IEEE ComSoc IEEE Communications Society | IEEE COMPUTER SOCIETY | IEEE Signal Processing Society | VTS | IEEE Transactions on Machine Learning in Communications and Networking

Received 29 September, 2023; revised 9 March, 2024 and 17 May, 2024; accepted 20 June, 2024; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

*Digital Object Identifier 10.1109/TMLCN.2023.1234567*

# Fair Probabilistic Multi-Armed Bandit with Applications to Network Optimization

**Zhiwu Guo[1], Student Member, IEEE, Chicheng Zhang[2], Ming Li[1], Fellow, IEEE, and Marwan Krunz[1], Fellow, IEEE**

[1]Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ 87521 USA
[2]Department of Computer Science, University of Arizona, Tucson, AZ 85721 USA

Corresponding author: Zhiwu Guo (email: zhiwuguo@arizona.edu).

An abridged version of this paper appeared in the Proceedings of IEEE WiOpt 2023 conference, August 2023. This research was supported in part by NSF (grants # 2229386 and 1822071) and by the Broadband Wireless Access & Applications Center (BWAC). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of NSF.

**ABSTRACT** Online learning, particularly Multi-Armed Bandit (MAB) algorithms, has been extensively adopted in various real-world networking applications. In certain applications, such as fair heterogeneous networks coexistence, multiple links (individual arms) are selected in each round, and the throughputs (rewards) of these arms depend on the chosen set of links. Additionally, ensuring fairness among individual arms is a critical objective. However, existing MAB algorithms are unsuitable for these applications due to different models and assumptions. In this paper, we introduce a new fair probabilistic MAB (FP-MAB) problem aimed at either maximizing the minimum reward for all arms or maximizing the total reward while imposing a fairness constraint that guarantees a minimum selection fraction for each arm. In FP-MAB, the learning agent probabilistically selects a meta-arm, which is associated with one or multiple individual arms in each decision round. To address the FP-MAB problem, we propose two algorithms: Fair Probabilistic Explore-Then-Commit (FP-ETC) and Fair Probabilistic Optimism In the Face of Uncertainty (FP-OFU). We also introduce a novel concept of regret in the context of the max-min fairness objective. We analyze the performance of FP-ETC and FP-OFU in terms of the upper bound of average regret and average constraint violation. Simulation results demonstrate that FP-ETC and FP-OFU achieve lower regrets (or higher objective values) under the same fairness requirements compared to existing MAB algorithms.

**INDEX TERMS** Probabilistic multi-armed bandit, Max-min fairness, Fairness constraint, Explore-then-commit, Optimism in the face of uncertainty, Online learning.

## I. INTRODUCTION

Online learning, especially MAB algorithms, is widely applied in various real-world networking applications, including cognitive radio networks [1], shortest path routing [2], and internet advertising [3]. In some applications, such as shortest path routing [2], a combination of multiple individual arms is played in each round, and exploring one set of arms can benefit the exploitation of other sets, as the reward of an individual arm is independent of the selected set. However, this independence does not hold in other applications, such as fair heterogeneous networks coexistence [4], wireless scheduling [5], and energy harvesting [6], which will be discussed in more detail later. Additionally, ensuring fairness among arms is a crucial objective in these applications.

In this paper, we introduce a probabilistic multi-armed bandit (MAB) problem in which we are given a collection of individual arms, and a learning agent *probabilistically* plays a meta-arm associated with one or multiple individual arms in each decision round. The reward of an individual arm *depends* on the specific meta-arm played and cannot be accurately estimated by pulling other meta-arms due to this dependence. The agent's objective is to either maximize its cumulative reward while adhering to a fairness constraint for each arm or to optimize a specific fairness objective, such as max-min fairness. This novel problem setup is applicable

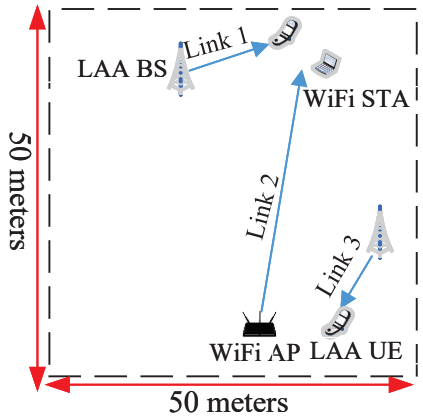



FIGURE 1: Motivating example: fair heterogeneous networks coexistence.

to numerous real-world networking scenarios. We describe three example applications as follows.

**Fair heterogeneous networks coexistence [4].** Different wireless technologies coexist in the same unlicensed band, such as LTE Licensed Assisted Access (LTE-LAA) [7] and WiFi [8] coexistence in the 5 GHz unlicensed spectrum. To improve spectrum utilization, previous works [4], [9] have proposed enabling simultaneous transmissions and adopting interference cancellation techniques to decode multiple signals, rather than using collision avoidance. Fig. 1 illustrates a heterogeneous LAA/WiFi coexistence scenario, with successive interference cancellation (SIC) [4] adopted in receivers. In this context, a transmission strategy is defined as a policy for selecting a concurrent transmission set (CTS) in each transmission period. Table 1 shows the normalized throughput of individual links under different transmission strategies. In this example, an arm is a link, a meta-arm is a CTS, and the reward of an arm corresponds to the link throughput. We observe that the reward of each arm depends on which meta-arm is selected by the learning agent since each arm's reward is influenced by the locations of other arms (i.e., interfering links). For instance, from Table 1, when arm 1 and arm 2 are concurrently played, the reward of arm 1 is 0.734, whereas it is 0.668 if arm 1 and arm 3 are concurrently played. Furthermore, ensuring fair and harmonious LAA/WiFi coexistence in the unlicensed spectrum is crucial for designing their transmission protocols [10], [11]. Thus, fairness objectives and constraints among individual arms (i.e., links) must be considered when designing online learning algorithms. To balance the reward of each arm, various meta-arms (transmission strategies) should be selected with different probabilities in each transmission period.

**Wireless scheduling [5].** Fair allocation of resources (e.g., bandwidth, power, transmission opportunities) is crucial for ensuring the Quality of Service (QoS) of users in wireless adhoc networks. Total user satisfaction is often improved if all users obtain an equitable quality of service, rather

TABLE 1: Normalized throughput of different transmission strategies.

| Transmission strategy | Link 1 | Link 2 | Link 3 |
|---|---|---|---|
| Link 1 transmits alone | 1 | 0 | 0 |
| Link 2 transmits alone | 0 | 1 | 0 |
| Link 3 transmits alone | 0 | 0 | 1 |
| Links 1 and 2 concurrently transmit | 0.734 | 0.277 | 0 |
| Links 1 and 3 concurrently transmit | 0.668 | 0 | 0.618 |
| Links 2 and 3 concurrently transmit | 0 | 0.799 | 0.685 |
| All three links concurrently transmit | 0.481 | 0.210 | 0.079 |

than some users benefiting at the expense of others. Fig. 2 illustrates an example of wireless network with four flows, where each flow can be either a single-hop link or a set of multi-hop links. Flows that share common nodes with other flows are considered as contending flows, meaning they cannot transmit packets simultaneously. For instance, flows 3 and 4 are contending flows as they share a common node $N_4$. In Fig. 2, there are three sets of non-contending flows: (flow 4, flow 1), (flow 4, flow 2), and flow 3. In this example, an arm represents a flow, a meta-arm represents a set of non-contending flows, and reward corresponds to throughput. The reward of each arm depends on the selected meta-arm because of interference from other flows. For instance, the reward of flow 4 differs between the meta-arms (flow 4, flow 1) and (flow 4, flow 2) due to interference from flow 1 and flow 2, respectively. The goal of the wireless scheduler is to either maximize the overall reward while adhering to a fairness constraint (e.g., minimum selection fraction) for each flow, or to optimize a specific fairness objective (e.g., max-min throughput). To achieve this, different sets of non-contending flows should be probabilistically selected in each transmission period.

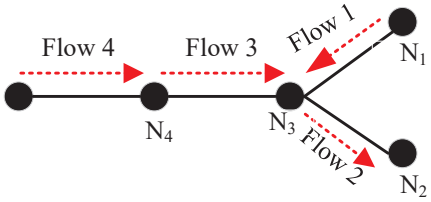



FIGURE 2: Motivating example: wireless scheduling.

**Energy harvesting [6].** Recent advances in wireless energy harvesting enable sensor nodes to extend their lifespan by remotely charging their batteries. The longevity of the sensor network depends on the minimum amount of energy a node can harvest in the network. Fig. 3 shows an example of energy harvesting where an energy source wirelessly charges 5 energy harvesting nodes using 3 channels. The amount of energy harvested by any node is stochastic and depends on the distance from the energy source and the channel allocated by the energy source. In each time slot, the energy source decides which channel to use for wireless charging,

aiming either to maximize the total harvested energy with a fairness constraint (e.g., minimum selection fraction) for each node or to maximize a specific fairness objective (e.g., max-min harvested energy). In this context, an arm represents an energy harvesting node, a meta-arm represents a mapping of node-channel associations, and the harvested energy corresponds to the reward. Different channels (meta-arms) present diverse propagation conditions for wireless waveform. For instance, if channel 1 has a higher frequency than channel 2, the reward of each arm on channel 1 will generally be less than that on channel 2 due to higher propagation loss [12]. Consequently, the reward of each arm depends on the meta-arm selected by the energy source. To meet the fairness objective or constraint, different meta-arms (node-to-channel assignments) may be chosen with varying probabilities.

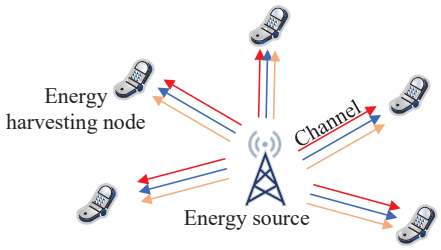



FIGURE 3: Motivating example: energy harvesting.

It is important to note that, although the aforementioned applications have a combinatorial arm structure, they cannot be solved using classic combinatorial multi-armed bandit (CMAB) formulations [2], [13], [14]. In CMAB, the learning agent plays a super arm consisting of multiple individual arms in each round, and the reward of a super arm can be expressed as a linear function of the individual arms' rewards, which remain constant regardless of the selected super arm. However, in our applications, the reward of an (individual) arm *depends* on the meta-arm that the learning agent selects due to environmental differences caused by meta-arms (e.g., locations of nodes, RF frequencies of channels). Thus, information obtained about an arm from one meta-arm is not useful for another meta-arm containing the same arm. Additionally, the reward of a meta-arm is no longer a linear function of the rewards of the individual arms within that meta-arm. Furthermore, while traditional MAB formulations with fairness considerations have been studied [15], [16], they are inapplicable to our problem as they lack a similar combinatorial structure and their solutions are not probabilistic.

Therefore, in this paper, we introduce a new online learning problem called the fair probabilistic multi-armed bandit (FP-MAB). Due to its unique combinatorial structure with dependent arm rewards and fairness considerations, this problem presents a novel type of MAB setup that has been largely unexplored in the literature. Our main contributions are summarized as follows.

(1) Inspired by a category of network optimization applications, we formulate a new fair probabilistic MAB (FP-MAB) problem. This problem aims to either maximize the minimum reward for all arms or maximize the total reward while adhering to fairness constraints that enforce a minimum selection fraction for each arm. We also introduce a novel notion of regret considering the max-min fairness objective.

(2) To address the formulated FP-MAB problem, we propose a Fair Probabilistic Explore-Then-Commit (FP-ETC) algorithm. This algorithm is designed to handle problems with both the max-min fairness objective and total reward maximization subject to fairness constraints. For the max-min fairness version of FP-ETC, we derive a sublinear upper bound $O\big(T^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\big)$ for the average regret, where $T$ is the time horizon and $K$ is the number of meta-arms. For the fairness constraint version of FP-ETC, we obtain an upper bound $O\big(NT^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\big)$ for the average regret and an upper bound $O\big(T^{-\frac{1}{3}}(K\log T)^{\frac{1}{3}}\big)$ for the average constraint violation of each arm, where $N$ is the number of arms.

(3) To further reduce the regret of FP-ETC, we introduce a Fair Probabilistic Optimism In the Face of Uncertainty (FP-OFU) algorithm. This algorithm is also applicable to problems with both the max-min fairness objective and total reward maximization subject to fairness constraints. For the max-min fairness version of FP-OFU, we derive an upper bound $O(K\sqrt{T}\log T)$ for the average regret. For the fairness constraint version of FP-OFU, we obtain an upper bound $O(NK\sqrt{T}\log T)$ for the average regret.

(4) We conduct extensive simulations to evaluate the performance of the proposed algorithms in the context of heterogeneous network coexistence application and compare them with state-of-the-art baselines. Simulation results demonstrate that the proposed algorithms significantly outperform the baseline algorithms in terms of average regret and fairness.

## II. RELATED WORK

### A. Combinatorial Multi-Armed Bandit

Combinatorial multi-armed bandit (CMAB) was first studied in [17], where the learning agent can play a combination of arms in each round. Since then, the CMAB model has found widespread application in various real-world networking scenarios, including shortest path routing [2], [18], channel allocation in cognitive radio networks [2], and recommendation systems [19]. Existing CMAB literature typically considers two types of reward structures: linear reward [2], [14] and non-linear reward [13], [20], [21]. This classification depends on whether a super-arm's reward can be expressed as a linear function of the rewards of individual arms belonging to it. However, it is essential to note that the reward structure of CMAB differs from our setting. In CMAB, the expected reward of each arm remains independent of the selected super-arm, resulting in the same expected reward regardless of which super-arm is played.

Consequently, in CMAB, information about the same arm can be exploited across different super-arms to expedite learning. In our FP-MAB framework, this is not the case. Although fairness considerations were studied in [16] under the CMAB setting, it is inapplicable to solving our problem since it does not have a similar combinatorial structure and the proposed solution is not probabilistic.

### B. Multi-Armed Bandit with Fairness Constraints

Online MAB algorithms with fairness constraints have been studied in the existing literature to address real-world networking problems. For instance, the work [22] introduced contextual MAB with fairness constraints, defining fairness as a minimum rate at which tasks or resources are allocated to users. In [23], the Fair-LinUCB algorithm was proposed to enhance the traditional LinUCB algorithm, aiming to achieve group-level fairness in personalized recommendation systems. While these studies focus on developing fair MAB algorithms in contextual bandit settings, they have limited relevance to our work. In [24], the authors studied a multi-player multi-armed bandit game where players collaborate to optimize arm allocation, maximizing the minimum expected reward received by any player. Similarly, the work [25] introduced team fairness, a group-based fairness measure in cooperative and single-objective multi-agent learning problems. However, our focus lies in single-agent applications, such as considering the entire network as a single agent in contexts like energy harvesting or wireless scheduling.

The authors in [26] introduced a definition of individual fairness, asserting that similar individuals should be treated similarly. Inspired by this concept, several fair MAB algorithms have been proposed. For instance, the FAIRBANDIT algorithm, outlined in [27], employs a strategy of playing all arms with equal probability until they can be distinguished with a high degree of confidence. Additionally, works such as [15], [16] have proposed fair MAB algorithms that ensure each arm is pulled at least a pre-specified fraction of the time. Furthermore, [28] presented online MAB algorithms that consider proportional fairness, aiming to maximize the sum of logarithmic utility functions of all arms. However, the aforementioned works do not explicitly address fairness objectives. In contrast, other works such as [29]–[31] have designed fair MAB algorithms utilizing the Nash Social Welfare (NSW) as the fairness objective. Unlike the objective of maximizing the minimum average reward in max-min fairness, NSW aims to maximize the product of all arms' average rewards.

The work most related to ours is [6]. This work proposed Maxmin-UCB, integrating the max-min objective into the UCB algorithm. Initially, Maxmin-UCB identifies the minimum UCB value (denoted as $\text{UCB}_{min}$) among all individual arms within a meta-arm, and subsequently selects the meta-arm with the maximum $\text{UCB}_{min}$. However, Maxmin-UCB may converge to deterministically playing a meta-arm over time, potentially leading to a max-min objective value that is no better (and possibly worse) than that achieved by our proposed fair probabilistic MAB algorithms in this paper.

In the conference version of this paper [32], we introduced the fair probabilistic explore-then-commit (FP-ETC) algorithm, which solely considers the max-min fairness objective. In this paper, we extend FP-ETC to address not only the max-min fairness objective but also fairness constraints. Additionally, to further reduce the regret of FP-ETC, we introduce the Fair Probabilistic Optimism In the Face of Uncertainty (FP-OFU) algorithm.

## III. PROBLEM FORMULATION

We begin with an overview of the multi-armed bandit (MAB) problem. This problem has received significant attention in the fields of statistics and machine learning [33], serving as a cornerstone for sequential decision-making in uncertain environments. In its fundamental setup, there is a set of arms (i.e., actions), available to the learning agent. Each arm presents an unknown reward distribution to the agent. With each passing time step, the agent selects an arm from this set and subsequently receives a stochastic reward drawn from the corresponding distribution. Ultimately, the objective of the learning agent is to maximize its expected cumulative reward.

Next, we describe the setup of FP-MAB. The frequently used notations are outlined in Table 2 for convenience.

TABLE 2: Main notations.

| Notation | Representation |
|---|---|
| $t$ | decision round |
| $N$ | number of arms |
| $n$ | index of arm |
| $A$ | meta-arm |
| $K$ | number of meta-arms |
| $\mathbf{p}_t$ | meta-arm selection vector at decision round $t$ |
| $r(A,n,t)$ | reward of arm $n$ associated with meta-arm $A$ at $t$ |
| $g(A,n)$ | true mean of reward for arm $n$ associated with meta-arm $A$ |

Illustrated in Fig. 4, we consider a discrete-time system with $N$ individual arms. We denote $\mathcal{N} = \{1, 2, \cdots, N\}$ as the set of all (individual) arms. In decision round $t (1 \leq t \leq T)$, the learning agent plays a *meta-arm* $A$, associated with one or multiple arms, where $A \in \mathcal{F}$ and $\mathcal{F}$ is the feasible set of meta-arms. There are $K$ meta-arms in $\mathcal{F}$. The learning agent then receives reward $r(A,n,t)$ for arm $n \in \mathcal{N}$ with $(n,A) \in E$, where $E$ is the edge set $\{\forall n, \forall A, (n,A) | \text{arm } n \text{ is associated with } A\}$. $E$ is determined by the actual application. For convenience, we normalize $r(A,n,t) \in [0,1]$. $r(A,n,t)$ is randomly sampled from an unknown distribution $\mathcal{D}_{A,n}$, dependent on the individual arm $n$ and meta-arm $A$. We assume that $\{r(A,n,t) : 1 \leq t \leq T\}$ is independent and identically distributed. The online decision-making process of selecting meta-arms is detailed in Alg. 1.

