Efficient Fair Probabilistic Multi-armed Bandit for Real-Time Resource Allocation in Spectrum Sharing

Zhiwu Guo* Chicheng Zhang† Ming Li* Marwan Krunz*
* Electrical and Computer Engineering, The University of Arizona, Tucson, Arizona, 85721, USA
† Computer Science, The University of Arizona, Tucson, Arizona, 85721, USA
Email: zhiwuguo@arizona.edu, chichengz@cs.arizona.edu, {lim, krunz}@arizona.edu

***Abstract*—Spectrum sharing is critical to addressing spectrum shortages driven by the rapid growth in mobile users and data traffic. Interference mitigation poses a significant challenge to realize efficient and fair spectrum resource allocation in dynamic environments. In this paper, we introduce an efficient fair probabilistic multi-armed bandit (EFP-MAB) algorithm, which optimizes the concurrent link transmission schedules to enhance coexisting networks' throughput, by exploiting interference cancellation techniques. We provide a comprehensive performance analysis of EFP-MAB, deriving a bound on its expected regret. The expected regret of EFP-MAB improves by $O(\sqrt{K\log(T)})$ compared to the state-of-the-art FP-OFU algorithm, where T is the time horizon and K is the number of meta-arms. The computational complexity of EFP-MAB is $poly(N, K)$, where N represents the number of arms, marking a significant efficiency improvement over the FP-OFU algorithm, which lacks polynomial time global optimality guarantees. Extensive simulations in a heterogeneous technology coexistence scenario demonstrate that EFP-MAB significantly outperforms FP-OFU in both expected regret and efficiency.**

***Index Terms*—Spectrum sharing, online learning, multi-armed bandit, fairness, efficiency.**

I. Introduction

The rapid growth in mobile users and their data demands has placed significant strain on next-generation wireless networks. A key challenge in addressing this issue lies in the limited availability of spectrum resources. Spectrum sharing has emerged as a promising solution to optimize spectrum utilization across diverse applications. Two prominent use cases illustrate the potential of spectrum sharing, as outlined below. (1) *5G Heterogeneous Networks* (HetNets), which involve sharing licensed spectrum among different network technologies within a 5G network, such as macrocells and small cells. HetNet, introduced in the 3rd Generation Partnership Project (3GPP) Release 12 [1], enables various types of small cells to coexist while overlapping with macrocell networks in the same location and using the same spectrum band. (2) *Heterogeneous Technology Coexistence*, which focuses on sharing unlicensed spectrum among various wireless technologies, including Wi-Fi, LTE, and 5G, to enhance network capacity and coverage. For example, 3GPP has proposed extending LTE License Assisted Access (LAA) [2] and 5G New Radio (NR) [3] to operate in the 5 GHz unlicensed spectrum, which is widely utilized by Wi-Fi. Other examples include IoT and cellular networks coexistence [4], and Wi-Fi/Bluetooth/ZigBee coexistence [5].

One of the critical challenges in efficiently sharing spectrum resources is to mitigate the mutual interference among different networks using the same band (e.g., between macrocells and small cells in HetNets, or LTE-LAA and Wi-Fi) [6]. The prevalent approach has been interference avoidance, which allocates different resource blocks to different networks' users, including scheduling them in different sub-carriers or time slots [7], [8]. For spectrum sharing in unlicensed bands, listen-before-talk (LBT) [2] has been introduced in the medium access control (MAC) layer of LTE-LAA and 5G unlicensed NR networks, similar to carrier sense multiple access with collision avoidance (CSMA/CA) [9] in Wi-Fi systems. These protocols require networks to sense the shared channel and transmit only if the detected energy level is below a predefined threshold. Despite their widespread applications, a significant limitation of interference avoidance-based spectrum-sharing resource allocation mechanisms is their low spectrum utilization efficiency. In addition, *fairness* is another key issue incurred by existing spectrum sharing paradigms. For example, in LAA/Wi-Fi coexistence scenarios, Wi-Fi performance can be severely degraded in the presence of LAA [10], [11]. This occurs because Wi-Fi frequently defers to LAA transmissions, leaving limited opportunities for Wi-Fi to transmit. The imbalance arises from the asymmetric carrier sensing parameter settings between Wi-Fi and LAA networks, which often result in the LAA networks dominating the shared spectrum.

