$ by (d) of Lemma F.1, and

(iv) $V_{M-K+1,t} \subset Z_{t+1}$ and $\subset \cup_{j=0}^{L} W_{j,t+1}$.

We prove item (iv) holds. Since the induction hypothesis holds at time t, we have $|V_{M-K+1,t}| \leq L$, which implies that $V_{M-K+1,t} \subset Z_{t+1}$ because $V_{M-K+1,t}$ is the set of arms with $|V_{M-K+1,t}|$ highest positive unfairness arms and Fair-MMAB-MF(K) select arms with n_F highest positive unfairness indices; Let k be an arm k in$V_{M-K+1,t}$. If $k \in C(t+1)$, then $k \in Z_{t+1}$. Suppose that $k \notin C(t+1)$. Then $k \in F(t+1) \cap V_{M-K+1,t}$. Hence k is one of the arms with $|F(t) \cap V_{M-K+1,t}|$ highest positive unfairness indices. Obviously $|F(t) \cap V_{M-K+1,t}| \leq n_F$ since $|V_{M-K+1,t}| \leq L$. Hence $k \in Z_{t+1}$ By (b) of Lemma F.1, it holds that $V_{M-K+1,t} \subset \cup_{j=0}^{L} W_{j,t+1}$. By (i)-(iv), it obviously holds that $[M] = V_{0,t+1}$.

We will show part (b), $|V_{j+1,t}| \leq M - K + L - j$ for $L \leq j \leq M - K$. Consider l such that $L \leq j \leq M - K$. Note that $V_{j+1,t} = \cup_{l=j+1}^{M-K} W_{l,t} \cup V_{M-K+1,t}$ by the definition of $V_{l,t}$. Using item (iv) and Lemma F.1, we know that $V_{j+1,t+1} \subset V_{j+1-L,t}$. Then, it holds that $V_{j+1,t+1} \cap V_{M-K+1,t} = \emptyset$ by (iv) and $L \leq j \leq M - K$, which implies that $V_{j+1,t+1} \subset (V_{j+1-L,t} \setminus V_{M-K+1,t}) = \cup_{l=j+1-L}^{M-K} W_{l,t}$. Note that $\cup_{l=j+1-L}^{M-K} W_{l,t} \cap Z_{t+1} \subset W_{0,t+1}$ by (d) of Lemma F.1. Hence $V_{j+1-L,t+1} \subset V_{j+1-L,t} - Z_{t+1}$.

If $|V_{j+1-L,t}| \leq L$, then $V_{j+1-L,t} \subset Z_{t+1}$ according to Fair-MMAB(K)-MF. Hence $|V_{j+1-L,t+1}| = 0$.

If $|V_{j+1-L,t}| > L$, then Fair-MMAB(K)-MF selects at least L arms among $V_{j+1-L,t}$. That is $|Z_{t+1} \cap V_{j+1-L,t}| \geq L$. Therefore, $|V_{j+1-L,t} - Z_{t+1}| \leq |V_{j+1-L,t}| - L \leq M - K + L - j$.

We have proved the part (b). □

G. Additional Experiments

G.1. Brief Summary of LFG, UCB-LP, and UCB-PLLP

We provide a brief summary of LFG of (Li et al., 2019), UCB-LP and UCB-PLLP in (Liu et al., 2022). LFG, UCB-LP, and UCB select multiple arms, but no more than K arms and they seek long-term fairness requirements, $\liminf_{T\to\infty} \frac{\mathsf{E}[n_{i,T}]}{T} \geq c_i$ for all $i \in [M]$.