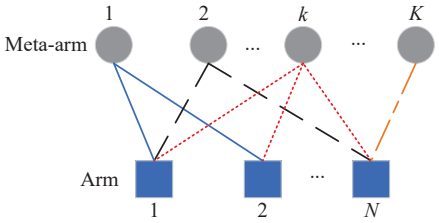



FIGURE 4: Problem setup of FP-MAB.

**Algorithm 1** Online decision-making process of selecting meta-arms

1: **Parameters**: $K$ meta-arms and $T$ decision rounds (both are known); reward distribution $\mathcal{D}_{A,n}$ (unknown), $A \in \mathcal{F}$; edge set $E$
2: **for** $t = 1$ to $T$ **do**
3: The learning agent plays a meta-arm $A$
4: It observes reward $r(A, n, t) \sim \mathcal{D}_{A,n}$ for all $n$, where $(n, A) \in E$
5: **end for**

**Remark 1:** The number of meta-arms, denoted as $K$, does not necessarily grow exponentially with the number of arms, $N$. In fact, $K$ only grows exponentially with $N$ in the worst-case scenario. To illustrate this point, we consider the three networking applications introduced in the Introduction section. Table 3 showcases the number of arms and meta-arms in the worst-case scenario for each application. For instance, in the context of energy harvesting, the number of arms remains fixed, equating to the number of energy harvesting nodes. Similarly, the number of meta-arms remains constant, corresponding to the number of channels. Notably, the number of meta-arms in this application is independent of the number of arms.

TABLE 3: The number of arms and meta-arms for three networking applications, in the worst scenario.

| Networking application | The number of arms | The number of meta-arms |
|---|---|---|
| Heterogeneous networks coexistence | $N$ | $2^N - 1$ |
| Wireless scheduling | $N$ | $2^{\frac{N}{2}}$ |
| Energy harvesting | Equal to the number of nodes | Equal to the number of channels |

In practical applications, the number of meta-arms may not be large. For example, in scenarios such as fair heterogeneous networks coexistence [4], where there are a limited number of users sharing the same unlicensed band with SIC, the number of meta-arms is typically constrained. For networking scenarios where the number of meta-arms increases exponentially with the number of arms, we can divide all arms into multiple orthogonal domains and allocate different resources to these orthogonal domains. This can decrease the number of arms sharing the same resource. Additionally, to mitigate the impact of potentially unfavorable meta-arms, we can eliminate a set of poor-performing meta-arms by probing all meta-arms before executing the FP-MAB algorithms. Poor meta-arms are those that yield low rewards for all arms associated with them.

To facilitate the discussion on optimizing fairness objectives and constraints, we define a policy as a probabilistic selection rule for meta-arms, rather than a deterministic choice. This is formally defined as follows:

**Definition 1** (Meta-arm Selection Vector)**.** *Meta-arm selection vector is denoted as* $\mathbf{p} = (p_1, \ldots, p_K)$*, where* $p_i (1 \leq i \leq K)$ *represents the probability of selecting meta-arm* $A_i$ *in each decision round.*

### A. Max-min Fairness Objective

We begin by formulating our problem with consideration for fairness objectives. Two commonly studied fairness objectives in wireless communications and networks are max-min fairness [34] and proportional fairness [35]. Integrating these fairness objectives into online MAB problems poses a non-trivial challenge, as traditional MAB algorithms are typically designed without fairness considerations. This necessitates the redesign of MAB algorithms, the redefinition of performance metrics (e.g., regret), and the derivation of new analyses to achieve fairness. In this work, we concentrate on determining the optimal $\mathbf{p}$ to maximize the minimum expected reward of arms, providing valuable insights for the design of other fair MAB algorithms.

Given a meta-arm $A \in \mathcal{F}$, we denote $g(A, n)$ as the true mean of the reward for arm $n$, where $n \in \mathcal{N}$. If $(n, A)$ is not in $E$, let $g(A, n) = 0$. $\forall A \in \mathcal{F}, \forall n \in \mathcal{N}$, if $g(A, n)$ is known, one can obtain the optimal $\mathbf{p}$ by solving the following max-min optimization problem:

$$\textbf{Opt-min}: \max_{\mathbf{p}} \quad f(\mathbf{p}) \\ \text{s.t.} \quad 0 \leq p_i \leq 1, i \in [K], \\ \sum_{i, i \in [K]} p_i = 1, \tag{1}$$

where $f(\mathbf{p}) = \min_{n \in \mathcal{N}} \{\sum_{i \in [K]} \big(p_i \times g(A_i, n)\big)\}$, $p_i$ represents the $i$-th element of $\mathbf{p}$, $A_i$ is the $i$-th meta-arm.

For comparison, we also present **Opt-total**, which seeks to *maximize the total expected rewards of all arms*:

$$\textbf{Opt-total}: \max_{\mathbf{p}} \quad \sum_{n \in \mathcal{N}} \sum_{i \in [K]} \big(p_i \times g(A_i, n)\big) \\ \text{s.t.} \quad 0 \leq p_i \leq 1, i \in [K], \\ \sum_{i, i \in [K]} p_i = 1. \tag{2}$$

However, $\forall i \in [K], n \in \mathcal{N}, g(A_i, n)$ is unknown to the learning agent beforehand. Hence, the learning agent must explore all meta-arms to learn and obtain accurate estimations of $g(A_i, n), n \in \mathcal{N}, i \in [K]$. Let $\hat{g}(A_i, n, t)$ represent the empirical average reward of arm $n \in \mathcal{N}$ until decision round $t$. We outline the process of obtaining $\hat{g}(A_i, n, t)$ as follows. If meta-arm $A_i$ is played in decision round $t$, the reward of arm $n, (n, A_i) \in E$, denoted as $r(A_i, n, t)$ is observed. For any other arm $n'$, where $(n', A_i) \notin E$,

namely, arm $n'$ is not played in $A_i$, we set $r(A_i, n', t) = 0$ for generalization purpose. Denote $n_t(A_i)$ as the number of times that $A_i$ has been played until decision round $t$, which can be represented as $n_t(A_i) = \sum_{\tau=1}^{t} \mathbf{1}\{a_\tau = A_i\}$, where $a_\tau$ is the index of meta-arm played in decision round $\tau$. Then, $\hat{g}(A_i, n, t)$ can be represented as follows:

$$\hat{g}(A_i, n, t) = \frac{1}{n_t(A_i)} \sum_{\tau=1}^{t} \mathbf{1}\{a_\tau = A_i\} r(A_i, n, \tau). \quad (3)$$

As previously mentioned, the learning agent explores all meta-arms to learn a more accurate $\hat{g}(A_i, n, t)$, while also exploiting the currently-known information to make the best action. Both exploration and exploitation incur a loss compared to the best action. We refer to this as loss *regret*, which measures how much suboptimal the learning agent is compared with the optimal strategy when the environment $\{g(A, n) : (A, n) \in E\}$ is known in advance. The goal of the learning agent is to minimize the incurred regret. However, unlike classic maximization problems, **Opt-min** is a max-min optimization problem, which necessitates a redefinition of regret. Inspired by the definition of regret in the regular MAB problem (e.g., Equation (1.1) of [36]), the regret of **Opt-min** is defined as:

$$R_T = \min_{n \in \mathcal{N}} \sum_{t=1}^{T} r(b_t, n, t) - \min_{n \in \mathcal{N}} \sum_{t=1}^{T} r(a_t, n, t), \quad (4)$$

where $b_1, \ldots, b_T$ is an independent and identically distributed sequence of meta-arms drawn from $\mathbf{p}^*$, which is the optimal solution of **Opt-min** in Equation (1) assuming the environment $\{g(A, n) : (A, n) \in E\}$ is known in advance; $a_1, \ldots, a_T$ is the sequence of meta-arms chosen by the learning agent.

Accordingly, define the average regret to be

$$\mathbb{E}[R_T] = \mathbb{E}\left[\min_{n \in \mathcal{N}} \sum_{t=1}^{T} r(b_t, n, t)\right] - \mathbb{E}\left[\min_{n \in \mathcal{N}} \sum_{t=1}^{T} r(a_t, n, t)\right], \quad (5)$$

where the expectation is with respect to (1) the choices of $b_1, \ldots, b_T$; (2) the random rewards drawn from the environment; (3) the random choices of $a_1, \ldots, a_T$ selected by the learning agent.

### B. Reward Maximization Subject To Fairness Constraint

We next formulate our problem with reward maximization while subject to fairness constraint. We model the fairness constraint as the targeted minimum selection fraction for each arm. Denote $\mathbf{d} = (d_1, d_2, \cdots, d_N)$ as the constraint vector, where $d_n$ denotes the targeted minimum selection fraction for arm $n$. If $\{g(A, n) : (A, n) \in E\}$ is known, the optimal $\mathbf{p}$ can be obtained by solving the following optimization problem:

$$\begin{aligned} \textbf{Opt-cons}: \max_{\mathbf{p}} \quad & \sum_{n \in \mathcal{N}} \sum_{i \in [K]} \big(p_i \times g(A_i, n)\big) \\ \text{s.t.} \quad & 0 \le p_i \le 1, i \in [K], \\ & \sum_{i, i \in [K]} p_i = 1, \\ & \sum_{i \in [K]} \big(p_i \times \mathbf{1}\{(n, A_i) \in E\}\big) \ge d_n, \forall n \in \mathcal{N}. \end{aligned} \quad (6)$$

Inspired by the definition of regret in the regular MAB problem (e.g., Equation (1.1) of [36]), the regret of **Opt-cons** is defined as:

$$R_T = \sum_{n \in \mathcal{N}} \sum_{t=1}^{T} r(b_t, n, t) - \sum_{n \in \mathcal{N}} \sum_{t=1}^{T} r(a_t, n, t), \quad (7)$$

where $b_1, \ldots, b_T$ is an independent and identically distributed sequence of meta-arms drawn from $\mathbf{p}^*$, which is the optimal solution of **Opt-cons** assuming the environment $\{g(A, n) : (A, n) \in E\}$ is known in advance; $a_1, \ldots, a_T$ is the sequence of meta-arms chosen by the learning agent.

The average regret is defined to be

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{n \in \mathcal{N}} \sum_{t=1}^{T} r(b_t, n, t)\right] - \mathbb{E}\left[\sum_{n \in \mathcal{N}} \sum_{t=1}^{T} r(a_t, n, t)\right]. \quad (8)$$

**Remark 2:** The objective of **Opt-cons** is to maximize the summation of average rewards among all arms while satisfying the fairness constraint. On the other hand, the objective of **Opt-min** is to maximize the minimum average reward of arms. Therefore, their regrets are defined differently (see Equations (4) and (7)).

Any constraint vector $\mathbf{d} = \{d_n\}_{n=1}^{N} (0 \le d_n \le 1)$ will be feasible for **Opt-cons** if there exists one meta-arm which is associated with all arms. For instance, the meta-arm $k$ in Fig. 4 is associated with all arms. In this case, $\mathbf{p} = (0, 0, 1, \cdots, 0)$, with $p_k = 1$ and $p_j = 0, j \ne k$, is a feasible meta-arm selection vector for any constraint vector $\mathbf{d} = \{d_n\}_{n=1}^{N}, 0 \le d_n \le 1$, since the constraint for each arm is satisfied according to Equation (6).

We would like to highlight that the problem formulation of FP-MAB would be meaningless if we solely prioritize maximizing the accumulated reward without checking the constraint violations for **Opt-cons**. Given the incorporation of fairness constraints within **Opt-cons**, it becomes essential to evaluate both regret and constraint violations simultaneously at each time step. Therefore, for **Opt-cons**, besides the regret, constraint violation also needs to be measured as a performance metric. For $\forall n \in \mathcal{N}$, the constraint violation of arm $n$ is defined as:

$$v_{n,T} = \max\{d_n - \frac{\sum_{t=1}^{T} \sum_{i \in [K]} p_{t,i} \mathbf{1}\{(n, A_i) \in E\}}{T}, 0\}, \quad (9)$$

where $p_{t,i}$ is the selection probability of meta-arm $A_i$ in decision round $t$. The average constraint violation of arm $n$ is

$$\mathbb{E}[v_{n,T}] = \mathbb{E}[\max\{d_n - \frac{\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i}\mathbf{1}\{(n,A_i)\in E\}}{T}, 0\}], \quad (10)$$

where the expectation is with the respect to $\mathbf{p}_t$.

## IV. FAIR PROBABILISTIC EXPLORE-THEN-COMMIT ALGORITHM

The explore-then-commit (ETC) algorithm [33], which consists of an exploration phase followed by an exploitation phase, is one of the most widely used MAB algorithms in a variety of online decision applications. The basic idea of the standard ETC algorithm is that the learning agent explores all arms uniformly in a round-robin manner during predefined exploration rounds, regardless of previous observations. In the remaining rounds, the learning agent selects the arm deemed empirically best for exploitation.

To address the formulated problems outlined in Section III, we introduce the Fair Probabilistic Explore-Then-Commit (FP-ETC) algorithm and analyze its performance in this section. FP-ETC is an extension of standard ETC algorithm. A key difference between FP-ETC and the standard ETC algorithm is that FP-ETC probabilistically selects arms during exploitation rounds instead of choosing the empirically best arm, as done in the standard ETC algorithm. The introduction of probabilistic exploitation in FP-ETC aims to satisfy desired fairness objectives or constraints.

The procedure of FP-ETC algorithm is outlined in Alg. 2, where the input $m$ is a pre-defined fixed positive integer representing the number of rounds that each meta-arm is explored in the exploration phase and will be optimized in the regret analysis later. $K$ is the number of meta-arms. If $t \leq mK$, the algorithm is in exploration phase (i.e., seeking better options) as shown from Step 3 to Step 5. Specifically, the FP-ETC algorithm first plays each meta-arm $m$ times in a round-robin fashion to update the empirical average reward for each arm associated with the corresponding meta-arm. Once $t > mK$, the algorithm enters the exploitation phase (i.e., staying with the currently-known best option) starting from Step 7. If the max-min fairness objective is considered, the minimum empirical average reward of all arms is maximized to obtain the estimated $\hat{\mathbf{p}}$ in Step 8, where $\mathcal{F}_1$ is the feasible set of $\mathbf{p}$ in **Opt-min**. Note that $\mathbf{p}$ does not change as $t$ increases in the exploitation phase of FP-ETC. If reward maximization subject to fairness constraint is considered, the summation of empirical average reward of all arms is maximized to obtain the estimated $\hat{\mathbf{p}}$ in Step 10, where $\mathcal{F}_2$ is the feasible set of $\mathbf{p}$ in **Opt-cons**. After that, FP-ETC sticks to the currently-known best option (i.e., $\hat{\mathbf{p}}$) and samples out a meta-arm $a_t$ based on the categorical distribution of $\hat{\mathbf{p}}$ as shown in Step 12. The selected meta-arm is played in Step 13.

It is important to note that Alg. 2 is utilized to determine the selected meta-arm in each decision round, and it interacts with line 3 of Alg. 1.

**Algorithm 2** Fair Probabilistic Explore-Then-Commit (FP-ETC)

1: **Input** : Positive integers $m$, $K$
2: **for** $t = 1$ to $T$ **do**
3: **if** $t \leq mK$ **then**
4: $a_t = t \bmod K + 1$
5: Play meta-arm $a_t$ in decision round $t$
6: **else**
7: **if** Max-min fairness objective is considered **then**
8: Solve the optimization problem $\hat{\mathbf{p}} = \arg\max_{\mathbf{p}\in\mathcal{F}_1} \min_{n\in\mathcal{N}}\{\sum_{i\in[K]} p_i \hat{g}(A_i, n, mK)\}$
9: **else if** Reward maximization subject to fairness constraint is considered **then**
10: Solve the optimization problem $\hat{\mathbf{p}} = \arg\max_{\mathbf{p}\in\mathcal{F}_2} \sum_{n\in\mathcal{N}}\sum_{i\in[K]} p_i \hat{g}(A_i, n, mK)$
11: **end if**
12: Sample out a meta-arm $a_t$ based on the categorical distribution of $\hat{\mathbf{p}}$
13: Play meta-arm $a_t$ in decision round $t$
14: **end if**
15: **end for**

We emphasize the novelty of our theoretical contributions in analyzing FP-ETC. The introduction of the meta-arm selection vector $\mathbf{p}$ in FP-ETC is crucial for meeting the fairness requirements. It is important to note that $\mathbf{p}$ is a continuous random variable within the feasible sets $\mathcal{F}_1$ and $\mathcal{F}_2$. As such, the general logic of regret analysis in traditional MAB algorithms with finite arms is not directly applicable to FP-ETC.

### A. Performance Analysis of FP-ETC Under Max-min Fairness Objective

To analyze the regret of FP-ETC under max-min fairness objective, we first present a concentration bound (Lemma 1) using the Hoeffding's inequality, from the perspective of each meta-arm. To bridge the gap between $\mathbf{p}$ and meta-arms, we then obtain a corresponding concentration bound with respect to any $\mathbf{p}$ (Lemma 2).

#### 1) Concentration Bounds

**Lemma 1.** *$\forall$ meta-arm $i$, $\forall n \in \mathcal{N}$, define event $E_{i,n}$ : $\left|\hat{g}(A_i, n, mK) - g(A_i, n)\right| \leq \sqrt{\frac{2\log(T)}{m}}$, where $m$ is the number of rounds that each meta-arm is played in the exploration phase, $K$ is the number of meta-arms, $T$ is the total number of rounds. Then $Pr(E_{i,n}) \geq 1 - \frac{2}{T^4}$.*

Lemma 1 is a direct application of the Hoeffding's inequality (Theorem A.1 of [33] by setting $\alpha = 2, \beta = 1$), given $r(A_i, n, t) \in [0, 1]$. It shows that the estimated

$\hat{g}(A_i, n, mK)$ concentrates around its true mean $g(A_i, n)$ after the exploration phase.

Obtaining a concentration bound per meta-arm is insufficient to derive the regret bound of FP-ETC under max-min fairness objective. Based on Lemma 1, we obtain a concentration bound for any $\mathbf{p}$, which is presented as follows.

**Lemma 2** (Concentration Bound for Any $\mathbf{p}$ Under Max-min Fairness Objective)**.** *Define function* $f(\mathbf{p}) = \min_{n\in\mathcal{N}}\{\sum_{i\in[K]}\big(p_i \times g(A_i, n)\big)\}$, *where* $p_i$ *is the* $i$*-th element of* $\mathbf{p}$. *For FP-ETC, if the max-min fairness objective is considered,* $\forall\ \mathbf{p} \in \mathcal{F}_1, Pr\Big(\Big|\hat{f}(\mathbf{p}) - f(\mathbf{p})\Big| \leq \sqrt{\frac{2\log(T)}{m}}\Big) \geq 1 - \frac{2NK}{T^4}$, *where* $N$ *is the number of arms,* $\hat{f}(\mathbf{p}) = \min_{n\in\mathcal{N}}\{\sum_{i\in[K]}\big(p_i \times \hat{g}(A_i, n, mK)\big)\}$.