Recently, physical-layer interference cancellation (IC) techniques have been proposed as promising alternatives to interference avoidance. For example, multiple-input and multiple-output (MIMO) [12] and successive interference cancellation (SIC) [13] enable two or more links to transmit simultaneously in the same frequency, while canceling out their interference. IC techniques have been shown to significantly enhance the performance of homogeneous wireless (e.g., Wi-Fi, cellular networks, etc [14], [15]), but also adopted to mitigate the interference across heterogeneous network protocols (e.g., Wi-Fi/ZigBee coexistence [16], LAA/Wi-Fi coexistence [13]).

Existing approaches to resource allocation in spectrum-sharing applications can be divided into either offline optimization [17] or online learning techniques [18], [19]. However, these methods have notable limitations. Offline optimization cannot easily adapt to the dynamic changes of wireless channels (e.g., due to fading or user mobility), as offline training-based channel/link quality estimation incurs high overhead.

On the other hand, online learning-based resource allocation algorithms (such as Multi-armed bandit (MAB) or Deep Reinforcement Learning (DRL)) dynamically interact with the wireless environment and optimize the policy on-the-fly [20], [21]. While DRL approaches can make real-time decisions, they often suffer from slow convergence [22] and lack reliable performance guarantees. Recently, several works also studied fair resource allocation under the MAB framework, with applications to network optimization [23], [24] or spectrum coexistence [25]–[27]. Specifically, Guo et. al. proposed a fair probabilistic MAB framework for online resource scheduling in spectrum sharing with interference cancellation techniques [25], [27]. A unique challenge of modeling and solving this problem is the combinatorial nature of multiple concurrent transmitting sets (arms), whose rewards structure is non-linear and depends on the combination itself. Therefore, previous fair MAB schemes are either not applicable to our problem [23], [24] , or incur significant regret and computational overheads [25], [27], which makes them difficult to be adopted in practice.

In reality, resource allocation or scheduling in shared spectrum applications often has stringent latency requirements, making *efficiency* a critical consideration. For instance, scheduling for ultra-reliable low-latency communications (URLLC) in 5G systems is 1 ms [7]. Additionally, real-world channel conditions remain highly correlated only within the channel coherence time, which is less than 50 ms at 5.9 GHz in rural and suburban environments [28]. Hence, it is essential to design efficient resource allocation or scheduling algorithms that can adapt to channel dynamics in real-time, thereby enhancing the performance of spectrum-sharing systems.

A. Contributions

To address the above requirements and challenges, in this work, we propose an efficient fair probabilistic multi-armed bandit (EFP-MAB) algorithm, which has significant practical applicability in resource allocation across various spectrum-sharing applications, including addressing fairness concern in NR/LAA/Wi-Fi heterogeneous network coexistence [2], [3]. Unlike traditional MAB frameworks, which deterministically select the optimal arm upon convergence, EFP-MAB selects each arm probabilistically to ensure that fairness requirements are met. EFP-MAB is a novel and generalized fair probabilistic MAB framework that incorporates arm fairness either as an objective or a constraint, while also supporting a more comprehensive combinatorial structure with dependent arm rewards. Our main contributions are summarized as follows:

(1) We introduce an efficient fair probabilistic MAB (EFP-MAB) algorithm, which incorporates either a max-min fairness objective or fairness constraints that enforce a minimum selection fraction for each arm.

(2) We analyze the regret of EFP-MAB. For the max-min fairness version of EFP-MAB, we derive a sublinear upper bound of $O(\sqrt{KT\log(T)})$ for the expected regret, where T is the time horizon and K is the number of meta-arms. For the fairness constraints version of EFP-MAB, we obtain an upper bound $O(N\sqrt{KT\log(T)})$ for the expected regret, where N represents the number of arms. Notably, the expected regret of EFP-MAB achieves an improvement of $O(\sqrt{K\log(T)})$ compared to the state-of-the-art FP-OFU algorithm [27].