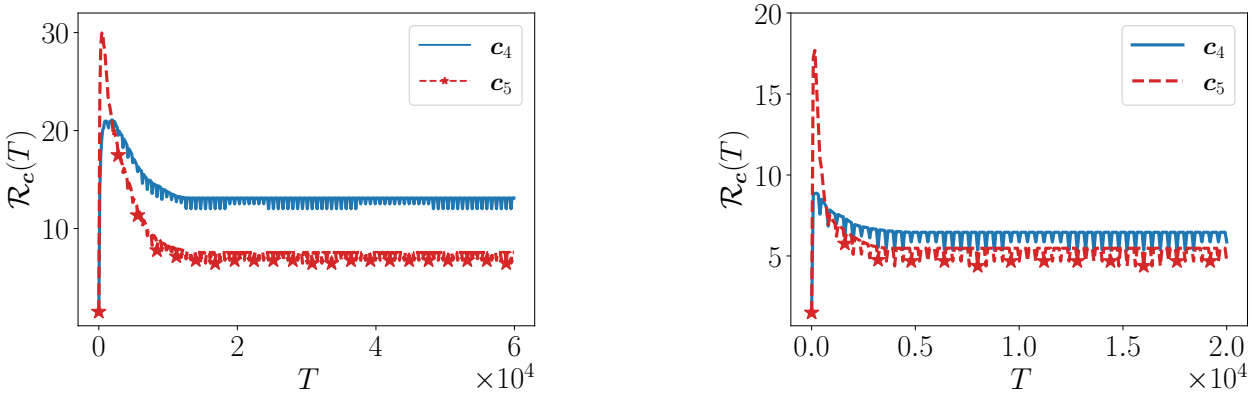
- LFG integrates virtual queue techniques (Neely, 2010) and UCB algorithm to address the MMAB with the fairness constraints. LFG has $O(\sqrt{KMLT\ln T})$ regret bound. For the plots of LFG, we use the value of $\eta = 50$ for $\boldsymbol{\theta}_1$.
- UCB-LP is a UCB-based randomized algorithm. It consists of two stages at each time step t. At the first stage, it finds an optimal sampling distribution over arms, $y_i(t)$ for arm i with the fairness constraint $y_i(t) \geq c_i$. At the second stage, it constructs a distribution $\pi_{\boldsymbol{b}}(t)$ over the set of super-arms, $\mathcal{B} = \{\boldsymbol{b} \mid \boldsymbol{b} \in \{0,1\}^M,\ ||\boldsymbol{b}||_1 \leq K\}$, such that $\sum_{\boldsymbol{b}\in\mathcal{B}} \pi_{\boldsymbol{b}}(t)\boldsymbol{b} = y_i(t)$ with $\sum_{\boldsymbol{b}\in\mathcal{B}} \pi_{\boldsymbol{b}}(t) = 1$ for all $i \in [M]$. UCB-LP selects $\boldsymbol{b}$ according to the distribution $\pi_{\boldsymbol{b}}(t)$. UCB-LP has $O(1)$ regret upper bound.
- UCB-PLLP is a low-complexity version of UCB-LP and has $O(m\sqrt{T\ln T})$ regret upper bound. UCB-PLLP also uses virtual queues as LFG does. For the plots of UCB-PLLP, we use $\alpha_t = \frac{0.9}{\sqrt{t}}$ and $\epsilon_t = \frac{0.2}{\sqrt{t}}$ for $\boldsymbol{\theta} = \boldsymbol{\theta}_1$.

G.2. Experiments for non-constant

We run Fair-MMAB(K) algorithms with two different $\boldsymbol{c}$ settings and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$;

- $\boldsymbol{c}_4 = [0.03, 0.05, 0.07, 0.09, 0.11, 0.13, 0.15, 0.17]$,
- $\boldsymbol{c}_5 = [0.17, 0.15, 0.13, 0.11, 0.09, 0.07, 0.05, 0.03]$.

In the choice of $\boldsymbol{c}_4$, the fairness requirement c_j of bad arms (arms 4-8)is higher than that of good arms (arms 1-3), while in the choice of $\boldsymbol{c}_5$, the fairness requirement c_j of bad arms (arms 4-8)is lower than that of good arms (arms 1-3). Since $\boldsymbol{c}_4$ impose high exposure on bad arms, the regrets of $\boldsymbol{c}_4$ is higher than that of $\boldsymbol{c}_5$ for large T, which is observed in Figure 7.