*Proof outline*: we first rewrite $f(\mathbf{p}) = \min_{n\in\mathcal{N}} h_n(\mathbf{p})$ and $\hat{f}(\mathbf{p}) = \min_{n\in\mathcal{N}} \hat{h}_n(\mathbf{p})$. Then $\Big|h_n(\mathbf{p}) - \hat{h}_n(\mathbf{p})\Big|$ is upper-bounded for any given arm $n \in \mathcal{N}$. After that, we prove that $\Big|\hat{f}(\mathbf{p}) - f(\mathbf{p})\Big|$ is upper-bounded by utilizing Lemma 3. The detailed proof of Lemma 2 and Lemma 3 are shown in Section A of Appendix.

#### 2) Upper Bound on Average Regret of FP-ETC Under Max-min Fairness Objective

After obtaining the concentration bound with regard to any $\mathbf{p} \in \mathcal{F}_1$. We are ready to upper bound the regret of FP-ETC under max-min fairness objective. We state the results in Theorem 1.

**Theorem 1.** *For FP-ETC, if the max-min fairness objective is considered, the average regret* $\mathbb{E}[R_T]$ *defined in Equation (5) is upper bounded by* $O\Big(T^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\Big)$.

*Proof outline*: We first utilize Lemma 2 to upper bound $R_T^f = \sum_{t=1}^{T}[f(\mathbf{p}^*) - f(\mathbf{p_t})]$, where $\mathbf{p_t}$ is the $\mathbf{p}$ vector in decision round $t$. $R_T^f$ represents the summation of instantaneous performance gap between the optimal policy and policy chosen by the algorithm. However, there is still a gap between $R_T^f$ and $R_T$ of Equation (4). To bridge the gap, we make use of Hoeffding's inequality, union bound, and Lemma 4. The detailed proof of Theorem 1 is shown in Section B of Appendix. Lemma 4 is presented in Section C of Appendix.

### *B. Performance Analysis of FP-ETC Under Reward Maximization Subject to Fairness Constraint*

For FP-ETC, the regret analysis under reward maximization subject to fairness constraint shares similar reasoning to the analysis of max-min fairness objective since there are only differences of objectives (max-min reward versus max total reward) and constraints ($\mathcal{F}_1$ versus $\mathcal{F}_2$) between **Opt-min** and **Opt-cons**. However, constraint violation may happen under the consideration of fairness constraint. We present the results in the following Theorem.

**Theorem 2.** *For FP-ETC, if reward maximization subject to fairness constraint is considered, the average regret* $\mathbb{E}[R_T]$ *defined in Equation (8) is upper bounded by* $O\Big(NT^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\Big)$. *Therefore, the average constraint violation* $\mathbb{E}[v_{n,t}]$ *defined in Equation (10) is upper bounded by* $O\Big(T^{-\frac{1}{3}}(K\log T)^{\frac{1}{3}}\Big)$ *for any arm* $n \in \mathcal{N}$.

*Proof outline*: Define $f_1(\mathbf{p}) = \sum_{n\in\mathcal{N}}\sum_{i\in[K]}\big(p_i \times g(A_i, n)\big)$ and $\hat{f}_1(\mathbf{p}) = \sum_{n\in\mathcal{N}}\sum_{i\in[K]}\big(p_i \times \hat{g}(A_i, n, mK)\big)$, we can obtain that $\Big|\hat{f}_1(\mathbf{p}) - f_1(\mathbf{p})\Big| \leq N\sqrt{\frac{2\log(T)}{m}}$ is upper-bounded with high probability, similar to Lemma 2. The difference is that the confidence interval under reward maximization subject to fairness constraint is $N$ times as that of Lemma 2 due to the summation of rewards for all arms, instead of minimization of those rewards. Following similar proof line of Theorem 1, the average regret of FP-ETC under reward maximization subject to fairness constraint is upper-bounded as $O\Big(NT^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\Big)$ by selecting $m = O\Big((\frac{T}{K})^{\frac{2}{3}}(\log T)^{\frac{1}{3}}\Big)$. For any arm $n \in \mathcal{N}$, constraint violation only happens during the exploration phase of FP-ETC under reward maximization subject to fairness constraint. This is because FP-ETC adopts a round-robin fashion to explore each meta-arm without considering fairness constraint in the exploration phase. In the exploitation phase, no constraint violation happens since the optimized $\hat{\mathbf{p}}$ naturally lies in the feasible set $\mathcal{F}_2$ as indicated by line 10 of Alg. 2. FP-ETC contributes at most $d_n$ of constraint violation in each round of exploration phase for each arm. Therefore, $\mathbb{E}[v_{n,T}] \leq \frac{d_n \times mK + 0 \times (T-mK)}{T} \leq O\Big(T^{-\frac{1}{3}}(K\log T)^{\frac{1}{3}}\Big)$ for any arm $n \in \mathcal{N}$.

## V. FAIR PROBABILISTIC OPTIMISM IN THE FACE OF UNCERTAINTY ALGORITHM

Many existing MAB algorithms are designed based on the underlying principle of *optimism in the face of uncertainty (OFU)* [37]. The OFU principle is that the learning agent is always optimistic about the uncertainty of the environment. Despite lacking complete knowledge about all actions, the learning agent forms an optimistic estimate of how rewarding each action might be and selects the action with the highest estimated reward. If the estimate turns out to be incorrect, the learning agent adjusts its strategy accordingly. However, if the learning agent's choice is successful, it can exploit that action and minimize regret. Thus, the OFU principle helps balance exploration and exploitation.

Utilizing the principle of OFU for the proposed probabilistic MAB problem, we introduce a Fair Probabilistic OFU (FP-OFU) algorithm. Denote $\mathbf{g} \in [0, 1]^{K\times N}$, with $g_{i,n}$ representing the true mean of reward for arm $n$ associated with meta-arm $i$. We define two functions:

$$U_1(\mathbf{p}, \mathbf{g}) = \min_{n\in\mathcal{N}}\{\mathbf{p} \times \mathbf{g} \times \mathbf{e}_n\}, \text{where } \mathbf{p} \in \mathcal{F}_1, \qquad (11)$$

$$U_2(\mathbf{p},\mathbf{g}) = \sum_{n\in\mathcal{N}} (\mathbf{p}\times\mathbf{g}), \text{where } \mathbf{p}\in\mathcal{F}_2, \tag{12}$$

where $\mathbf{e}_n \in \{0,1\}^{N\times 1}$, with $n$-th element being 1 and 0 for other elements. $U_1(\mathbf{p},\mathbf{g})$ is designed for the consideration of the max-min fairness objective (i.e., **Opt-min**), while $U_2(\mathbf{p},\mathbf{g})$ is designed for the consideration of the fairness constraint (i.e., **Opt-cons**).

The basic idea of FP-OFU is that the algorithm maintains a confidence set $M_t \subseteq [0,1]^{K\times N}$ for the parameter $\mathbf{g}$ in decision round $t$, such that $\mathbf{g}\in M_t$ with a high probability. $M_t$ can be calculated and updated based on the past actions $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_t$ and respective reward $r(A_{a_1},n,1), r(A_{a_2},n,2),\cdots,r(A_{a_t},n,t),\ n\in\mathcal{N}$. The algorithm chooses a pair $(\mathbf{p},\mathbf{g})$ that jointly maximizes $U_1(\mathbf{p},\mathbf{g})$ or $U_2(\mathbf{p},\mathbf{g})$, depending on the consideration of either aforementioned fairness objective or fairness constraint.

The procedure of FP-OFU is given in Alg. 3. In step 1, the algorithm initializes $M_0$ as a hypercube with each dimension lower-bounded by 0 and upper-bound by 1. This is because each element in $\mathbf{g}$ ranges from $[0,1]$. If $1\le t\le T$, FP-OFU obtains an optimal pair $(\mathbf{p}_t,\tilde{\mathbf{g}}_t)$ that jointly maximizes $U_1(\mathbf{p}\in\mathcal{F}_1,\mathbf{g}\in M_t)$ if max-min fairness objective is considered, as shown in Step 4, or jointly maximizes $U_2(\mathbf{p}\in\mathcal{F}_2,\mathbf{g}\in M_t)$ if reward maximization subject to fairness constraint is considered as shown in Step 6. Note that the optimal pair $(\mathbf{p}_t,\tilde{\mathbf{g}}_t)$ changes as $t$ increases. After that, it samples out a meta-arm $a_t$ based on the categorical distribution of $\mathbf{p}_t$ as shown in Step 8 and plays meta-arm $a_t$ in Step 9. $M_t$ is then updated in Step 10. The updating rule for $M_t$ will be provided in Theorem 3.

**Algorithm 3** Fair Probabilistic Optimism In the Face of Uncertainty (FP-OFU)

1: Initialize $M_0 = [0,1]^{K\times N}$
2: **for** $t=1$ to $T$ **do**
3: **if** Max-min fairness objective is considered **then**
4: Solve the optimization problem $(\mathbf{p}_t,\tilde{\mathbf{g}}_t) = \arg\max_{(\mathbf{p},\mathbf{g})\in\mathcal{F}_1\times M_t} U_1(\mathbf{p},\mathbf{g})$
5: **else if** Reward maximization subject to fairness constraint is considered **then**
6: Solve the optimization problem $(\mathbf{p}_t,\tilde{\mathbf{g}}_t) = \arg\max_{(\mathbf{p},\mathbf{g})\in\mathcal{F}_2\times M_t} U_2(\mathbf{p},\mathbf{g})$
7: **end if**
8: Sample out a meta-arm $a_t$ based on the categorical distribution of $\mathbf{p}_t$
9: Play meta-arm $a_t$ in decision round $t$
10: Update $M_t$.
11: **end for**

### A. Performance Analysis of FP-OFU Under Max-min Fairness Objective

In this subsection, we analyze the regret of FP-OFU algorithm under the consideration of max-min fairness objective. As the construction of confidence sets $M_t$ is the key to Alg. 3. We first present how to construct these confidence sets in the following.

At time step $t$, denote $\mathbf{P}_t$ and $\mathbf{r}_{n,t}$ as follows:

$$\mathbf{P}_t = \begin{bmatrix}\mathbf{p}_1\\ \mathbf{p}_2\\ \cdots\\ \mathbf{p}_t\end{bmatrix},\quad \mathbf{r}_{n,t} = \begin{bmatrix} r(A_{a_1},n,1)\\ r(A_{a_2},n,2)\\ \cdots\\ r(A_{a_t},n,t)\end{bmatrix}, \tag{13}$$

$\mathbf{P}_t$ is a matrix with the dimension of $t\times K$, $\mathbf{r}_{n,t}$ is a reward vector with the dimension of $t\times 1$. For any arm $n\in\mathcal{N}$, define $\hat{\mathbf{g}}_{t,n}$ as the $n$-th column of matrix $\hat{\mathbf{g}}_t$, which is the $l^2$-regularized least-squares estimation of matrix $\mathbf{g}$ in decision round $t$, with regularization parameter $\lambda>0$. $\hat{\mathbf{g}}_{t,n}$ can be expressed as

$$\hat{\mathbf{g}}_{t,n} = (\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{P}_t^\top\mathbf{r}_{n,t}. \tag{14}$$

For any arm $n\in\mathcal{N}$, denote $\mathbf{g}_n$ as the $n$-th column of matrix $\mathbf{g}$. The following Theorem states that $\mathbf{g}_n$ lies with a high probability in an ellipsoid with the center at $\hat{\mathbf{g}}_{t,n}$.

**Theorem 3.** *Let $\lambda>0$, for any arm $n\in\mathcal{N}$, for any $0<\delta<1$, with probability at least $1-\delta$, for all $t>0$, $\mathbf{g}_n\in[0,1]^{1\times K}$ lies in the set*

$$M_t = \{\mathbf{g}'_n\in[0,1]^{1\times K} : \|\hat{\mathbf{g}}_{t,n}-\mathbf{g}'_n\|_{\bar{V}_t} \le \sqrt{K\log(1+\frac{t}{\lambda\delta})} + (\lambda K)^{\frac{1}{2}}\}, \tag{15}$$

*where $\bar{V}_t = \mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I}$.*

*Proof outline*: We first represent the instantaneous reward $r(A_{a_t},n,t)$ as a noisy linear product of $\mathbf{p}_t$ and $\mathbf{g}_n^\top$. After that, we measure the norm of $\hat{\mathbf{g}}_{t,n}-\mathbf{g}_n$ ($\hat{\mathbf{g}}_{t,n}$ is shown in Equation (14)) weighted by matrix $\bar{V}_t$, where $\bar{V}_t = \mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I}$. The detailed proof of Theorem 3 is provided in Section D of Appendix.

Based on Theorem 3, we upper bound the average regret of FP-OFU under max-min fairness objective, which is shown in Theorem 4.

**Theorem 4.** *For FP-OFU algorithm, if the max-min fairness objective is considered, $\forall T>0$, the average regret $\mathbb{E}[R_T]$ defined in Equation (5) is upper bounded by $O(K\sqrt{T}\log T)$.*

*Proof outline*: We first upper bound $R_T^{U_1} = \sum_{t=1}^{T}(U_1(\mathbf{p}^*,\mathbf{g}) - U_1(\mathbf{p}_t,\mathbf{g}))$, which is the summation of instantaneous performance gap between the optimal policy and policy chosen by the algorithm. However, there is still a gap between $R_T^{U_1}$ and $R_T$ of Equation (4). To bridge the gap, we make use of Azuma's inequality, union bound, and Lemma 4. The detailed proof of Theorem 4 is shown in Section E of Appendix.

### B. Performance Analysis of FP-OFU Under Reward Maximization Subject to Fairness Constraint

In this subsection, we analyze the performance of FP-OFU under reward maximization subject to fairness constraint.

**Theorem 5.** *For FP-OFU algorithm, if reward maximization subject to fairness constraint is considered, $\forall T > 0$, the average regret $\mathbb{E}[R_T]$ defined in Equation (8) is upper bounded by $O(NK\sqrt{T}\log T)$. In addition, there are no constraint violations for this case.*

**Remark 3:** The regret analysis of Theorem 5 is similar to that of Theorem 4 and its detailed proof is shown in Section F of Appendix. We observe that the average regret of Theorem 5 is $O(N)$ times of the average regret of Theorem 4. This is because there is a minimization function over all meta-arms for the regret definition of **Opt-min** as shown in Equation (4), whereas the regret is summed among all meta-arms for the regret definition of **Opt-cons** as shown in Equation (7). Furthermore, if reward maximization subject to fairness constraint is considered in FP-OFU, as indicated by line 6 of Alg. 3, no constraint violation happens in each round since the optimized $\hat{\mathbf{p}}_t$ naturally lies in the feasible set $\mathcal{F}_2$.

## VI. MORE DISCUSSIONS ON FP-MAB PROBLEM AND ALGORITHMS

### A. Lower Bound on Average Regret

Regarding the lower bound of proposed FP-MAB problem, we are interested in lower bounds on regret that apply to all FP-MAB algorithms, rather than analyzing a specific FP-MAB algorithm. We would like to highlight the FP-MAB problem is an extension of the traditional MAB problem [36] by playing a meta-arm (i.e., an association of one or multiple individual arms) at each time step while considering fairness objectives/constraints for each individual arm. First, if we set $N = 1$ in the FP-MAB problem, FP-MAB reduces to a traditional MAB problem, where the meta-arm in FP-MAB becomes a regular arm, and the max-min reward objective simplifies to reward maximization due to $N = 1$. Second, under this condition, if we further set $d_n = 0, \forall n \in \mathcal{N}$ in FP-MAB, the fairness constraints in FP-MAB can be neglected, and FP-MAB reduces to a traditional MAB problem without fairness constraints. Hence, the lower bound for the traditional MAB problem still applies to the FP-MAB problem. Based on the Theorem 2.1 of [33], for any bandit algorithm, there exists a problem instance such that $\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT})$, where $T$ is the time horizon and $K$ is the number of arms. $\Omega(\sqrt{KT})$ is the lower bound for the traditional MAB problem. For FP-OFU algorithm proposed in this paper, we can achieve an upper bound $O(K\sqrt{T}\log T)$ on the average regret. Therefore, the achieved upper bound has an $O(\sqrt{K}\log T)$ gap with its lower bound.

### B. Differences in Proof Techniques Compared to Those Used in Combinatorial Bandit

If we treat FP-MAB as a general combinatorial bandit framework, where the player generates a distribution over the regular arms and plays a regular arm (meta-arm in this paper) in each time step, existing fair MAB algorithms still cannot solve the FP-MAB problem. This is because FP-MAB focuses on the fairness of individual arms, rather than the fairness of meta-arms. In general combinatorial bandit frameworks, as studied in [2], [38], the received reward in each time step is a linear combination of individual arms' rewards. However, this is not the case for FP-MAB due to the max-min fairness objective. This key difference means that traditional proof techniques used in general combinatorial bandit works cannot be applied to analyze FP-MAB. Instead, we need to derive concentration bounds for non-linear rewards based on the concentration bounds for the rewards of all individual arms.

### C. Solving Optimization Problems

There are mainly two types of optimizations involved in the FP-ETC and FP-OFU algorithms. One is the online version of **Opt-min** under FP-ETC and FP-OFU, shown in line 8 of Algorithm 2 and line 4 of Algorithm 3. Another one is the online version of **Opt-cons** under FP-ETC and FP-OFU, shown in line 10 of Algorithm 2 and line 6 of Algorithm 3. We show how these optimization problems are solved in the following.

Both **Opt-min** of FP-ETC and **Opt-min** of FP-OFU are solved using *fminimax* function in MATLAB. *fminimax* seeks a point that minimizes the maximum of a set of objective functions. All the objective functions and constraints can be linear or non-linear. The idea of *fminimax* is that it converts a min-max (or max-min) problem into a goal attainment problem [39], which can be solved using standard goal attainment method introduced in [39]. Note that **Opt-min** of FP-ETC is a convex optimization problem, whereas **Opt-min** of FP-OFU is a non-convex and non-linear optimization problem as its objective function involves the multiplication of two optimization variables, $\mathbf{p}$ and $\mathbf{g}$.