(3) The computational complexity of EFP-MAB is $poly(N, K)$, a significant improvement in efficiency compared to the FP-OFU algorithm, which lacks polynomial time global optimality guarantees.

(4) We conduct extensive simulations to evaluate the performance of EFP-MAB in the context of heterogeneous technology coexistence and compare it with the state-of-the-art baseline FP-OFU algorithm. The results demonstrate that EFP-MAB significantly outperforms the baseline in terms of expected regret and efficiency. Our algorithm can achieve ms-level run time, which is applicable to real-time online spectrum resource allocation.

II. Related Work

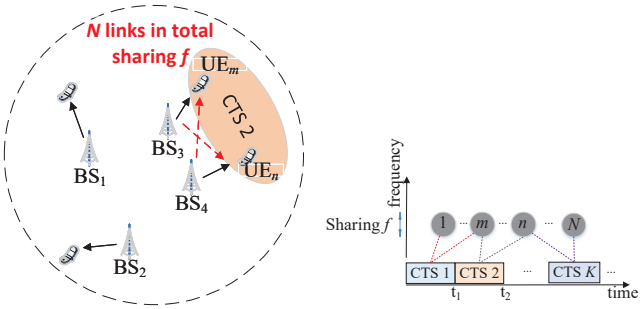
The authors in [29] addressed the joint resource allocation problem in 5G HetNets, incorporating both user-level and cell-level fairness. However, their solution relied on offline optimizations, which introduce significant overhead and delays due to the need to collect necessary information (e.g., channel conditions) beforehand. Similarly, the work in [30] focused on high-mobility scenarios in 5G HetNets, proposing a deep reinforcement learning-based resource allocation algorithm to configure time-division duplex patterns. However, this approach overlooks fairness in its modeling. In [18], authors applied reinforcement learning techniques to tune contention window size for both LTE and Wi-Fi nodes in LTE/Wi-Fi coexistence scenarios. Both approaches in [18], [30] lack theoretical performance guarantees and complexity analysis.

Most existing bandit algorithms with fairness considerations focus exclusively on either fairness objectives or fairness constraints, rather than addressing both. For instance, [24] proposed fair MAB algorithms that ensure each arm is selected at least a pre-specified fraction of the time, while Maxmin UCB, introduced in [31] integrates a max-min fairness objective into the UCB algorithm. Furthermore, most existing bandit algorithms such as UCB [32] and Maxmin UCB [31], converge to deterministically selecting an optimal arm over time. In contrast, Guo et al. [27] introduced the Fair Probabilistic Optimism in the Face of Uncertainty (FP-OFU) algorithm, which incorporates both a max-min fairness objective and fairness constraints that enforce a minimum selection fraction for each arm. The idea of probabilistically selecting arms, as introduced in [27] and this work, has been similarly applied in the traditional fair MAB problem [24], combinatorial sleeping bandits with fairness constraints [23], and combinatorial bandits with probabilistically triggered arms [33]. However, FP-OFU suffers from a high-order expected regret with respect to the number of meta-arms and lacks polynomial-time global optimality guarantees.

III. System Model and Problem Formulation

A. System Model

We consider a spectrum sharing system where N individual links share the same frequency band in a common geographical area, as illustrated in Fig. 1(a). The downlink transmissions from BS_3 to UE_m and BS_4 to UE_n are referred to as link m and link n, respectively. A Concurrent Transmitting Set (CTS) consists of one or more links scheduled for simultaneous transmission. For instance, in Fig. 1(a), CTS 2 includes link m and link n. When CTS 2 is scheduled for transmission, each UE in this CTS receives the signal from its intended transmitter (illustrated by black solid arrows) while experiencing interference from other links in the CTS at the same time (indicated by red dashed arrows). Various IC techniques can be applied to mitigate the interference (either fully or partially). In general, the throughput of each link in a CTS is a function of the SINR, which depends on the CTS topology and channels. However, since the channels are assumed to be unknown, the performance of each link under each CTS must be learned from data. We assume there are K CTSes in total, as shown Fig. 1(b), the objective of the spectrum sharing system is to determine an optimal scheduling policy that maximizes the overall spectrum efficiency (e.g., long-term network throughput), while balancing individual link-level fairness across all time slots.