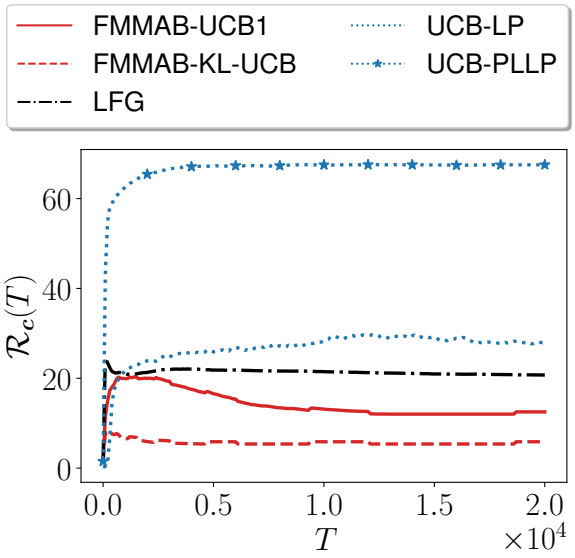


(a) Fair-MMAB(K)-UCB1 *(b)* Fair-MMAB(K)-KL-UCB

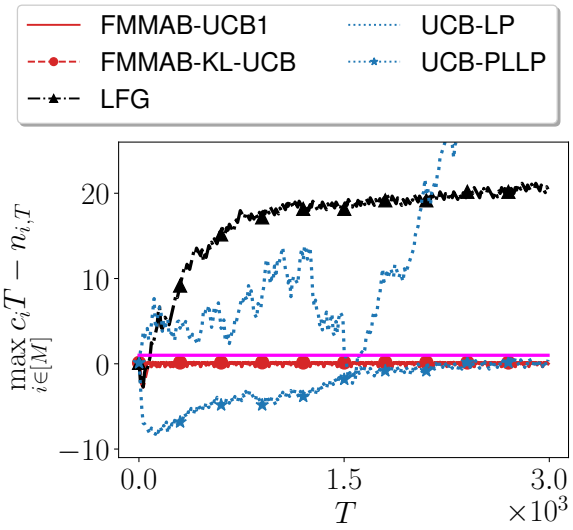
Figure 7. $\mathcal{R}_{\boldsymbol{c}}(T)$ for non-constant fairness constraints when $\boldsymbol{\theta} = \boldsymbol{\theta}_1$

Figure 8 show $\mathcal{R}_{\boldsymbol{c}}$ and $\max_{i\in[M]} c_i t - n_{i,t}$ for Fair-MMAB(K)-UCB1/KL-UCB, LFG, UCB-LP, and UCB-PLLP for $\boldsymbol{c} = \boldsymbol{c}_4$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. Figure 9 is for $\boldsymbol{c} = \boldsymbol{c}_5$. We observe similar behaviors of the graphs as for the case of constant fairness constraints in the figures. Our Fair-MMAB(K) algorithms outperform the existing fair multiple-play MABs, LFG, UCB-LP, and UCB-PLLP.

s

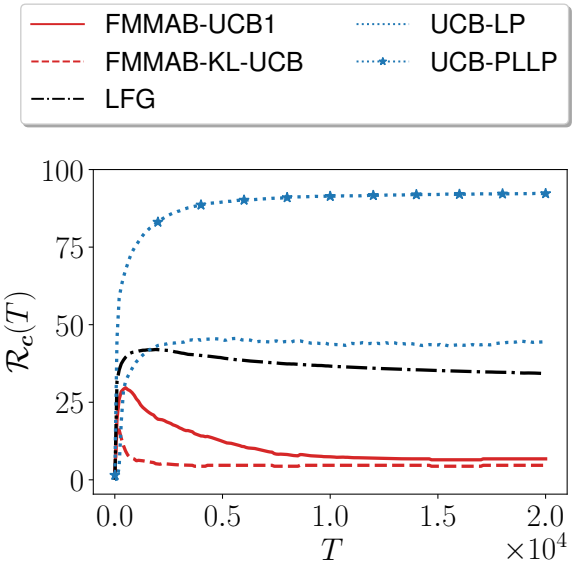


(a) $\mathcal{R}_c(T)$

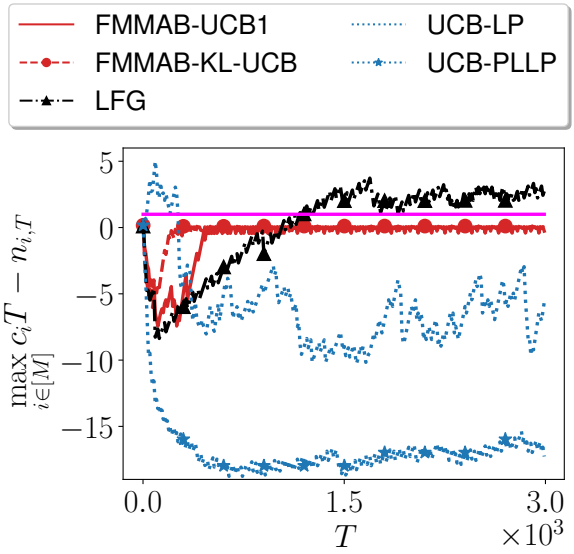


(b) Uniform Fairness

Figure 8. Performance Comparison for $\boldsymbol{c} = \boldsymbol{c}_4,\ \boldsymbol{\theta} = \boldsymbol{\theta}_1$



(a) $\mathcal{R}_c(T)$



(b) Uniform Fairness

Figure 9. Performance Comparison for $\boldsymbol{c} = \boldsymbol{c}_5,\ \boldsymbol{\theta} = \boldsymbol{\theta}_1$