Both **Opt-cons** of FP-ETC and **Opt-cons** of FP-OFU are solved using *fmincon* function in MATLAB, which is a nonlinear programming solver. It addresses the minimization (or maximization) optimization problems with linear or non-linear objective functions and constraints. *fmincon* adopts standard interior-point optimization algorithm [40] to solve the optimization problem. Note that **Opt-cons** of FP-ETC is a convex optimization problem, whereas **Opt-cons** of FP-OFU is a non-convex and non-linear optimization problem since its objective function involves the multiplication of two optimization variables, $\mathbf{p}$ and $\mathbf{g}$, which are jointly solved using *fmincon* function in MATLAB.

We discuss the computation complexity and scalability of implementing FP-ETC and FP-OFU algorithms in the following.

### D. Computation Complexity and Scalability

*Computation complexity:* First, we analyze the computation complexity of implementing FP-ETC. The main computation cost for FP-ETC is solving the optimization problems in line 8 and line 10 of Algorithm 2. We observe that the objective functions of both optimization problems are linear functions

with the meta-arm selection vector $\mathbf{p}$ as the optimization variable. The constraints $\mathcal{F}_1$ and $\mathcal{F}_2$ are also linear functions. Therefore, the optimization problems of FP-ETC can be treated as linear programming problems or converted to them, which can be solved in polynomial time.

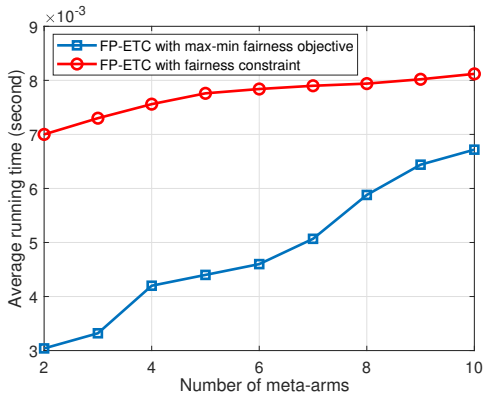
Next, we next analyze the computation complexity of implementing FP-OFU. The main computation cost for FP-OFU is solving the optimization problems in line 4 and line 6 of Algorithm 3. Even though the constraints $\mathcal{F}_1$ and $\mathcal{F}_2$ are linear functions, the objective functions of both optimization problems in FP-OFU are non-convex and non-linear since they involve the multiplication of two optimization variables, $\mathbf{p}$ and $\mathbf{g}$. For this type of optimization problems, only local optima may be obtained based on the related literature, and the complexity analysis for this non-convex problem is not well understood yet [41].

*Scalability:* We implemented the FP-ETC and FP-OFU algorithms in MATLAB within the context of the LAA/WiFi coexistence scenario. We simulated 500 random LAA/WiFi coexistence topologies with $T = 5000$ in both algorithms. Fig. 5 shows the average running time of FP-ETC and FP-OFU in each decision round. From Fig. 5(a), we observe that the time complexity of FP-ETC algorithm seems like scaling linearly with the number of meta-arms. Fig. 5(b) illustrates that the time complexity of the FP-OFU algorithm exhibits a superlinear scale with the number of meta-arms.

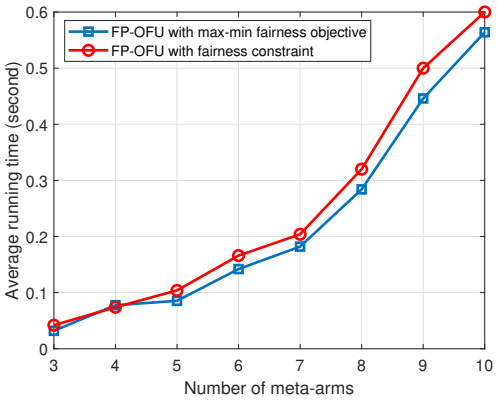
In practical applications, the number of meta-arms may not be large. For example, in scenarios such as fair heterogeneous networks coexistence [4], where there are a limited number of users sharing the same unlicensed band with SIC, the number of meta-arms is typically constrained. Additionally, to mitigate the impact of potentially unfavorable meta-arms, we can eliminate a set of poor-performing meta-arms by probing all meta-arms before executing the FP-MAB algorithms. Poor meta-arms are those that yield low rewards for all arms associated with them.

## VII. SIMULATION RESULTS

In this section, we evaluate the performance of the PF-ETC and FP-OFU algorithms via simulations under the application of heterogeneous LAA/WiFi coexistence, where SIC is enabled in each receiver to cancel possible interference and the successful decoding SINR threshold is set to 10 dB. Rayleigh channel is considered for each link. Due to space limitation, we only present the results for $N = 2$ and $N = 3$ coexisting links, as other scenarios have similar observations. There are 3 meta-arms and 7 meta-arms for the scenario of $N = 2$ and $N = 3$, respectively. To evaluate the performance of proposed algorithms under the consideration of the max-min fairness objective, we use the average regret defined in Equation (5), minimum link throughput, Jain fairness index (JFI) [42] as the performance metrics. Specifically, the throughput is normalized and represents the effective channel utilization. Let $x_n$ be the throughput of link $n \in \mathcal{N}$, $JFI(x_1, \cdots, x_N) = \frac{(\sum_{n=1}^{N} x_n)^2}{N \times \sum_{n=1}^{N} x_n^2}$, which ranges from $\frac{1}{N}$ (worst case) to 1 (best case), it is maximized when all links have the same throughput. To evaluate the performance of proposed algorithms under the consideration of the fairness constraint, we use the average regret defined in Equation (8), total throughput, the minimum selection fraction of arm as performance metrics. We simulate 500 randomized LAA/WiFi coexistence topology, where all nodes in each topology are uniformly distributed in a $50 \times 50$ $m^2$ area.




(a) FP-ETC




(b) FP-OFU

FIGURE 5: Average running time per decision round.

### A. FP-MAB with the Max-min Fairness Objective

First, we show the average regret of FP-ETC for different choices of $m$ and FP-OFU by setting $\lambda = 1, \delta = 0.01$ in Equation (15). We also compare them with the Maxmin UCB algorithm [6], where it selects meta-arm in each decision round according to the following rule: $a_t = \arg\max_{i \in [K]} [\min_{n \in \mathcal{N}} \hat{g}(A_i, n, t-1) + \sqrt{\frac{2 \log T}{n_{t-1}(A_i)}}]$. These results are presented in Fig. 6. We observe that Maxmin UCB performs worse than proposed FP-ETC and FP-OFU algorithms. This is because the Maxmin UCB algorithm aims to identify one best meta-arm and it will converge to deterministically playing a meta-arm after a sufficient time instead of a probabilistic meta-arm selection strategy. On the other hand, both FP-ETC and FP-OFU select a combination of different meta-arms to maximize the reward

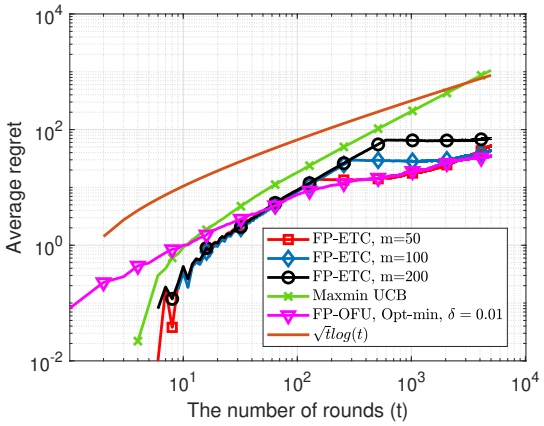
and satisfy the max-min fairness objective at the same time. The probabilistic MAB algorithm is actually a generalized version of the corresponding basic MAB algorithm. It is also observed that for FP-ETC, higher $m$ generally results in higher average regret as FP-ETC incurs significant regret during the exploration phase. For two-link LAA/WiFi coexistence scenario (i.e., $N = 2$), from Fig. 6(a), we can see that FP-OFU has similar performance with FP-ETC when $m$ is set to 50. This is because the number of meta-arms (i.e., $K$) is small in this case. However, as $K$ increases, FP-OFU outperforms FP-ETC, as indicated in Fig. 6(b). This is reasonable since FP-OFU avoids the round-robin exploration of suboptimal meta-arms, which happens during the exploration phase of FP-ETC.

It is possible to obtain the local optimal solutions for both online **Opt-min** and online **Opt-cons** in FP-OFU algorithm as they are non-convex and non-linear optimization problem. Their objective functions involve the multiplication of two optimization variables, $\mathbf{p}$ and $\mathbf{g}$. For the max-min fairness version of FP-OFU, the average regret is upper bounded by $O(K\sqrt{T}\log T)$. For the fairness constraint version of FP-OFU, the average regret is upper bounded by $O(NK\sqrt{T}\log T)$. To compare the average regret obtained through experiments implemented in MATLAB (which calls optimization functions to obtain meta-arm selection vector $\mathbf{p}_t$ at time $t$ in FP-OFU), we also show $\sqrt{t}\log(t)$ (theoretical trend of average regret of FP-OFU) in Fig. 6. We can see that the slope of average regret for FP-OFU is almost the same as $\sqrt{t}\log(t)$, indicating that the optimization problems of FP-OFU are well solved.

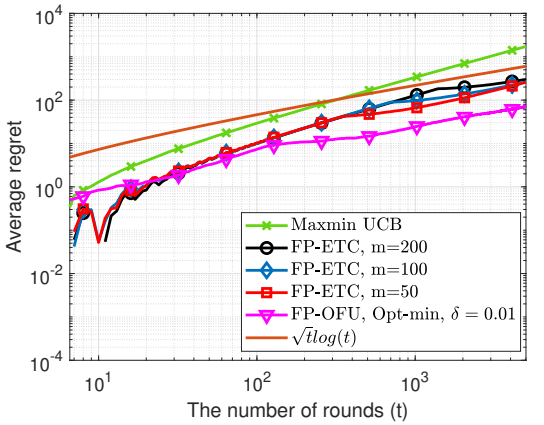
Next, we compare FP-ETC and FP-OFU with the following five baselines: UCB for **Opt-total**, ETC for **Opt-total** (adopt the same $m$ as FP-ETC), Maxmin UCB [6], ETC with NSW [29], UCB with NSW [29]. **Opt-total** is defined in Equation (2). The reasons why we select NSW as a comparison with the max-min fairness objective are as follows. NSW is a popular fairness objective in recently MAB works [29]–[31]. Different from maximizing the minimum average reward in max-min fairness objective, NSW aims to maximize the product of all arms' average rewards. This idea is the same as proportional fairness [35], another commonly studied fairness objective in wireless communications and networks. The algorithms of ETC with NSW and UCB with NSW in [29] are developed in a multi-agent MAB setting. To adapt it to our problem, ETC with NSW obtains the meta-arm selection vector $\hat{\mathbf{p}} = \arg\max_{\mathbf{p}\in\mathcal{F}_1} \prod_{n\in\mathcal{N}}\{\sum_{i\in[K]} p_i\hat{g}(A_i,n,mK)\}$; UCB with NSW updates $\mathbf{p}_t$ based on the following rule:

$$\mathbf{p}_t \leftarrow \arg\max_{\mathbf{p}\in\mathcal{F}_1}[\prod_{n\in\mathcal{N}}\sum_{i\in[K]} p_i\hat{g}(A_i,n,t-1) + N\sum_{i=1}^{K} p_i\sqrt{\frac{\log(NKt)}{n_{t-1}(A_i)}}]. \tag{16}$$

The UCB for **Opt-total** selects meta-arm in decision round $t$ according to the following rule: $a_t = \arg\max_{i\in[K]}\ [\sum_{n\in\mathcal{N}}\hat{g}(A_i,n,t-1) + N\sqrt{\frac{2\log t}{n_{t-1}(A_i)}}]$.




(a) Two-link LAA/WiFi coexistence ($N = 2$)




(b) Three-link LAA/WiFi coexistence ($N = 3$)

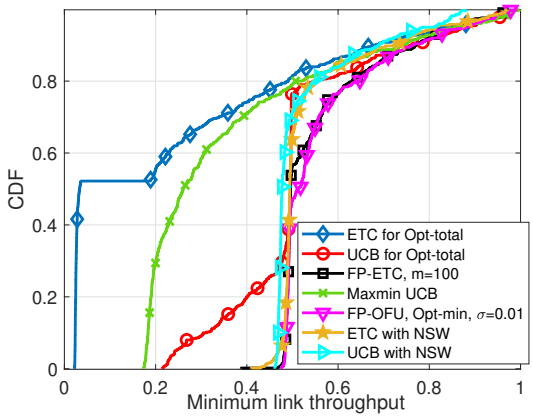
FIGURE 6: Average regret vs the number of rounds $t$.

The minimum link throughput and Jain fairness index (JFI) are compared for these algorithms.

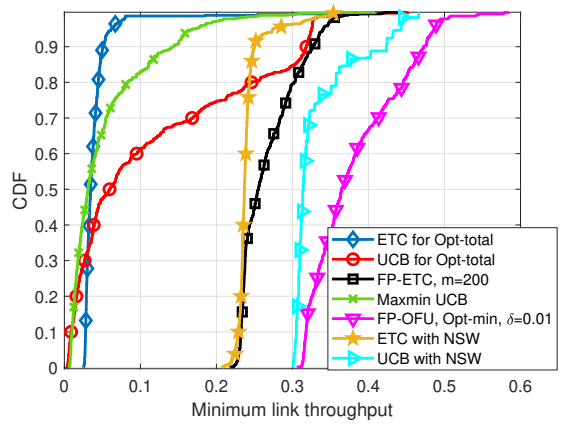
The cumulative distribution functions (CDFs) of the minimum link throughput for the aforementioned algorithms are presented in Fig. 7. The simulation is based on 500 randomized LAA/WiFi coexistence topologies with $T = 5000$. As we can see, both FP-ETC and FP-OFU achieve much higher minimum link throughput than other baselines. This is attributed to the additional meta-arm selection vector $\mathbf{p}$ in FP-ETC and FP-OFU algorithms, which allows for tuning the selection probability of each meta-arm to satisfy the max-min fairness objective.

From Fig. 7, we observe that the CDF of FP-ETC performs slight worse than that of FP-ETC in the case of $N = 2$. However, FP-OFU significantly outperforms the FP-ETC in the case of $N = 3$. We provide the intuition below. For FP-ETC, the explorations for meta-arms only happen during the exploration phase ($t \geq m * K$). When $N = 2$ (i.e., two LAA/WiFi link coexistence), there are only $K = 3$ meta-arms, $m = 100$ rounds of explorations for each meta-arm may be enough to obtain accurate successful decoding

probabilities for each link in FP-ETC algorithm. For $N = 3$, there are $K = 7$ meta-arms, the dependency of links' rewards in the same meta-arm is stronger with higher $N$. Therefore, even $m = 200$ rounds of explorations for each meta-arm may not be enough to obtain accurate successful decoding probabilities for each link in FP-ETC algorithm in the case of $N = 3$. However, for FP-OFU, the explorations for meta-arms happen during all decision rounds ($1 \leq t \leq T$), where $T$ is set to 5000.



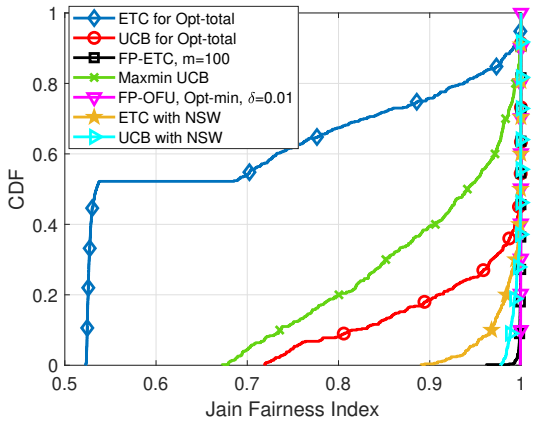

(a) Two-link LAA/WiFi coexistence ($N = 2$)



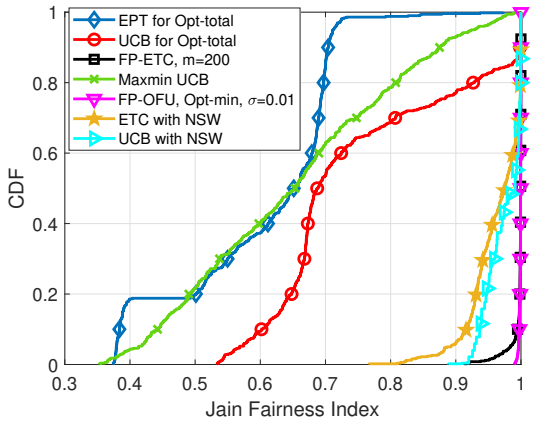

(b) Three-link LAA/WiFi coexistence ($N = 3$)

FIGURE 7: CDF of minimum link throughput.

Fig. 8 shows the CDFs of JFI for the aforementioned algorithms. It is observed that both FP-ETC and FP-OFU clearly perform better than other baselines. It is worth noting that FP-ETC and FP-OFU almost guarantee the same throughput for all links under all LAA/WiFi coexisting topologies, which is demonstrated by the fact that all JFI values of FP-ETC and FP-OFU are close to 1. Fig. 8 also indicates that FP-OFU slightly performs better than FP-ETC in some LAA/WiFi coexisting topologies. These results confirm that the proposed FP-ETC and FP-OFU algorithms are effective solutions to achieve the max-min fairness objective.




(a) Two-link LAA/WiFi coexistence ($N = 2$)




(b) Three-link LAA/WiFi coexistence ($N = 3$)

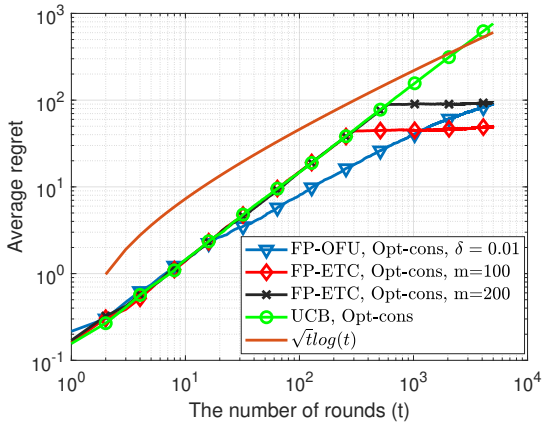
FIGURE 8: CDF of Jain fairness index (JFI).