(a) System Model: N links in sharing the same frequency band. (b) Frequency-time representation of total(b) the scheduling problem.

Fig. 1: System model of spectrum sharing system that allows concurrent transmissions scheduling.

B. Problem Formulation

In this paper, we model the CTS scheduling problem in the aforementioned spectrum sharing systems using a multi-armed bandit (MAB) approach. Traditional MAB algorithms, such as the Upper Confidence Bound (UCB) algorithm [32], are not inherently designed with fairness considerations. Furthermore, deterministically selecting a single CTS for transmission in each slot may incur fairness issues for individual links. To address these challenges, we formulate a *fair probabilistic MAB problem*. This approach extends the traditional MAB framework by incorporating fairness either as part of the objective function or as constraints, achieved by probabilistically selecting each arm. For the fairness objective, we adopt the *max-min fairness* criterion [34], [35], a widely used metric in wireless networking that ensures fair resource allocation among competing links. Regarding fairness constraints, we enforce a minimum transmission requirement for each link, ensuring that each link is transmitted at least a specified number of times over the total transmission period.

We consider a discrete-time system consisting of T time steps and denote $\mathcal{N} = \{1, 2, \cdots, N\}$ as the set of all individual links (referred to as arms). In time step $t (1 \leq t \leq T)$, a CTS (referred to as a meta-arm) is probabilistically selected according to the meta-arm selection vector $\mathbf{p}_t$, where $p_{t,i}$ represents the probability of selecting meta-arm i in time step t. A meta-arm represents an association of one or more arms. We denote a meta-arm as A and an arm as a. Let $\mathcal{F}$ represent the feasible set of all meta-arms, and K denote the number of meta-arms in $\mathcal{F}$. When meta-arm A is selected in time step t, the learning agent receives rewards $r(A, a, t)$ (e.g., success packet delivery or not, which is binary) for all arms $a \in \mathcal{N}$. The reward $r(A, a, t)$ is assumed to be randomly sampled from an unknown distribution $\mathcal{D}_{A,a}$, which depends on both the arm a and meta-arm A. Furthermore, we assume that $\{r(A, a, t) : 1 \leq t \leq T\}$ are independent and identically distributed (i.i.d.). The two problems (max-min fairness objective and fairness constraints) are formulated respectively in the following sections.

1) Max-min Fairness Objective: Given a meta-arm $A \in \mathcal{F}$, let $g(A, a)$ denote the true mean reward for arm a (which represents the successful decoding probability of a link), where $a \in \mathcal{N}$. If arm a is not associated with meta-arm A, we denote $g(A, a) = 0$. If the environment $(g(A, a))_{A \in \mathcal{F}, a \in \mathcal{N}}$ is known in advance, the optimal meta-arm selection vector $\mathbf{p}$ can be determined by solving the 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 $f(\mathbf{p}) = \min_{a \in \mathcal{N}} \{\sum_{i \in [K]} \big(p_i \times g(A_i, a)\big)\}$, p_i represents the i-th element of $\mathbf{p}$, and A_i denote the i-th meta-arm. The objective function $\sum_{i \in [K]} \big(p_i \times g(A_i, a)\big)$ represents the total long-term throughput of link (i.e., arm) a.

Let $\hat{g}(A_i, a, t)$ denote the empirical average reward of arm $a \in \mathcal{N}$ up to t. The number of times meta-arm A_i has been selected by t, denoted as $n_t(A_i)$, is given by $n_t(A_i) = \sum_{\tau=1}^{t} \mathbb{I}(c_\tau = A_i)$, where c_τ is the index of meta-arm selected in time step τ, $\mathbb{I}(E) = 1$ if predicate E is true, and equal to 0 otherwise. The empirical average reward $\hat{g}(A_i, a, t)$ is computed as $\hat{g}(A_i, a, t) = \frac{1}{n_t(A_i)} \sum_{\tau=1}^{t} \mathbb{I}(c_\tau = A_i) r(A_i, a, \tau)$. During the learning process, the agent explores all meta-arms to obtain more accurate estimates of $\hat{g}(A_i, a, t)$ while simultaneously trying to make the best possible decisions based on currently available information. Both exploration and exploitation can lead to a cumulative reward compared to the optimal probabilistic selection that would be chosen if the environment $\{g(A_i, a) : \forall i \in [K], a \in \mathcal{N}\}$ were known in advance. This difference is referred to as *regret*. The objective of the learning agent is to minimize this incurred regret.