### *B. FP-MAB with Fairness Constraint*

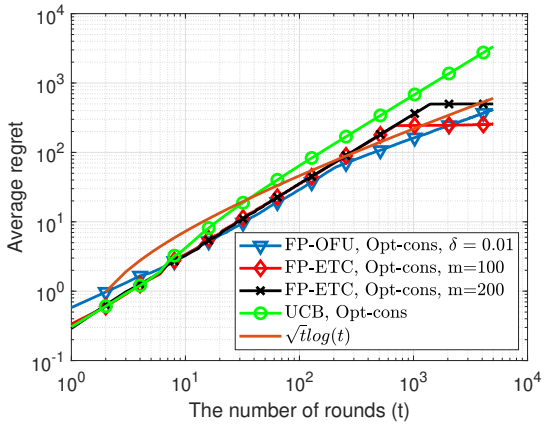
In this subsection, we present the simulation results of FP-ETC and FP-OFU algorithms under **Opt-cons**. We also compare them with the UCB for **Opt-cons**, which is similar to the proposed fair MAB algorithm in [15]. It selects meta-arm in each decision round according to the following rule: (1) determine $\Delta_d = \big(c_{t-1}(1) - d_1 \times (t-1), \cdots, c_{t-1}(n) - d_n \times (t-1), \cdots, c_{t-1}(N) - d_N \times (t-1)\big)$, where $c_{t-1}(n)$ is the number of times that link $n$ has been chosen until $t-1$, $d_n$ is the targeted minimum selection fraction for link $n$; (2) if all the elements of $\Delta_d$ are greater than 0, select meta-arm $a_t = \arg\max_{i \in [K]} [\sum_{n \in \mathcal{N}} \hat{g}(A_i, n, t-1) + N\sqrt{\frac{2 \log T}{n_{t-1}(A_i)}}]$, otherwise select meta-arm $a_t = \arg\max_{\{i | (\tilde{n}, A_i) \in E\}} [\sum_{n \in \mathcal{N}} \hat{g}(A_i, n, t-1) + N\sqrt{\frac{2 \log T}{n_{t-1}(A_i)}}]$, where $\tilde{n} = \min_{n \in \mathcal{N}} \Delta_d$. The simulation is based on 500 randomized LAA/WiFi coexistence topologies with $T = 5000$.

First, we show the average regret of proposed algorithms and baseline, which are presented in Fig. 9. For $N = 2$ and $N = 3$, we set the minimum selection fraction for each link as 0.6 and 0.4 respectively. We observe that FP-ETC and FP-

OFU significantly outperform UCB due to the probabilistic meta-arm selection strategy. It is expected higher $m$ incurs higher average regret for FP-ETC. It is interesting to observe that the regret of FP-OFU increases with a lower rate as $t$ when $t$ is small, compared with FP-ETC. This is again attributed to the significant regret of round-robin exploration of suboptimal meta-arms during the exploration phase of FP-ETC. When $t$ is large, FP-OFU has similar performance with FP-ETC when $m$ is set between 100 and 200.



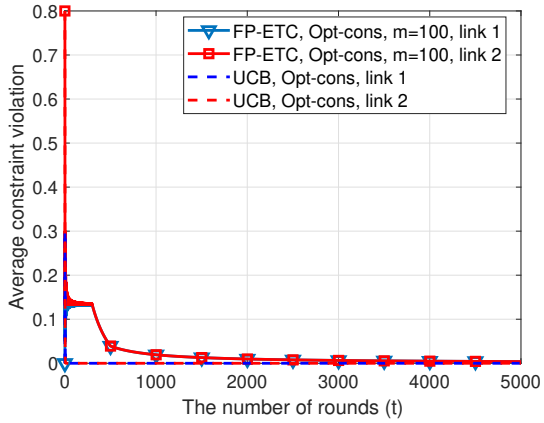

(a) Two-link LAA/WiFi coexistence ($N = 2$)



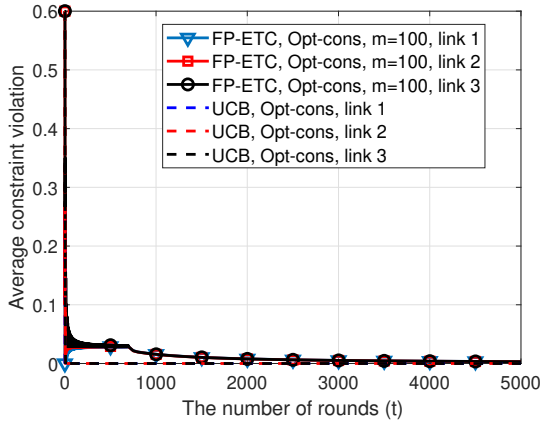

(b) Three-link LAA/WiFi coexistence ($N = 3$)

FIGURE 9: Average regret vs the number of rounds $t$.

Next, we present the average constraint violation of proposed algorithms and baseline, as shown in Fig. 10. For $N = 2$ and $N = 3$, we set the minimum selection fraction for each link as 0.8 and 0.6 respectively. Note that, as we stated in Theorem 5, there are no constraint violations in each round of FP-OFU, we only illustrate the results of FP-ETC and UCB in Fig. 10. We observe that, for FP-ETC, there are indeed constraint violations for each link (arm) in its exploration phase. However, the constraint violation will decrease with $t$ and converge to 0 when $t$ increases to infinity. For UCB, constraint violation can be quickly reduced to almost 0. This is because UCB can play the meta-arm, which




(a) Two-link LAA/WiFi coexistence ($N = 2$)



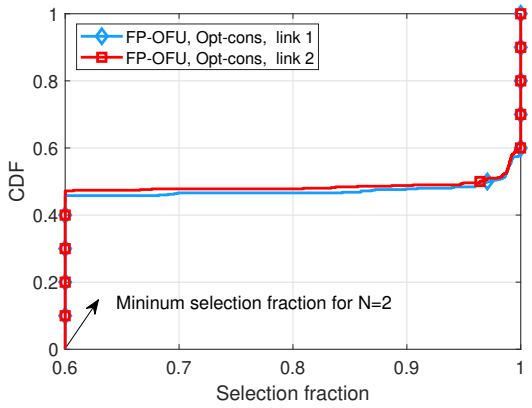

(b) Three-link LAA/WiFi coexistence ($N = 3$)

FIGURE 10: Constraint violation vs the number of rounds $t$.

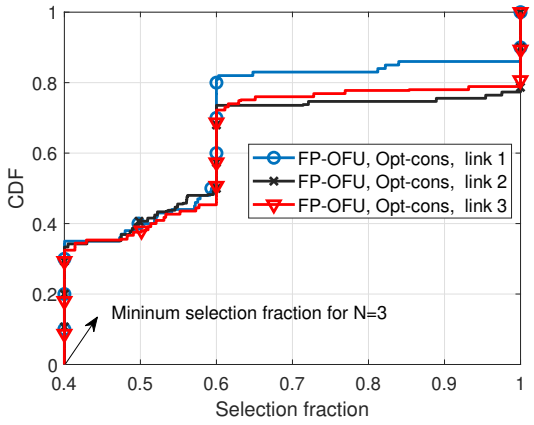
is associated with all individual arms, at the worst case to satisfy the fairness constraint. Even though UCB seems to have less constraint violation, its regret is much larger than FP-ETC (as shown in Fig. 9) and its objective values are also generally much smaller than FP-ETC, as shown later.

For FP-ETC, the constraint violation converge to 0 only when $t$ increases to infinity. To verify if the fairness constraint is satisfied for FP-OFU under a finite time horizon, we present the CDFs of selection fraction of different links (arms) for FP-OFU algorithm when $T$ is set to 5000. The simulation is based on 500 randomized LAA/WiFi coexistence topologies and we set the minimum selection fraction for each link to 0.6 and 0.4 for $N = 2$ and $N = 3$ respectively, we observe from Fig. 11 that the fairness constraints of all links (arms) for all LAA/WiFi coexistence topologies are indeed satisfied under a finite time horizon.

Lastly, we show the CDFs of the total throughput for the aforementioned algorithms in Fig. 12. We observe that both FP-ETC and FP-OFU achieve much higher total throughput than UCB. This is attributed to the additional meta-arm

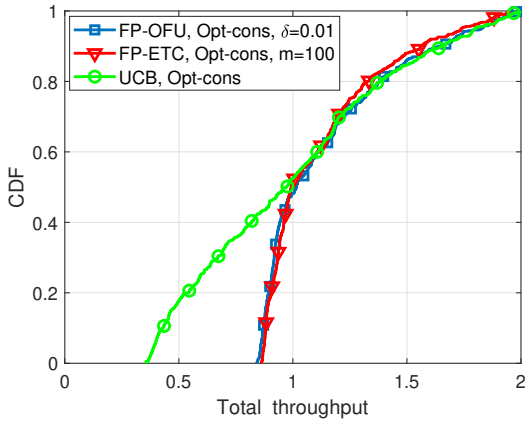



(a) Two-link LAA/WiFi coexistence ($N = 2$)

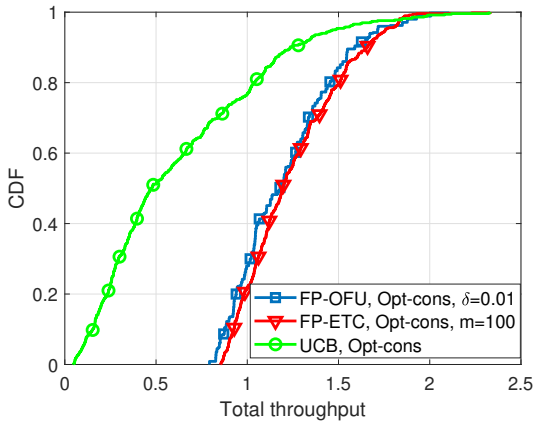



(b) Three-link LAA/WiFi coexistence ($N = 3$)

FIGURE 11: CDF of selection fraction for FP-OFU algorithm.




(a) Two-link LAA/WiFi coexistence ($N = 2$)




(b) Three-link LAA/WiFi coexistence ($N = 3$)

FIGURE 12: Total throughput.

selection vector $\mathbf{p}$ in FP-ETC and FP-OFU algorithms, which allows for tuning the selection probability of each meta-arm to satisfy desired fairness constraint. For both cases ($N = 2$ and $N = 3$), FP-ETC and FP-OFU achieve similar performance regarding the total throughput.

## VIII. Experimental Results

To validate the feasibility of the proposed algorithms in real-world LAA/WiFi coexistence scenarios, we established a wireless communication testbed. This testbed comprised three National Instruments (NI) USRP 2921 devices: one serving as the LAA transmitter, another as the WiFi transmitter, and the third functioning as the SIC receiver. We conducted over-the-air LAA/WiFi transmission and reception simultaneously, transmitting LAA and WiFi frames to a common SIC receiver at the 2.495 GHz unlicensed band. The SIC receiver was equipped with the capability to decode both LAA and WiFi frames. LAA and WiFi frames, utilizing different modulation and coding schemes, were generated using MATLAB's LTE toolbox and WLAN toolbox respectively. The LAA/WiFi coexistence topology is illustrated in Fig. 13, where "W", "L", and "R" denote the WiFi transmitter, LAA transmitter, and SIC receiver respectively.

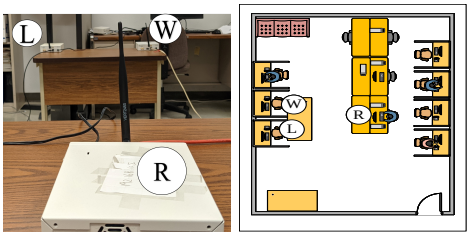



FIGURE 13: LAA/WiFi coexistence topology.

Let $\gamma = 10 \log \frac{P_L}{P_W}$, where $P_L$ and $P_W$ represent the transmission power of LAA and WiFi respectively. To explore various scenarios with different successful decoding probabilities for LAA and WiFi links, we manipulate $P_L$ and $P_W$ and consider $\gamma$ values in the set $\{-10, -5, 0, 5, 10\}$ in our experiments. We present the average regret under the experimental LAA/WiFi coexistence scenarios in Fig.

14, where the average regret is computed over all $\gamma$ values, with each $\gamma$ averaged over 10 experiments. Due to experimental constraints, we could only execute the three online algorithms (Maxmin UCB, FP-ETC, and FP-OFU) until $t = 161$. In this experiment, there are three meta-arms. For a fair comparison, we set $m = 25$ for FP-ETC, allowing it to explore all meta-arms from $t = 1$ to $t = 75$ before transitioning to exploitation. From Fig. 14, we observe that FP-OFU gradually outperforms FP-ETC as $t$ increases. This suggests that FP-OFU exhibits a lower slope in regret increase compared to FP-ETC over time. Notably, both FP-ETC and FP-OFU significantly outperform Maxmin UCB. These findings align with the simulation results, underscoring the efficacy and applicability of the proposed algorithms in real-world scenarios.

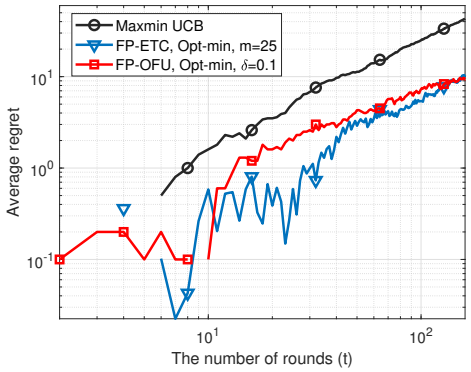



FIGURE 14: Average regret under experimental LAA/WiFi coexistence scenarios.

## IX. CONCLUSIONS

We clearly and concisely summarized our key findings and contributions in the following.

*Contributions:* Firstly, we formulate a new fair probabilistic MAB (FP-MAB) problem considering max-min fairness objective or fairness constraint (i.e., enforcing a minimum selection fraction for each arm); we also define a novel notion of regret considering the max-min fairness objective. Additionally, we propose fair probabilistic explore-then-commit (FP-ETC) and fair probabilistic optimism in the face of uncertainty (FP-OFU) algorithms, both applicable to solving the formulated FP-MAB problem. Performance metrics such as average regret and average constraint violation are analyzed for FP-ETC and FP-OFU algorithms. Lastly, we evaluate the performances of FP-ETC and FP-OFU algorithms in a practical networking scenario, such as heterogeneous LAA/WiFi networks coexistence in the 5 GHz unlicensed band, and compare the proposed algorithms with state-of-the-art baseline algorithms.

*Key findings:* Firstly, FP-ETC and FP-OFU achieve sub-linear regret and significantly outperform the state-of-the-art baselines (e.g., maxmin-UCB). Secondly, concerning the max-min fairness objective, FP-ETC and FP-OFU achieve much higher minimum link throughput in the scenario of LAA/WiFi coexistence, compared with the state-of-the-art baselines (e.g., maxmin-UCB [6], UCB for **Opt-total**, ETC for **Opt-total**, UCB with NSW [29]). Lastly, FP-ETC exhibits constraint violations during its exploration phase due to deterministic explorations, whereas FP-OFU has no constraint violations.

## X. APPENDIX

### A. Proof of Lemma 2

*Proof:*
First, we define event $E = \cap_{i\in[K],n\in\mathcal{N}} E_{i,n}$, where $E_{i,n}$ is defined in Lemma 1. Using the property of union bound, we have $Pr(E) \geq 1 - \sum_{\forall i,n} Pr(\bar{E}_{i,n}) \geq 1 - \sum_{i\in[K]} N\frac{2}{T^4} = 1 - \frac{2NK}{T^4}$. For the remainder of the proof, we condition on event $E$ happening.

For any $\mathbf{p} \in \mathcal{F}_1$, given $f(\mathbf{p}) = \min_{n\in\mathcal{N}}\{\sum_{i\in[K]}\big(p_i \times g(A_i,n)\big)\}$, the estimation of $f(\mathbf{p})$ at the end of exploration phase (i.e., $t = mK$) of FP-ETC is $\hat{f}(\mathbf{p}) = \min_{n\in\mathcal{N}}\{\sum_{i\in[K]}\big(p_i \times \hat{g}(A_i,n,mK)\big)\}$.

For the convenience of the proof, for every $n \in \mathcal{N}$, we define $h_n(\mathbf{p}) = \sum_{i\in[K]}\big(p_i \times g(A_i,n)\big)$ and $\hat{h}_n(\mathbf{p}) = \sum_{i\in[K]}\big(p_i \times \hat{g}(A_i,n,mK)\big)$, then we can obtain the relationship of $f(\mathbf{p})$ and $h_n(\mathbf{p})$, $\hat{f}(\mathbf{p})$ and $\hat{h}_n(\mathbf{p})$, respectively, which are $f(\mathbf{p}) = \min_{n\in\mathcal{N}} h_n(\mathbf{p})$ and $\hat{f}(\mathbf{p}) = \min_{n\in\mathcal{N}} \hat{h}_n(\mathbf{p})$. For every $n \in \mathcal{N}$, we bound $\left|h_n(\mathbf{p}) - \hat{h}_n(\mathbf{p})\right|$ :

$$\begin{aligned}
\left|h_n(\mathbf{p}) - \hat{h}_n(\mathbf{p})\right| &\leq \left|\sum_{i\in[K]}\big(p_i g(A_i,n)\big) - \sum_{i\in[K]}\big(p_i \hat{g}(A_i,n,mK)\big)\right| \\
&\leq \left|\sum_{i\in[K]} p_i\big(g(A_i,n) - \hat{g}(A_i,n,mK)\big)\right| \\
&\leq \left|\sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{m}}\right| \text{ (if event } E \text{ happens)} \\
&= \sqrt{\frac{2\log(T)}{m}} \text{ (since } \sum_{i\in[K]} p_i = 1).
\end{aligned} \tag{17}$$

Equation (17) shows that for every $n \in \mathcal{N}$, $\left|h_n(\mathbf{p}) - \hat{h}_n(\mathbf{p})\right|$ is upper-bounded by $\sqrt{\frac{2\log(T)}{m}}$ if event $E$ happens. Next, we utilize Equation (17) to further bound $\left|\hat{f}(\mathbf{p}) - f(\mathbf{p})\right|$ for any $\mathbf{p} \in \mathcal{F}_1$. To do so, we need another Lemma, which is presented as follows.

**Lemma 3.** *Denote a sequence* $A = \big(a(n)\big)_{n\in\mathcal{N}}$ *and a sequence* $B = \big(b(n)\big)_{n\in\mathcal{N}}$, *if* $\forall n, a(n) \leq b(n)$, *then* $\min_{n\in\mathcal{N}} a(n) \leq \min_{n\in\mathcal{N}} b(n)$.

The proof of Lemma 3 is straightforward. Let $\arg\min_{n\in\mathcal{N}} b(n) = n^*$. We have $\min_{n\in\mathcal{N}} b(n) = b(n^*) \geq a(n^*) \geq \min_{n\in\mathcal{N}} a(n)$. This completes the proof of Lemma 3.

By Equation (17), taking $a(n) = \hat{h}_n(\mathbf{p}) - \sqrt{\frac{2\log(T)}{m}}$ and $b(n) = h_n(\mathbf{p})$ give $a(n) \leq b(n)$, $\forall n \in \mathcal{N}$. According to Lemma 3,

$$\min_{n\in\mathcal{N}}\{\hat{h}_n(\mathbf{p}) - \sqrt{\frac{2\log(T)}{m}}\} \leq \min_{n\in\mathcal{N}} h_n(\mathbf{p}). \tag{18}$$

Similarly, taking $a(n) = h_n(\mathbf{p})$ and $b(n) = \hat{h}_n(\mathbf{p}) + \sqrt{\frac{2\log(T)}{m}}$ give $a(n) \leq b(n)$, $\forall n \in \mathcal{N}$. By Lemma 3,

$$\min_{n\in\mathcal{N}} h_n(\mathbf{p}) \leq \min_{n\in\mathcal{N}}\{\hat{h}_n(\mathbf{p}) + \sqrt{\frac{2\log(T)}{m}}\}. \tag{19}$$

Combining Equation (18) and Equation (19), we obtain

$$\min_{n\in\mathcal{N}} h_n(\mathbf{p}) - \sqrt{\frac{2\log(T)}{m}} \leq \min_{n\in\mathcal{N}} \hat{h}_n(\mathbf{p}) \leq \min_{n\in\mathcal{N}} h_n(\mathbf{p}) + \sqrt{\frac{2\log(T)}{m}}, \tag{20}$$

which is equivalently

$$\left|\hat{f}(\mathbf{p}) - f(\mathbf{p})\right| \leq \sqrt{\frac{2\log(T)}{m}}. \tag{21}$$

■

### B. Proof of Theorem 1

*Proof:*

Define a clean event $\xi := \{\forall \mathbf{p} \in \mathcal{F}_1, \left|\hat{f}(\mathbf{p}) - f(\mathbf{p})\right| \leq \sqrt{\frac{2\log(T)}{m}}\}$, where $\hat{f}(\mathbf{p})$ and $f(\mathbf{p})$ are defined in Lemma 2. According to Lemma 2, $Pr(\xi) \geq 1 - \frac{2NK}{T^4}$. We also define a bad event $\bar{\xi}$, which is the complement of $\xi$. Denote the optimal $\mathbf{p}^* = \arg\max_{\mathbf{p}\in\mathcal{F}_1} f(\mathbf{p})$.

We first analyze event $\xi$. If FP-ETC chooses $\mathbf{p}$, where $\mathbf{p} \neq \mathbf{p}^*$, under event $\xi$, we have $f(\mathbf{p}) + \sqrt{\frac{2\log(T)}{m}} > \hat{f}(\mathbf{p}) > \hat{f}(\mathbf{p}^*) \geq f(\mathbf{p}^*) - \sqrt{\frac{2\log(T)}{m}}$. Re-arranging these terms gives

$$f(\mathbf{p}^*) - f(\mathbf{p}) \leq 2\sqrt{\frac{2\log(T)}{m}}. \tag{22}$$

Therefore, under event $\xi$, FP-ETC contributes at most $2\sqrt{\frac{2\log(T)}{m}}$ regret in each round of the exploitation phase. In each round of the exploration phase, FP-ETC trivially contributes at most regret of 1 for each arm. Thus, under event $\xi$,

$$\begin{aligned} R_T^f = \sum_{t=1}^{T}[f(\mathbf{p}^*) - f(\mathbf{p_t})] &\leq mK + (T-mK)2\sqrt{\frac{2\log(T)}{m}} \\ &\leq mK + 2T\sqrt{\frac{2\log(T)}{m}}, \end{aligned} \tag{23}$$

where $\mathbf{p_t}$ is the $\mathbf{p}$ vector in decision round $t$. When $t \leq mK$, $\mathbf{p_t}$ is a standard basis vector, with element 1 indicating that the corresponding meta-arm is selected in the exploration phase. The total regret in the exploration phase of FP-ETC is upper-bounded by $mK$ as there are $K$ meta-arms and each meta-arm is played $m$ times.

Until now, we have upper bounded $R_T^f$ under event $\xi$ happens. However, there is still a gap between $R_T^f$ and $R_T$ of Equation (4). Observe that the first term in Equation (4) is $\min_{n\in\mathcal{N}} \sum_{t=1}^{T} r(b_t, n, t)$. Applying the Hoeffding's inequality, we obtain that $\forall n \in \mathcal{N}$, at least with probability $1-\delta$,

$$\left|\sum_{t=1}^{T} r(b_t, n, t) - T\sum_{i\in[K]} p_i^* g(A_i, n)\right| \leq \sqrt{\frac{T}{2}\log(\frac{2}{\delta})}. \tag{24}$$

Without loss of generality, we set $\delta = \frac{1}{T^2}$. Therefore,

$$\left|\min_{n\in\mathcal{N}}\sum_{t=1}^{T} r(b_t, n, t) - T\min_{n\in\mathcal{N}}\sum_{i\in[K]} p_i^* g(A_i, n)\right| \leq \sqrt{T\log(2T)}. \tag{25}$$

Note that $\mathbf{p_t}$ does not change when $t > mK$. Applying the Hoeffding's inequality, we obtain that $\forall n \in \mathcal{N}$, at least with probability $1 - \frac{1}{T^2}$,

$$\begin{aligned} &\left|\sum_{t=mK+1}^{T} r(a_t, n, t) - \sum_{t=mK+1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i, n)\right| \\ &\leq \sqrt{(T-mK)\log(2T)}, \end{aligned} \tag{26}$$

where $p_{t,i}$ is the $i$-th element of vector $\mathbf{p_t}$.

When $t \leq mK$, the reward for each arm is upper bounded by 1. Combining the two cases of $t \leq mK$ and $t > mK$ together, we obtain that $\forall n \in \mathcal{N}$, at least with probability $1 - \frac{1}{T^2}$,

$$\begin{aligned} &\left|\sum_{t=1}^{T} r(a_t, n, t) - \sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i, n)\right| \\ &\leq mK + \left|\sum_{t=mK+1}^{T} r(a_t, n, t) - \sum_{t=mK+1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i, n)\right| \\ &\leq mK + \sqrt{(T-mK)\log(2T)}, \end{aligned} \tag{27}$$

Therefore,

$$\begin{aligned} &\left|\min_{n\in\mathcal{N}}\sum_{t=1}^{T} r(a_t, n, t) - \min_{n\in\mathcal{N}}\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i, n)\right| \\ &\leq mK + \sqrt{(T-mK)\log(2T)}. \end{aligned} \tag{28}$$

Define event $\eta := \{\forall n \in \mathcal{N}$, Equation (24) and Equation (27) hold$\}$. Event $\bar{\eta}$ is the complement of $\eta$. Using union bound, we obtain that event $\eta$ happens at least with probability $1 - \frac{2N}{T^2}$.

Combining Equation (25) and Equation (28) together, when event $\xi \cap \eta$ happens, $R_T$ of Equation (4) can be upper

bounded:

$$\begin{aligned}
R_T &= \min_{n\in\mathcal{N}} \sum_{t=1}^{T} r(b_t, n, t) - \min_{n\in\mathcal{N}} \sum_{t=1}^{T} r(a_t, n, t) \\
&\leq T \min_{n\in\mathcal{N}} \sum_{i\in[K]} p_i^* g(A_i, n) - \min_{n\in\mathcal{N}} \sum_{t=1}^{T} \sum_{i\in[K]} p_{t,i} g(A_i, n) \\
&+ \sqrt{T\log(2T)} + mK + \sqrt{(T-mK)\log(2T)} \\
&\leq T f(\mathbf{p}^*) - f(\sum_{t=1}^{T} \mathbf{p_t}) + mK + 2\sqrt{T\log(2T)} \\
&\overset{(a)}{\leq} T f(\mathbf{p}^*) - \sum_{t=1}^{T} f(\mathbf{p_t}) + mK + 2\sqrt{T\log(2T)} \\
&= R_T^f + mK + 2\sqrt{T\log(2T)} \\
&\leq 2mK + 2T\sqrt{\frac{2\log(T)}{m}} + 2\sqrt{T\log(2T)},
\end{aligned} \tag{29}$$

where (a) is because of Lemma 4 of Section C of Appendix.

Recall that $m$ was given in advance in FP-ETC algorithm. Therefore, we can choose $m$ to minimize the right-hand side of Equation (29). Since the first two terms (i.e., $2mK$ and $2T\sqrt{\frac{2\log(T)}{m}}$) are monotonically increasing and monotonically decreasing with respect to $m$. We can set $m$ so that the two terms are approximately equal. By solving it, we obtain $m = O\big((\frac{T}{K})^{\frac{2}{3}}(\log T)^{\frac{1}{3}}\big)$. Plug it into Equation (29), we have $R_T \leq O\big(T^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\big)$, where $O\big(T^{\frac{1}{2}}(\log T)^{\frac{1}{2}}\big)$ is neglected as it has a lower order than $O\big(T^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\big)$.

Using union bound, $Pr(\xi \cap \eta) \geq 1 - \frac{2N}{T^2} - \frac{2NK}{T^4}$, averaging all the events, then $\mathbb{E}[R_T]$ of Equation (5) is

$$\begin{aligned}
\mathbb{E}\left[R_T\right] \leq& \mathbb{E}\left[R_T I(\xi\cap\eta)\right] + \mathbb{E}\left[R_T I(\overline{\xi\cap\eta})\right] \\
&\leq O\big(T^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\big) + \mathbb{E}\left[T\cdot I(\overline{\xi\cap\eta})\right] \\
&\leq O\big(T^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\big) + T\cdot(\frac{2N}{T^2} + \frac{2NK}{T^4}) \\
&\leq O\big(T^{\frac{2}{3}}(K\log T)^{\frac{1}{3}}\big),
\end{aligned} \tag{30}$$

where the last term $T\cdot(\frac{2N}{T^2} + \frac{2NK}{T^4})$ is neglected since it is the order of $T^{-1}$.

■

### C. Lemma 4 and Its Proof

**Lemma 4.** *Given* $f(\mathbf{p}) = \min_{n\in\mathcal{N}}\{\sum_{i\in[K]}\big(p_i \times g(A_i,n)\big)\}$, $f(\sum_{t=1}^{T}\mathbf{p_t}) \geq \sum_{t=1}^{T} f(\mathbf{p_t})$.

*Proof:*

Denote $n^* = \arg\min_{n\in\mathcal{N}} \sum_{i\in[K]}\big(\sum_{t=1}^{T} p_{t,i}\times g(A_i,n)\big)$, where $p_{t,i}$ is the $i$-th element of $\mathbf{p_t}$, we have

$$\begin{aligned}
f(\sum_{t=1}^{T}\mathbf{p_t}) &= \min_{n\in\mathcal{N}}\{\sum_{i\in[K]}\big(\sum_{t=1}^{T} p_{t,i}\times g(A_i,n)\big)\} \\
&= \sum_{i\in[K]}\big(\sum_{t=1}^{T} p_{t,i}\times g(A_i,n^*)\big) \\
&= \sum_{t=1}^{T}\sum_{i\in[K]}\big(p_{t,i}\times g(A_i,n^*)\big) \\
&\geq \sum_{t=1}^{T}\min_{n\in\mathcal{N}}\{\sum_{i\in[K]}\big(p_{t,i}\times g(A_i,n)\big)\} \\
&= \sum_{t=1}^{T} f(\mathbf{p_t}),
\end{aligned} \tag{31}$$

■

### D. Proof of Theorem 3

*Proof:*

Since each meta-arm is randomly sampled based on the categorical distribution of $\mathbf{p}_t$ in each decision round, we obtain that for any arm $n \in \mathcal{N}$, for all $t > 0$,

$$\mathbb{E}[r(A_{a_t}, n, t)|\mathbf{p}_t] = \langle \mathbf{p}_t, \mathbf{g}_n^\top \rangle, \tag{32}$$

where $a_t$ is the index of meta-arm chosen by the learning agent. The instantaneous reward $r(A_{a_t}, n, t)$ can be modelled as a noisy linear product of $\mathbf{p}_t$ and $\mathbf{g}_n^\top$, which is

$$r(A_{a_t}, n, t) = \langle \mathbf{p}_t, \mathbf{g}_n^\top \rangle + \eta_{n,t}, \tag{33}$$

where $\eta_{n,t}$ is a random variable representing the noise. Since both $r(A_{a_t}, n, t) \in [0,1]$ and $\langle \mathbf{p}_t, \mathbf{g}_n^\top \rangle \in [0,1]$, then $\eta_{n,t} \in [-1,1]$. Therefore, $\eta_{n,t}$ follows 1-subgaussian distribution [43].

Let $\eta_n = (\eta_{n,1}, \eta_{n,2}, \cdots, \eta_{n,t})^\top$. Plug Equation (33) into Equation (14), we obtain

$$\begin{aligned}
\hat{\mathbf{g}}_{t,n} &= (\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{P}_t^\top(\mathbf{P}_t\mathbf{g}_n + \eta_n) \\
&= (\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{P}_t^\top\mathbf{P}_t\mathbf{g}_n + (\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{P}_t^\top\eta_n \\
&\overset{(a)}{=} (\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}(\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})\mathbf{g}_n - \lambda(\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{g}_n \\
&+ (\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{P}_t^\top\eta_n \\
&= \mathbf{g}_n - \lambda(\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{g}_n + (\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{P}_t^\top\eta_n,
\end{aligned} \tag{34}$$

where (a) is because we add and subtract a term $\lambda(\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{g}_n$.

Subtract $\mathbf{g}_n$ and multiply $\mathbf{p}$ on both sides of Equation (34), we obtain

$$\begin{aligned}
\mathbf{p}(\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n) &= \mathbf{p}(\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{P}_t^\top\eta_n - \lambda\mathbf{p}(\mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I})^{-1}\mathbf{g}_n \\
&= \langle \mathbf{p}^\top, \mathbf{P}_t^\top\eta_n\rangle_{\bar{V}_t^{-1}} - \lambda\langle \mathbf{p}^\top, \mathbf{g}_n\rangle_{\bar{V}_t^{-1}},
\end{aligned} \tag{35}$$

where $\bar{V}_t = \mathbf{P}_t^\top\mathbf{P}_t + \lambda\mathbf{I}$.

Note that for $\lambda > 0$, both $\bar{V}_t$ and $\bar{V}_t^{-1}$ are positive definite, the above inner product is well-defined. Using Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\left|\mathbf{p}(\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n)\right| &\leq \|\mathbf{p}^\top\|_{\bar{V}_t^{-1}}\left(\|\mathbf{P}_t^\top \eta_n\|_{\bar{V}_t^{-1}} + \lambda\|\mathbf{g}_n\|_{\bar{V}_t^{-1}}\right) \\ &\overset{(a)}{\leq} \|\mathbf{p}^\top\|_{\bar{V}_t^{-1}}\left(\|\mathbf{P}_t^\top \eta_n\|_{\bar{V}_t^{-1}} + \lambda^{\frac{1}{2}}\|\mathbf{g}_n\|_2\right),\end{aligned} \tag{36}$$

where (a) is because $\|\mathbf{g}_n\|^2_{\bar{V}_t^{-1}} \leq \frac{1}{\lambda_{min}(\bar{V}_t)}\|\mathbf{g}_n\|_2^2 \leq \frac{1}{\lambda}\|\mathbf{g}_n\|_2^2$.

By the Theorem 1 of reference [44] and let $V = \lambda\mathbf{I}$ and $R = 1$ (since $\eta_{n,t}$ is 1-subgaussian), for any arm $n \in \mathcal{N}$, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t > 0$,

$$\|\mathbf{P}_t^\top \eta_n\|_{\bar{V}_t^{-1}} \leq \sqrt{2\log\left(\frac{\det(\bar{V}_t)^{\frac{1}{2}}\det(\lambda\mathbf{I})^{-\frac{1}{2}}}{\delta}\right)}. \tag{37}$$

Conditioning on the above event happens, plug Equation (37) into Equation (36), we have $\forall t > 0, \forall \mathbf{p}$,

$$\begin{aligned}&\left|\mathbf{p}(\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n)\right| \\ \leq&\|\mathbf{p}^\top\|_{\bar{V}_t^{-1}}\left(\sqrt{2\log\left(\frac{\det(\bar{V}_t)^{\frac{1}{2}}\det(\lambda\mathbf{I})^{-\frac{1}{2}}}{\delta}\right)} + \lambda^{\frac{1}{2}}\|\mathbf{g}_n\|_2\right).\end{aligned} \tag{38}$$

Since both $\mathbf{p}$ and $\mathbf{g}_n$ are $1 \times K$ vector and each element is upper bounded by 1, $\|\mathbf{p}\|_2 \leq \sqrt{K}$ and $\|\mathbf{g}_n\|_2 \leq \sqrt{K}$, based on the relationship of trace and determinant of a matrix, we get

$$\det(\bar{V}_t) \leq (\lambda + t)^K, \tag{39}$$

Therefore, Equation (38) can be reduced to

$$\left|\mathbf{p}(\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n)\right| \leq \|\mathbf{p}^\top\|_{\bar{V}_t^{-1}}\left(\sqrt{K\log(1 + \frac{t}{\lambda\delta})} + (\lambda K)^{\frac{1}{2}}\right). \tag{40}$$

Let $\mathbf{p}^\top = \bar{V}_t(\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n)$, Equation (40) is written as

$$\begin{aligned}\|\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n\|^2_{\bar{V}_t} \leq \|\bar{V}_t(\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n)\|_{\bar{V}_t^{-1}} \times \\ \left(\sqrt{K\log(1 + \frac{t}{\lambda\delta})} + (\lambda K)^{\frac{1}{2}}\right).\end{aligned} \tag{41}$$

Since $\|\bar{V}_t(\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n)\|_{\bar{V}_t^{-1}} = \|\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n\|_{\bar{V}_t}$, divide both sides by $\|\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n\|_{\bar{V}_t}$, we get

$$\|\hat{\mathbf{g}}_{t,n} - \mathbf{g}_n\|_{\bar{V}_t} \leq \sqrt{K\log(1 + \frac{t}{\lambda\delta})} + (\lambda K)^{\frac{1}{2}}. \tag{42}$$

■

### *E. Proof of Theorem 4*

*Proof:*

Denote $\beta_t(\delta) = \sqrt{K\log(1 + \frac{t}{\lambda\delta})} + (\lambda K)^{\frac{1}{2}}$. Define $E = \cap_{n\in\mathcal{N}} E_n$, where $E_n := \{\text{Equation (15) holds for arm } n\}$.