The regret for **Opt-min** is defined as follows [27]:

$$R_T = \min_{a\in\mathcal{N}} \sum_{t=1}^{T} r(b_t, a, t) - \min_{a\in\mathcal{N}} \sum_{t=1}^{T} r(c_t, a, t), \quad (2)$$

where $(b_t)_{t=1}^T$ is an i.i.d. sequence of meta-arms drawn from $\mathbf{p}^*$, the optimal solution of **Opt-min** in Equation (1); $(c_t)_{t=1}^T$ is the sequence of meta-arms chosen by the learning agent during the learning process.

Accordingly, the expected regret is defined as

$$\mathbb{E}[R_T] = \mathbb{E}[\min_{a\in\mathcal{N}} \sum_{t=1}^{T} r(b_t, a, t)] - \mathbb{E}[\min_{a\in\mathcal{N}} \sum_{t=1}^{T} r(c_t, a, t)], \quad (3)$$

where the expectation is taken over the following sources of randomness: (1) $(b_t)_{t=1}^T$; (2) $(c_t)_{t=1}^T$; (3) rewards r sampled from the environment.

2) Total Reward Maximization Subject To Fairness Constraints: We model the fairness constraint for each arm as a targeted minimum selection fraction, denoted by d_a for arm a. If the environment $(g(A,a))_{A\in\mathcal{F},a\in\mathcal{N}}$ is known in advance, the optimal selection vector $\mathbf{p}$ can be determined by solving the following optimization problem:

$$\begin{aligned} \textbf{Opt-cons}: \max_{\mathbf{p}} \quad & \sum_{a\in\mathcal{N}} \sum_{i\in[K]} \big(p_i \times g(A_i, a)\big) \\ \text{s.t.} \quad & 0 \le p_i \le 1, i \in [K], \\ & \sum_{i,i\in[K]} p_i = 1, \\ & \sum_{i\in[K]} \big(p_i \times \mathbb{I}(E_{a,A_i})\big) \ge d_a, \forall a \in \mathcal{N}, \end{aligned} \quad (4)$$

where E_{a,A_i} represents the event that arm a is associated with meta-arm A_i.

The regret for **Opt-cons** is defined as follows [27]:

$$R_T = \sum_{a\in\mathcal{N}} \sum_{t=1}^{T} r(b_t, a, t) - \sum_{a\in\mathcal{N}} \sum_{t=1}^{T} r(c_t, a, t), \quad (5)$$

where $(b_t)_{t=1}^T$ and $(c_t)_{t=1}^T$ retain similar meanings as defined in Equation (2).

The expected regret is defined to be

$$\mathbb{E}[R_T] = \mathbb{E}[\sum_{a\in\mathcal{N}} \sum_{t=1}^{T} r(b_t, a, t)] - \mathbb{E}[\sum_{a\in\mathcal{N}} \sum_{t=1}^{T} r(c_t, a, t)]. \quad (6)$$

Remark 1: For **Opt-cons**, in addition to regret, fairness constraint violation must also be analyzed. However, our proposed EFP-MAP algorithm guarantees no violations of the fairness constraints, as the optimization variable $\mathbf{p}_t$ always remains within its feasible set. Consequently, details regarding this performance metric are omitted in this paper. Refer to [27] for a detailed formulation of fairness constraint violations.

IV. Efficient Fair Probabilistic Multi-armed Bandit Algorithm

To address the above formulated problems, we propose an efficient fair probabilistic multi-armed bandit (EFP-MAB) algorithm and analyze its performance in this section.

The procedures of EFP-MAB algorithm are outlined in Alg. 1. The core idea of EFP-MAB is to construct a rectangular confidence interval for each combination of meta-arm A and arm a based on the empirical mean reward. To achieve this, the algorithm maintains two key variables for each meta-arm A and arm n at time step t: $n_t(A)$ and $\hat{g}(A,a,t)$, as described in Step 1. For $1 \le t \le T$, the algorithm calculates the upper confidence bound of arm a under meta-arm A, denoted as $\bar{g}(A,a,t)$, as the sum of its empirical mean $\hat{g}(A,a,t-1)$ and its confidence interval $\sqrt{\frac{2\log(T)}{n_{t-1}(A)+1}}$, as shown in Step 3. EFP-MAB computes the meta-arm selection vector $\mathbf{p}_t$ based on either the max-min fairness objective or total reward maximization subject to fairness constraints, as indicated in Steps 5 and 7, respectively. Specifically, in Step 5, with the goal of the max-min fairness objective, Oracle$_1$ is defined as:

$$\text{Oracle}_1((\bar{g}(A,a,t))_{A\in\mathcal{F},a\in\mathcal{N}}) = \arg\max_{\mathbf{p}\in\mathcal{F}_1} U_1(\mathbf{p}, \bar{\mathbf{g}}_t), \quad (7)$$

where $\mathcal{F}_1$ is the feasible set of **Opt-min** in Equation (1), $\bar{\mathbf{g}}_t \in [0,1]^{K\times N}$ is the matrix of UCBs for all meta-arms and arms at time step t, with element $\bar{g}_{i,a}$ representing the upper confidence bound for arm a under meta-arm A_i, $U_1(\mathbf{p}, \bar{\mathbf{g}}_t) = \min_{a\in\mathcal{N}} \sum_{i\in[K]} \big(p_i \times \bar{g}(A_i, a, t)\big)$. In Step 7, with the goal of total reward maximization subject to fairness constraints, Oracle$_2$ is defined as:

$$\text{Oracle}_2((\bar{g}(A,a,t))_{A\in\mathcal{F},a\in\mathcal{N}}) = \arg\max_{\mathbf{p}\in\mathcal{F}_2} U_2(\mathbf{p}, \bar{\mathbf{g}}_t), \quad (8)$$

where $\mathcal{F}_2$ is the feasible set of **Opt-cons** in Equation (4), $U_2(\mathbf{p}, \bar{\mathbf{g}}_t) = \sum_{a\in\mathcal{N}} \sum_{i\in[K]} \big(p_i \times \bar{g}(A_i, a, t)\big)$. In Step 9, the algorithm samples a meta-arm A based on the categorical distribution of $\mathbf{p}_t$. The selected meta-arm is played, the rewards of all arms are observed, $n_t(A)$ and all $\hat{g}(A,a,t)$, $\forall a \in \mathcal{N}$ are updated in Step 10.

The performance of the EFP-MAB algorithm is analyzed separately for the max-min fairness objective and total reward maximization subject to fairness constraints in the following sections.

A. Performance Analysis under Max-min Fairness Objective

We start by analyzing EFP-MAB for the max-min fairness objective. The standard MAB analysis methodology can not be directly applied to the analysis of the EFP-MAB algorithm, as traditional MAB is typically analyzed based on a discrete arm space. In contrast, the approach proposed in this work and in [27] is analyzed in terms of a continuous space defined by the meta-arm selection vector $\mathbf{p}$. Unlike [27], this work employs an upper confidence bound for each arm, with the random variable $\mathbf{p}$ appearing in the bounds of the utility functions, as shown in Equations (11) and (12). This introduces additional challenges to the analysis compared to [27].

Algorithm 1 Efficient Fair Probabilistic Multi-armed Bandit (EFP-MAB) Algorithm

For each meta-arm A and each arm a, maintain: (1) variable $n_t(A)$ as the number of times meta-arm A is played until t, initially 0; (2) variable $\hat{g}(A,a,t)$ as the empirical mean of arm a's reward under meta-arm A until t, initially 0.
for $t = 1$ to T **do**
  For each meta-arm A and each arm a, set $\bar{g}(A,a,t) = \min\{\hat{g}(A,a,t-1) + \sqrt{\frac{2\log(T)}{n_{t-1}(A)+1}}, 1\}$
  if Max-min fairness objective is considered **then**
    $\mathbf{p}_t = \text{Oracle}_1((\bar{g}(A,a,t))_{A\in\mathcal{F},a\in\mathcal{N}})$
  else if Total reward maximization subject to fairness constraints is considered **then**
    $\mathbf{p}_t = \text{Oracle}_2((\bar{g}(A,a,t))_{A\in\mathcal{F},a\in\mathcal{N}})$
  end if
  Sample out a meta-arm A based on the categorical distribution of $\mathbf{p}_t$
  Play meta-arm A, observe the rewards of all arms, and update $n_t(A)$ and all $\hat{g}(A,a,t)$, $\forall a \in \mathcal{N}$.
end for