For $\mathbf{p} \in \mathcal{F}_1$, for any arm $n \in \mathcal{N}$, define $h_n(\mathbf{p}, \mathbf{g}) = \sum_{i\in[K]}\left(p_i \times g(A_i, n)\right)$, then $h_n(\mathbf{p}, \hat{\mathbf{g}}_{t,n}) = \sum_{i\in[K]}\left(p_i \times \hat{g}(A_i, n, t)\right)$. Therefore,

$$U_1(\mathbf{p}, \mathbf{g}) = \min_{n\in\mathcal{N}} h_n(\mathbf{p}, \mathbf{g}), \tag{43}$$

$$U_1(\mathbf{p}, \hat{\mathbf{g}}_t) = \min_{n\in\mathcal{N}} h_n(\mathbf{p}, \hat{\mathbf{g}}_{t,n}). \tag{44}$$

We first bound $\left|h_n(\mathbf{p}, \mathbf{g}) - h_n(\mathbf{p}, \hat{\mathbf{g}}_{t,n})\right|$ for any $\mathbf{p}$ when $E$ happens (which, due to Theorem 3, happens with probability $1 - N\delta$):

$$\begin{aligned}\left|h_n(\mathbf{p}, \mathbf{g}) - h_n(\mathbf{p}, \hat{\mathbf{g}}_{t,n})\right| &= \left|\sum_{i\in[K]}\left(p_i g(A_i, n)\right) - \sum_{i\in[K]}\left(p_i \hat{g}(A_i, n, t)\right)\right| \\ &= \left|\langle \mathbf{p}, \mathbf{g}_n^\top\rangle - \langle \mathbf{p}, \hat{\mathbf{g}}_{t,n}^\top\rangle\right| \\ &= \left|\langle \mathbf{p}, (\mathbf{g}_n - \hat{\mathbf{g}}_{t,n})^\top\rangle\right| \\ &= \left|\|\mathbf{p}\|_{\bar{V}_t^{-1}}\|\mathbf{g}_n - \hat{\mathbf{g}}_{t,n}\|_{\bar{V}_t}\right| \\ &\overset{(a)}{\leq} \beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}},\end{aligned} \tag{45}$$

where (a) is due to the definition of event $E$.

Using Lemma 3 and taking $a(n) = h_n(\mathbf{p}, \hat{\mathbf{g}}_{t,n}) - \beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}}, n \in \mathcal{N}$ and $b(n) = h_l(\mathbf{p}, \mathbf{g}), n \in \mathcal{N}$, we obtain that for every $n \in \mathcal{N}$, $a(n) \leq b(n)$. Therefore,

$$\min_{n\in\mathcal{N}}\{h_n(\mathbf{p}, \hat{\mathbf{g}}_{t,n}) - \beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}}\} \leq \min_{n\in\mathcal{N}} h_n(\mathbf{p}, \mathbf{g}). \tag{46}$$

Similarly, taking $a(n) = h_n(\mathbf{p}, \mathbf{g}), n \in \mathcal{N}$ and $b(n) = h_n(\mathbf{p}, \hat{\mathbf{g}}_{t,n}) + \beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}}, n \in \mathcal{N}$, we obtain that for every $n \in \mathcal{N}$, $a(n) \leq b(n)$. Therefore,

$$\min_{n\in\mathcal{N}} h_n(\mathbf{p}, \mathbf{g}) \leq \min_{n\in\mathcal{N}}\{h_n(\mathbf{p}, \hat{\mathbf{g}}_{t,n}) + \beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}}\}. \tag{47}$$

Combine Equation (46) and Equation (47), we get $\min_{n\in\mathcal{N}} h_n(\mathbf{p}, \mathbf{g}) - \beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}} \leq \min_{n\in\mathcal{N}} h_l(\mathbf{p}, \hat{\mathbf{g}}_{t,n}) \leq \min_{n\in\mathcal{N}} h_n(\mathbf{p}, \mathbf{g}) + \beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}}$, which is equivalently

$$\left|U_1(\mathbf{p}, \mathbf{g}) - U_1(\mathbf{p}, \hat{\mathbf{g}}_t)\right| \leq \beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}}. \tag{48}$$

Therefore, for any $\delta > 0$ and $t > 0$, with probability at least $1 - N\delta$, Inequality (48) holds.

Following the same reasoning as Equation (45), we can bound $\left|h_n(\mathbf{p}, \tilde{\mathbf{g}}_{t,n}) - h_l(\mathbf{p}, \hat{\mathbf{g}}_{t,n})\right|$ for any $\mathbf{p}$ when $E$ happens (which happens with probability $1 - N\delta$):

$$\begin{aligned}&\left|h_n(\mathbf{p}, \tilde{\mathbf{g}}_{t,n}) - h_n(\mathbf{p}, \hat{\mathbf{g}}_{t,n})\right| \\ =&\left|\sum_{i\in[K]}\left(p_i \tilde{g}(A_i, n, t)\right) - \sum_{i\in[K]}\left(p_i \hat{g}(A_i, n, t)\right)\right| \\ =&\left|\langle \mathbf{p}, \tilde{\mathbf{g}}_{t,n}^\top\rangle - \langle \mathbf{p}, \hat{\mathbf{g}}_{t,n}^\top\rangle\right| \\ =&\left|\langle \mathbf{p}, (\tilde{\mathbf{g}}_{t,n} - \hat{\mathbf{g}}_{t,n})^\top\rangle\right| \\ =&\left|\|\mathbf{p}\|_{\bar{V}_t^{-1}}\|\tilde{\mathbf{g}}_{t,n} - \hat{\mathbf{g}}_{t,n}\|_{\bar{V}_t}\right| \\ \leq&\beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}},\end{aligned} \tag{49}$$

note that $\hat{\mathbf{g}}_{t,n}$ (shown in Equation (14)) is the center of an ellipsoid for the construction of confidence sets, $\tilde{\mathbf{g}}_{t,n}$ is estimated by the FP-OFU algorithm. Denote $U_1(\mathbf{p}, \tilde{\mathbf{g}}_t) = \min_{n\in\mathcal{N}} h_n(\mathbf{p}, \tilde{\mathbf{g}}_{t,n})$, then for any $\delta > 0$ and $t > 0$, with probability at least $1 - N\delta$,

$$\left|U_1(\mathbf{p}, \tilde{\mathbf{g}}_t) - U_1(\mathbf{p}, \hat{\mathbf{g}}_t)\right| \leq \beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}}. \tag{50}$$

Define $R_T^{U_1} = \sum_{t=1}^{T}(U_1(\mathbf{p}^*, \mathbf{g}) - U_1(\mathbf{p}_t, \mathbf{g}))$. We bound $U_1(\mathbf{p}^*, \mathbf{g}) - U_1(\mathbf{p}_t, \mathbf{g})$, which is

$$\begin{aligned}
&U_1(\mathbf{p}^*, \mathbf{g}) - U_1(\mathbf{p}_t, \mathbf{g}) \\
&\overset{(a)}{\leq} U_1(\mathbf{p}_t, \tilde{\mathbf{g}}_t) - U_1(\mathbf{p}_t, \mathbf{g}) \\
&= U_1(\mathbf{p}_t, \tilde{\mathbf{g}}_t) - U_1(\mathbf{p}_t, \hat{\mathbf{g}}_t) + U_1(\mathbf{p}_t, \hat{\mathbf{g}}_t) - U_1(\mathbf{p}_t, \mathbf{g}) \\
&\leq \beta_t(\delta)\|\mathbf{p}_t\|_{\bar{V}_t^{-1}} + \beta_t(\delta)\|\mathbf{p}_t\|_{\bar{V}_t^{-1}} \\
&= 2\beta_t(\delta)\|\mathbf{p}_t\|_{\bar{V}_t^{-1}}
\end{aligned} \tag{51}$$

where (a) is because of the optimality of $(\mathbf{p}_t, \tilde{\mathbf{g}}_t)$ guaranteed by the FP-OFU algorithm. Since both $U_1(\mathbf{p}^*, \mathbf{g})$ and $U_1(\mathbf{p}_t, \mathbf{g})$ are upper-bounded by 1, the difference of them is also upper-bounded by 1, combine this fact with Equation (51), we have

$$\begin{aligned}
&U_1(\mathbf{p}^*, \mathbf{g}) - U_1(\mathbf{p}_t, \mathbf{g}) \\
&\leq \min(1, 2\beta_t(\delta)\|\mathbf{p}_t\|_{\bar{V}_t^{-1}}) \\
&= 2\beta_t(\delta)\min(\frac{1}{2\beta_t(\delta)}, \|\mathbf{p}_t\|_{\bar{V}_t^{-1}}) \\
&\overset{(a)}{\leq} 2\beta_t(\delta)\min(1, \|\mathbf{p}_t\|_{\bar{V}_t^{-1}}),
\end{aligned} \tag{52}$$

where (a) is due to $\beta_t(\delta) > 1$.

Using union bound on the events of Inequality (48) and Inequality (50) hold, for any $\delta > 0$ and $t > 0$, with probability at least $1 - 2N\delta$,

$$\begin{aligned}
R_T^{U_1} &= \sum_{t=1}^{T}(U_1(\mathbf{p}^*, \mathbf{g}) - U_1(\mathbf{p}_t, \mathbf{g})) \\
&\leq \sum_{t=1}^{T} 2\beta_t(\delta)\min(\|\mathbf{p}_t\|_{\bar{V}_t^{-1}}, 1), \\
&\leq \sqrt{T\sum_{t=1}^{T}(2\beta_t(\delta)\min(\|\mathbf{p}_t\|_{\bar{V}_t^{-1}}, 1))^2} \\
&\leq \sqrt{4\beta_T^2(\delta)T\sum_{t=1}^{T}\min(\|\mathbf{p}_t\|^2_{\bar{V}_t^{-1}}, 1)} \\
&\leq \sqrt{8\beta_T^2(\delta)T\log\frac{\det(\bar{V}_T)}{\det(\lambda\mathbf{I})}} \\
&\leq \beta_T(\delta)\sqrt{8KT\log(1+\frac{T}{\lambda})} \\
&\leq (\sqrt{K\log(1+\frac{T}{\lambda\delta})} + (\lambda K)^{\frac{1}{2}})\sqrt{8KT\log(1+\frac{T}{\lambda})} \\
&\leq O\Big(K\sqrt{T\log(\frac{T}{\delta})\log(T)}\Big).
\end{aligned} \tag{53}$$

Without loss of generality, we set $\delta = \frac{1}{T}$. Therefore, with probability at least $1 - \frac{2N}{T}$,

$$R_T^{U_1} \leq O(K\sqrt{T}\log T). \tag{54}$$

There is still a gap between $R_T^{U_1}$ and $R_T$ of Equation (4). Observe that the first term in Equation (4) is $\min_{n\in\mathcal{N}}\sum_{t=1}^{T} r(b_t, n, t)$. Applying the Hoeffding's inequality, we obtain that $\forall n \in \mathcal{N}$, at least with probability $1 - \delta_1$,

$$\left|\sum_{t=1}^{T} r(b_t, n, t) - T\sum_{i\in[K]} p_i^* g(A_i, n)\right| \leq \sqrt{\frac{T}{2}\log(\frac{2}{\delta_1})}. \tag{55}$$

Without loss of generality, we set $\delta_1 = \frac{1}{T^2}$. Therefore,

$$\left|\min_{n\in\mathcal{N}}\sum_{t=1}^{T} r(b_t, n, t) - T\min_{n\in\mathcal{N}}\sum_{i\in[K]} p_i^* g(A_i, n)\right| \leq \sqrt{T\log(2T)}. \tag{56}$$

Next, we give the definition of *martingale* sequence and use the Azuma's inequality to obtain a concentration bound for the second term of Equation (4).

**Definition 2** (Martingale sequence)**.** *A sequence of random variables $X_1, X_2, X_3, \cdots$ is a martingale sequence if (1) $\forall i, \mathbb{E}[\![|X_i|]\!] < \infty$, and (2) $\forall i, \mathbb{E}[X_{i+1}|X_1, X_2, \cdots, X_i] = X_i$. That is, the conditional expected value of the next observation, given all the past observations, is equal to the most recent observation.*

$\forall n \in \mathcal{N}$, define $X_{n,j} = \sum_{t=1}^{j} r(A_{a_t}, n, t) - \sum_{t=1}^{j}\sum_{i\in[K]} p_{t,i} g(A_i, n)$, which is a random variable since $p_{t,i}(i \in [K])$ and $r(A_{a_t}, n, t)$ are random. In the following, we show that $\forall n \in \mathcal{N}$, $\{X_{n,j}\}_{j>0}$ is a martingale sequence. We first show that the first condition of martingale sequence is satisfied since $\forall n \in \mathcal{N}, \forall j > 0$,

$$\begin{aligned}
\mathbb{E}[\![|X_{n,j}|]\!] &= \mathbb{E}\Big[\Big|\sum_{t=1}^{j} r(A_{a_t}, n, t) - \sum_{t=1}^{j}\sum_{i\in[K]} p_{t,i} g(A_i, n)\Big|\Big] \\
&\leq \sum_{t=1}^{j}\mathbb{E}\Big[\Big|r(A_{a_t}, n, t) - \sum_{i\in[K]} p_{t,i} g(A_i, n)\Big|\Big],
\end{aligned} \tag{57}$$

note that $r(A_{a_t}, n, t) \in [0, 1]$ and $\sum_{i\in[K]} p_{t,i} g(A_i, n) \in [0, 1]$. Therefore, we have $\mathbb{E}\Big[\Big|r(A_{a_t}, n, t) - \sum_{i\in[K]} p_{t,i} g(A_i, n)\Big|\Big] \leq 1$ and $\mathbb{E}[\![|X_{n,j}|]\!] \leq j < \infty$. The first condition of Definition 2 is satisfied.

Next, we show that the second condition of martingale sequence is also satisfied. $\forall j$, denote $\mathcal{G}_{n,j} = \{X_{n,1}, X_{n,2}, \cdots, X_{n,j}\}$ and $\mathcal{H}_j = \{\forall n \in \mathcal{N}, \mathbf{p}_t, a_t, r(A_{a_t}, n, t)\}_{t=1}^{j}$, then we have

$$\begin{aligned}
&\mathbb{E}[X_{n,j+1}|X_{n,1}, X_{n,2}, \cdots, X_{n,j}] \\
=&\mathbb{E}[X_{n,j+1}|\mathcal{G}_{n,j}] \\
\overset{(a)}{=}&\mathbb{E}[\mathbb{E}[X_{n,j+1}|\mathcal{H}_j]|\mathcal{G}_{n,j}] \\
\overset{(b)}{=}&\mathbb{E}[\mathbb{E}[(X_{n,j} + \Delta_{n,j+1})|\mathcal{H}_j]|\mathcal{G}_{n,j}] \\
=&X_{n,j} + \mathbb{E}[\mathbb{E}[\Delta_{n,j+1}|\mathcal{H}_j]|\mathcal{G}_{n,j}],
\end{aligned} \tag{58}$$

where (a) is due to tower property of conditional expectation with $\mathcal{G}_{n,j} \subseteq \mathcal{H}_j$ and we denote $\Delta_{n,j+1} = r(A_{a_{j+1}}, n, j+1) - \sum_{i\in[K]} p_{j+1,i} g(A_i, n)$ in (b).

The second term of Equation (58) is equal to 0. This is because

$$\begin{aligned}
&\mathbb{E}[\Delta_{n,j+1}|\mathcal{H}_j] \\
=&\mathbb{E}[(r(A_{a_{j+1}},n,j+1)-\sum_{i\in[K]}p_{j+1,i}g(A_i,n))|\mathcal{H}_j] \\
=&\mathbb{E}[r(A_{a_{j+1}},n,j+1)|\mathcal{H}_j]-\sum_{i\in[K]}p_{j+1,i}g(A_i,n) \\
\overset{(a)}{=}&\sum_{i\in[K]}Pr(E_{j+1,i}|\mathcal{H}_j)\mathbb{E}[r(A_{a_{j+1}},n,j+1)|(\mathcal{H}_j,E_{j+1,i})] \\
&-\sum_{i\in[K]}p_{j+1,i}g(A_i,n) \\
=&\sum_{i\in[K]}p_{j+1,i}g(A_i,n)-\sum_{i\in[K]}p_{j+1,i}g(A_i,n) \\
=&0,
\end{aligned} \tag{59}$$

where we denote event $E_{j+1,i}$ as the event of selecting meta-arm $i$ in round $j+1$ in (a).

Combine Equations (57) and (59), we prove that $\{X_{n,j}\}_{j>0}$ is a martingale sequence. We use the Azuma's inequality to obtain a concentration bound for the second term of Equation (4).

**Lemma 5** (Azuma's Inequality)**.** *Suppose* $\{X_i : i = 1,2,\cdots\}$ *is a martingale sequence and* $|X_i - X_{i-1}| \leq c_i$ *almost surely. Then for all positive integer* $T$ *and all positive real* $\epsilon$, $Pr(|X_T - X_0| \geq \epsilon) \leq 2\exp(\frac{-\epsilon^2}{2\sum_{i=1}^{T}c_i^2})$.

$\forall n \in \mathcal{N}$, $\forall j > 0$, $-1 \leq X_j \leq 1$, therefore $\forall j > 0$, $\left|X_{j+1} - X_j\right| \leq 2$. Using Azuma's inequality and initializing $X_0 = 0$, we get that $\forall n \in \mathcal{N}$, $\forall T > 0$, $\forall \epsilon > 0$, $Pr(\left|\sum_{t=1}^{T} r(A_{a_t},n,t) - \sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i}g(A_i,n)\right| \geq \epsilon) \leq 2\exp(\frac{-\epsilon^2}{8T})$. Set $2\exp(\frac{-\epsilon^2}{8T}) = \frac{1}{T^2}$, then $\forall n \in \mathcal{N}$, $\forall T > 0$, at least with probability $1 - \frac{1}{T^2}$,

$$\left|\sum_{t=1}^{T} r(A_{a_t},n,t) - \sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i}g(A_i,n)\right| \leq 4\sqrt{T\log(2T)}. \tag{60}$$

Therefore,

$$\left|\min_{n\in\mathcal{N}}\sum_{t=1}^{T} r(A_{a_t},n,t) - \min_{n\in\mathcal{N}}\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i}g(A_i,n)\right| \leq 4\sqrt{T\log(2T)}. \tag{61}$$

Define event $\eta := \{\forall n \in \mathcal{N}, \text{Equation (55) and Equation (60) hold}\}$. Event $\bar{\eta}$ is the complement of $\eta$. Using union bound, we obtain that event $\eta$ happens at least with probability $1 - \frac{2N}{T^2}$. Define event $\xi := \{\forall T > 0, \text{Equation (54) holds}\}$, event $\bar{\xi}$ is the complement of $\xi$, we already obtained that event $\xi$ happens at least with probability $1 - \frac{2N}{T}$. When event $\xi \cap \eta$ happens, $R_T$ of Equation (4) can be upper bounded, which is

$$\begin{aligned}
R_T &= \min_{n\in\mathcal{N}}\sum_{t=1}^{T} r(b_t,n,t) - \min_{n\in\mathcal{N}}\sum_{t=1}^{T} r(a_t,n,t) \\
&\leq T\min_{n\in\mathcal{N}}\sum_{i\in[K]} p_i^* g(A_i,n) - \min_{n\in\mathcal{N}}\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i}g(A_i,n) \\
&+ \sqrt{T\log(2T)} + 4\sqrt{T\log(2T)} \\
&\leq Tf(\mathbf{p}^*) - f(\sum_{t=1}^{T}\mathbf{p_t}) + O(\sqrt{T\log(T)}) \\
&\overset{(a)}{\leq} Tf(\mathbf{p}^*) - \sum_{t=1}^{T} f(\mathbf{p_t}) + O(\sqrt{T\log(T)}) \\
&= R_T^{U_1} + O(\sqrt{T\log(T)}) \\
&\leq O(K\sqrt{T}\log T) + O(\sqrt{T\log(T)}) \\
&\leq O(K\sqrt{T}\log T),
\end{aligned} \tag{62}$$

where (a) is because of Lemma 4 of Section C of Appendix.

Using union bound, $Pr(\xi \cap \eta) \geq 1 - \frac{2N}{T^2} - \frac{2N}{T}$, averaging all the events, then $\mathbb{E}[R_T]$ of Equation (5) is

$$\begin{aligned}
\mathbb{E}\left[R_T\right] &\leq \mathbb{E}\left[R_T I(\xi\cap\eta)\right] + \mathbb{E}\left[R_T I(\overline{\xi\cap\eta})\right] \\
&\leq O(K\sqrt{T}\log T) + \mathbb{E}\left[T\cdot I(\overline{\xi\cap\eta})\right] \\
&\leq O(K\sqrt{T}\log T) + T\cdot(\frac{2N}{T^2} + \frac{2N}{T}) \\
&\leq O(K\sqrt{T}\log T).
\end{aligned} \tag{63}$$

■

### F. Proof of Theorem 5

*Proof:*

Define $E = \cap_{n\in\mathcal{N}} E_n$, where $E_n := \{\text{Equation (15) holds for arm } n\}$. Denote $\beta_t(\delta) = \sqrt{K\log(1+\frac{t}{\lambda\delta})} + (\lambda K)^{\frac{1}{2}}$. We first bound $\left|U_2(\mathbf{p},\mathbf{g}) - U_2(\mathbf{p},\hat{\mathbf{g}}_t)\right|$ for all $\mathbf{p}$ when $E$ happens (which, due to Theorem 3, happens with probability $1 - N\delta$):

$$\begin{aligned}
\left|U_2(\mathbf{p},\mathbf{g}) - U_2(\mathbf{p},\hat{\mathbf{g}}_t)\right| &= \left|\sum_{n\in\mathcal{N}}(\mathbf{p}\times\mathbf{g}) - \sum_{n\in\mathcal{N}}(\mathbf{p}\times\hat{\mathbf{g}}_t)\right| \\
&= \left|\sum_{n\in\mathcal{N}}\langle\mathbf{p},\mathbf{g}_n^\top\rangle - \sum_{n\in\mathcal{N}}\langle\mathbf{p},\hat{\mathbf{g}}_{t,n}^\top\rangle\right| \\
&= \left|\sum_{n\in\mathcal{N}}\langle\mathbf{p},(\mathbf{g}_n - \hat{\mathbf{g}}_{t,n})^\top\right| \\
&= \left|\sum_{n\in\mathcal{N}}\left(\|\mathbf{p}\|_{\bar{V}_t^{-1}}\|\mathbf{g}_n - \hat{\mathbf{g}}_{t,n}\|_{\bar{V}_t}\right)\right| \\
&\overset{(a)}{\leq} N\beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}},
\end{aligned} \tag{64}$$

where (a) is due to the definition of event $E$.

Similarly (this time, instead using the definition of event $E$, we use the fact that $\tilde{\mathbf{g}}_t \in M_t$), we can also obtain that ,

for all $\mathbf{p}$,

$$\left|U_2(\mathbf{p}, \tilde{\mathbf{g}}_t) - U_2(\mathbf{p}, \hat{\mathbf{g}}_t)\right| \leq N\beta_t(\delta)\|\mathbf{p}\|_{\bar{V}_t^{-1}}. \tag{65}$$

Combining Equations (64) and (65) gives that when $E$ happens,

$$\begin{aligned} & U_2(\mathbf{p}^*, \mathbf{g}) - U_2(\mathbf{p}_t, \mathbf{g}) \\ & \overset{(a)}{\leq} U_2(\mathbf{p}_t, \tilde{\mathbf{g}}_t) - U_2(\mathbf{p}_t, \mathbf{g}) \\ & \leq U_2(\mathbf{p}_t, \tilde{\mathbf{g}}_t) - U_2(\mathbf{p}_t, \hat{\mathbf{g}}_t) + U_2(\mathbf{p}_t, \hat{\mathbf{g}}_t) - U_2(\mathbf{p}_t, \mathbf{g}) \\ & \leq N\beta_t(\delta)\|\mathbf{p}_t\|_{\bar{V}_t^{-1}} + N\beta_t(\delta)\|\mathbf{p}_t\|_{\bar{V}_t^{-1}} \\ & \leq 2N\beta_t(\delta)\|\mathbf{p}_t\|_{\bar{V}_t^{-1}}, \end{aligned} \tag{66}$$

where (a) is because of the optimality of $(\mathbf{p}_t, \tilde{\mathbf{g}}_t)$ guaranteed by the FP-OFU algorithm. Since both $U_2(\mathbf{p}^*, \mathbf{g})$ and $U_2(\mathbf{p}_t, \mathbf{g})$ are upper-bounded by $N$, the difference of them is also upper-bounded by $N$, combine this fact with Equation (66), we have

$$\begin{aligned} & U_2(\mathbf{p}^*, \mathbf{g}) - U_2(\mathbf{p}_t, \mathbf{g}) \\ & \leq \min(N, 2N\beta_t(\delta)\|\mathbf{p}_t\|_{\bar{V}_t^{-1}}) \\ & = 2N\beta_t(\delta)\min(\frac{1}{2\beta_t(\delta)}, \|\mathbf{p}_t\|_{\bar{V}_t^{-1}}) \\ & \overset{(a)}{\leq} 2N\beta_t(\delta)\min(1, \|\mathbf{p}_t\|_{\bar{V}_t^{-1}}), \end{aligned} \tag{67}$$

where (a) is due to $\beta_t(\delta) > 1$.

Following the same reasoning as Equation (53), we obtain that for any $\delta > 0$ and $t > 0$, with probability at least $1 - 2N\delta$,

$$\begin{aligned} R_T^{U_2} &= \sum_{t=1}^{T}(U_2(\mathbf{p}^*, \mathbf{g}) - U_2(\mathbf{p}_t, \mathbf{g})) \\ &\leq O\Big(NK\sqrt{T\log(\frac{T}{\delta})\log(T)}\Big). \end{aligned} \tag{68}$$

Without loss of generality, we set $\delta = \frac{1}{T}$. Therefore, with probability at least $1 - \frac{2N}{T}$,

$$R_T^{U_2} \leq O(NK\sqrt{T}\log T). \tag{69}$$

By the similar reasoning as Equation (55), using union bound over all $n \in \mathcal{N}$ and setting $\delta_1 = \frac{1}{T^2}$, at least with probability $1 - \frac{N}{T^2}$,

$$\left|\sum_{n\in\mathcal{N}}\sum_{t=1}^{T} r(b_t, n, t) - T\sum_{n\in\mathcal{N}}\sum_{i\in[K]} p_i^* g(A_i, n)\right| \leq N\sqrt{T\log(2T)}. \tag{70}$$

By Equation (60) and union bound, at least with probability $1 - \frac{N}{T^2}$,

$$\left|\sum_{n\in\mathcal{N}}\sum_{t=1}^{T} r(A_{a_t}, n, t) - \sum_{n\in\mathcal{N}}\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i, n)\right| \leq 4N\sqrt{T\log(2T)}. \tag{71}$$

Define event $\eta := \{\text{Equation (70) and Equation (71) hold}\}$. Event $\bar{\eta}$ is the complement of $\eta$. Using union bound, we obtain that event $\eta$ happens at least with probability $1 - \frac{2N}{T^2}$. Define event $\xi := \{\forall T > 0, \text{Equation (69) holds}\}$, event $\bar{\xi}$ is the complement of $\xi$. We already obtained that event $\xi$ happens at least with probability $1 - \frac{2N}{T}$. When event $\xi \cap \eta$ happens, $R_T$ of Equation (7) is upper bounded:

$$\begin{aligned} R_T &= \sum_{n\in\mathcal{N}}\sum_{t=1}^{T} r(b_t, n, t) - \sum_{n\in\mathcal{N}}\sum_{t=1}^{T} r(a_t, n, t) \\ &\leq T\sum_{n\in\mathcal{N}}\sum_{i\in[K]} p_i^* g(A_i, n) - \sum_{n\in\mathcal{N}}\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i, n) \\ &\quad + N\sqrt{T\log(2T)} + 4N\sqrt{T\log(2T)} \\ &= R_T^{U_2} + O(N\sqrt{T\log(T)}) \\ &\leq O(NK\sqrt{T}\log T) + O(N\sqrt{T\log(T)}) \\ &\leq O(NK\sqrt{T}\log T). \end{aligned} \tag{72}$$

Following the same reasoning as Equation (63), $\mathbb{E}[R_T]$ of Equation (8) is upper-bounded by $O(NK\sqrt{T}\log T)$.

■

## REFERENCES


[1] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[2] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466–1478, 2012.

[3] P. Rusmevichientong and D. P. Williamson, "An adaptive algorithm for selecting profitable keywords for search-based advertising services," in *Proceedings of the 7th ACM Conference on Electronic Commerce*, 2006, pp. 260–269.

[4] Z. Guo, M. Li, and M. Krunz, "Exploiting successive interference cancellation for spectrum sharing over unlicensed bands," *IEEE Transactions on Mobile Computing*, pp. 1–18, 2023.

[5] L. Tassiulas and S. Sarkar, "Maxmin fair scheduling in wireless networks," in *Proceedings of Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, 2002, pp. 763–772.

[6] D. Ghosh, A. Verma, and M. K. Hanawal, "Learning and fairness in energy harvesting: A maximin multi-armed bandits approach," in *Proc. of International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[7] 3GPP, "Feasibility study on licensed-assisted access to unlicensed spectrum," Standard (TR) 36.889, V13.0.0, 2015.

[8] IEEE, "Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," IEEE Standard 802.11, 2012.

[9] S. Yun and L. Qiu, "Supporting WiFi and LTE coexistence," in *Proc. of IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 810–818.

[10] A. Keyhanian, A. Leivadeas, I. Lambadaris, and I. Marmokos, "Analyzing the coexistence of Wi-Fi and LAA-LTE towards a proportional throughput fairness," in *Proceedings of the 16th ACM International Symposium on Mobility Management and Wireless Access*, 2018, pp. 95–101.

[11] X. Wang, T. Q. Quek, M. Sheng, and J. Li, "Throughput and fairness analysis of Wi-Fi and LTE-U in unlicensed band," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 1, pp. 63–78, 2016.

[12] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.

[13] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. of International conference on machine learning*. PMLR, 2013, pp. 151–159.

[14] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits," in *Proc. of Artificial Intelligence and Statistics*. PMLR, 2015, pp. 535–543.

[15] V. Patil, G. Ghalme, V. Nair, and Y. Narahari, "Achieving fairness in the stochastic multi-armed bandit problem," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 7885–7915, 2021.
[16] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1799–1813, 2019.
[17] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.I.D. rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 1987.
[18] C. A. Oliveira and P. M. Pardalos, "A survey of combinatorial optimization problems in multicast routing," *Computers & Operations Research*, vol. 32, no. 8, pp. 1953–1981, 2005.
[19] X. Hu, D. Ngo, A. Slivkins, and S. Z. Wu, "Incentivizing combinatorial bandit exploration," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 173–37 183, 2022.
[20] W. Chen, W. Hu, F. Li, J. Li, Y. Liu, and P. Lu, "Combinatorial multi-armed bandit with general reward functions," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
[21] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Combinatorial cascading bandits," *Advances in Neural Information Processing Systems*, vol. 28, 2015.
[22] Y. Chen, A. Cuellar, H. Luo, J. Modi, H. Nemlekar, and S. Nikolaidis, "Fair contextual multi-armed bandits: Theory and experiments," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 181–190.
[23] W. Huang, K. Labille, X. Wu, D. Lee, and N. Heffernan, "Achieving user-side fairness in contextual bandits," *Human-Centric Intelligent Systems*, vol. 2, no. 3, pp. 81–94, 2022.
[24] I. Bistritz, T. Baharav, A. Leshem, and N. Bambos, "My fair bandit: Distributed learning of max-min fairness with multi-player bandits," in *International Conference on Machine Learning*. PMLR, 2020, pp. 930–940.
[25] N. A. Grupen, B. Selman, and D. D. Lee, "Cooperative multi-agent fairness and equivariant policies," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 9, 2022, pp. 9350–9359.
[26] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.
[27] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth, "Fairness in learning: Classic and contextual bandits," *Advances in neural information processing systems*, vol. 29, 2016.
[28] M. S. Talebi and A. Proutiere, "Learning proportionally fair allocations with low regret," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 2, no. 2, pp. 1–31, 2018.
[29] S. Hossain, E. Micha, and N. Shah, "Fair algorithms for multi-agent multi-armed bandits," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 005–24 017, 2021.
[30] S. Barman, A. Khan, A. Maiti, and A. Sawarni, "Fairness and welfare quantification for regret in multi-armed bandits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, 2023, pp. 6762–6769.
[31] M. Jones, H. Nguyen, and T. Nguyen, "An efficient algorithm for fair multi-agent multi-armed bandit with low regret," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 8159–8167.
[32] Z. Guo, C. Zhang, M. Li, and M. Krunz, "Fair coexistence of heterogeneous networks: A novel probabilistic multi-armed bandit approach," in *21st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. IEEE, 2023, pp. 1–8.
[33] A. Slivkins *et al.*, "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.
[34] X. L. Huang and B. Bensaou, "On max-min fairness and scheduling in wireless ad-hoc networks: Analytical framework and implementation," in *Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing*, 2001, pp. 221–231.
[35] L. B. Jiang and S. C. Liew, "Proportional fairness in wireless LANs and ad hoc networks," in *Proc. of IEEE Wireless Communications and Networking Conference*, vol. 3, 2005, pp. 1551–1556.
[36] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.
[37] C. Wu, T. Li, Z. Zhang, and Y. Yu, "Bayesian optimistic optimization: Optimistic exploration for model-based reinforcement learning," *Advances in neural information processing systems*, vol. 35, pp. 14 210–14 223, 2022.
[38] J. Zimmert, H. Luo, and C.-Y. Wei, "Beating stochastic and adversarial semi-bandits optimally and simultaneously," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7683–7692.
[39] F. Gembicki, "Vector optimization for control with performance and parameter sensitivity indices," *Ph. D. thesis, Case Western Reserve Univ.*, 1974.
[40] F. S. Hillier, *Introduction to operations research*. McGrawHill, 2001.
[41] C. Daskalakis, S. Skoulakis, and M. Zampetakis, "The complexity of constrained min-max optimization," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 1466–1478.
[42] A. B. Sediq, R. H. Gohary, R. Schoenen, and H. Yanikomeroglu, "Optimal tradeoff between sum-rate efficiency and Jain's fairness index in resource allocation," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3496–3509, 2013.
[43] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.
[44] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," *Advances in neural information processing systems*, vol. 24, 2011.

**Zhiwu Guo** received the B.S. and M.S. degrees in electrical and computer engineering from University of Electronic Science and Technology of China, in 2014 and in 2017, respectively. He is currently working towards the Ph.D. degree at the Department of Electrical and Computer Engineering, the University of Arizona. His research interests include heterogeneous network spectrum sharing, interference cancellation, and artificial intelligence for wireless communication and networking.

**Chicheng Zhang** is an Assistant Professor in the Department of Computer Science at the University of Arizona. His research interests lie on the theory and applications of interactive machine learning, with paradigms including active learning, contextual bandits, reinforcement learning, imitation learning, etc. He constantly serves as a program committee member and publishes in many conferences and journals in machine learning and learning theory, such as ICML, COLT, NeurIPS, ICLR, AISTATS, ALT, TMLR. His paper received an outstanding paper runner-up award at ICML 2022.

**Ming Li** (M'11, SM'17, F'24) is currently a Professor of ECE at University of Arizona (also affiliated with CS). He was an Assistant Professor in the CS Department at Utah State University from 2011 to 2015. He received his Ph.D. in ECE from Worcester Polytechnic Institute, MA, in 2011. His research interests include wireless network optimization and machine learning, network security and privacy, and cyber-physical system security. He has published more than 135 journal and conference papers, with an h-index of 45. He received the NSF CAREER Award in 2014, the ONR YIP Award in 2016, and several paper awards, including the best paper award from ACM WiSec 2020. He served on the editorial boards of IEEE TMC and TDSC and is currently an associate editor for IEEE TIFS. He was a TPC Co-chair of IEEE CNS 2022. He is a Fellow of IEEE, and a member of ACM.

**Marwan Krunz** (Fellow, IEEE) is currently a Regents Professor of Electrical and Computer Engineering with The University of Arizona. He also holds a joint appointment as a Professor of computer science. He directs the Broadband Wireless Access and Applications Center (BWAC), a multi-university NSF/industry center that focuses on next-generation wireless technologies. He is an Affiliated Faculty of the UA Cancer Center. Previously, he served as the Site Director for the Connection One Center. From 2015 to 2023, he was the Kenneth VonBehren Endowed Professor in electrical and computer engineering. He served as the chief scientist for two startup companies that focus on 5G and beyond systems and machine learning for wireless communications. He has published more than 330 journal articles and peer-reviewed conference papers and is a named inventor on ten patents. His latest H-index is 62. His research interests include wireless communications and protocols, network security, and machine learning. He was an Arizona Engineering Faculty Fellow and an IEEE Communications Society Distinguished Lecturer. He received the NSF CAREER Award. He was the TPC Chair for several conferences and symposia, INFOCOM in 2004, SECON in 2005, WoWMoM in 2006, and Hot Interconnects 9. He was the General Chair of WiOpt'23, the Vice Chair of WiOpt'16, and the General Co-Chair of WiSec'12. He has served as the Editor-in-Chief for IEEE TRANSACTIONS ON MOBILE COMPUTING and an editor for numerous IEEE journals.