Theorem 1. *For EFP-MAB algorithm, if the max-min fairness objective is considered, $\forall T > 0$, the expected regret $\mathbb{E}[R_T]$ defined in Equation (3) is upper bounded by $O(\sqrt{KT\log(T)})$.*

Remark 2: The most closely related state-of-the-art algorithm to EFP-MAB is the FP-OFU algorithm proposed in [27]. Under the max-min fairness objective, the expected regret of FP-OFU is $O(K\sqrt{T}\log(T))$, EFP-MAB achieves an improvement of $O(\sqrt{K\log(T)})$ for the expected regret. Refer to Section VII-A in Appendix for the detailed Proof of Theorem 1.

B. Performance Analysis under Total Reward Maximization Subject to Fairness Constraints

We next analyze EFP-MAB for total reward maximization subject to fairness constraints.

Theorem 2. *For EFP-MAB algorithm, if the total reward maximization subject to fairness constraints is considered, $\forall T > 0$, the expected regret $\mathbb{E}[R_T]$ defined in Equation (6) is upper bounded by $O(N\sqrt{KT\log(T)})$.*

Remark 3: Under the total reward maximization subject to fairness constraints, the expected regret of FP-OFU is $O(NK\sqrt{T}\log(T))$ [27], EFP-MAB achieves an improvement of $O(\sqrt{K\log(T)})$ for the expected regret. Refer to Section VII-B in Appendix for the detailed Proof of Theorem 2.

C. Analysis of Computational Complexity

As mentioned earlier, the most closely related state-of-the-art algorithm to EFP-MAB is the FP-OFU algorithm proposed in [27]. We highlight the key improvements that EFP-MAB introduces in solving the associated optimization problems compared to FP-OFU. In FP-OFU, both the online **Opt-min** and **Opt-cons** problems are non-convex and non-linear, as their objective functions involve the product of two optimization variables, $\mathbf{p}$ and $\mathbf{g}$. For such optimization problems, only local optima are generally obtainable, as finding global optima remains a challenging open problem in the literature. Additionally, the online **Opt-cons** in FP-OFU includes a complex non-linear constraint stemming from the elliptic confidence interval of $\mathbf{g}$. In contrast, EFP-MAB simplifies these challenges by first computing the upper confidence bound (UCB) of $\mathbf{g}$ before addressing the optimization problems in Oracle_1 and Oracle_2, thereby reducing $\mathbf{p}$ to the sole optimization variable. Moreover, EFP-MAB uses rectangular confidence intervals for $\mathbf{g}$, simplifying the constraint structure. These modifications transform Oracle_1 and Oracle_2 into linear programming problems, enabling the use of standard convex solvers to efficiently obtain global optimal solutions.

To compare their computational complexity numerically, note that both $\mathbf{p}$ (a $1\times K$ vector) and $\mathbf{g}$ (a $K\times N$ matrix) are optimization variables in FP-OFU, there are $K + K\times N$ optimization variables in FP-OFU. As the optimization problems in FP-OFU are non-convex and non-linear, we do not have polynomial-time global optimality guarantees. In contrast, EFP-MAB treats $\mathbf{p}$ as the sole optimization variable, and both Oracle_1 and Oracle_2 in EFP-MAB are linear programming problems with K optimization variables and N constraints, the computational complexity of EFP-MAB is $poly(N,K)$. This distinction makes EFP-MAB significantly more scalable and suitable for scenarios with larger K.

V. Simulation Results

A. Simulation Setup

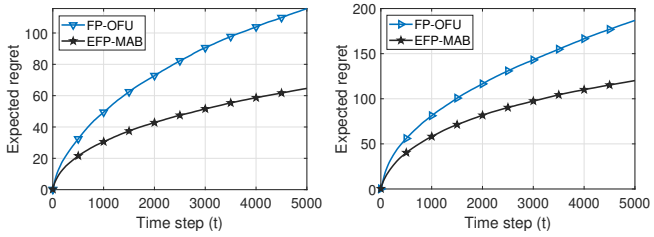
We evaluate the performance of EFP-MAB algorithm under the heterogeneous technology coexistence of LTE, NR (5G), and Wi-Fi, where there are N LTE/NR/Wi-Fi links in total sharing the same channel in the 5 GHz unlicensed band within the same geographical area. Without loss of generality, we focus on downlink transmissions (i.e., LTE/NR BS to UE, Wi-Fi AP to STA), as the proposed approach can similarly be extended to uplink transmissions. To enhance overall spectrum efficiency, we assume concurrent transmissions are permitted in the MAC layer protocol of the coexisting technologies and adopt successive interference cancellation (SIC) to cancel out interference when decoding desired signals.

We simulate 100 randomized LTE/NR/Wi-Fi coexistence topologies, each consisting of $N = 4$ links. The nodes within each topology are uniformly distributed across a $100m\times 100m$ area. Rayleigh channel model is considered. The successful decoding SINR threshold is set to 10 dB. As demonstrated in [27], the FP-OFU algorithm outperforms other baseline methods (e.g., UCB, Maxmin UCB [31], FP-ETC [27]) in terms of regret. Therefore, in this paper, we focus on comparing the performance of our proposed EFP-MAB algorithm against the state-of-the-art FP-OFU algorithm [27].

B. EFP-MAB with Max-min Fairness objective

We show the performance of EFP-MAB algorithm with max-min fairness objective in the following.

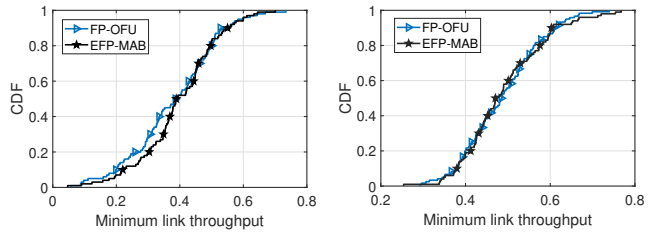
Fig. 2 illustrates the expected regret of EFP-MAB and FP-OFU algorithms for 8 CTSes ($K = 8$) and 10 CTSes ($K = 10$) [1]. These K CTSes are selected as follows: $\text{CTS}_1 = \{\text{link 1}\}$, $\text{CTS}_2 = \{\text{link 2}\}$, $\text{CTS}_3 = \{\text{link 3}\}$, and $\text{CTS}_4 = \{\text{link 4}\}$ are always included; the remaining $K - 4$ CTSes are randomly chosen from the remaining feasible set of CTSes for each topology. The results show that the EFP-MAB algorithm consistently achieves significantly lower regret than FP-OFU in both scenarios. This outcome aligns with theoretical expectations, as the expected regret of EFP-MAB is $O(\sqrt{KT\log(T)})$, as derived in Equation (32), which is asymptotically lower than the expected regret $O(K\sqrt{T}\log(T))$ of FP-OFU.



(a) $K = 8$ (b) $K = 10$

Fig. 2: Expected regret vs time step t.

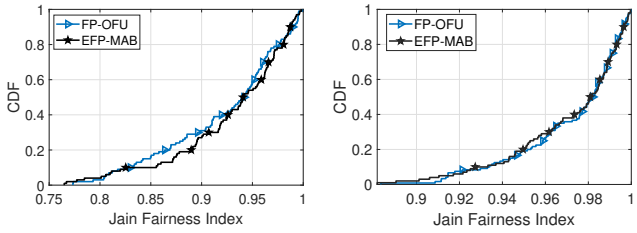
Fig. 3 depicts the cumulative distribution functions (CDFs) of the minimum link throughput for EFP-MAB and FP-OFU under $K = 8$ and $K = 10$, across the 100 random topologies. The results indicate that EFP-MAB slightly outperforms FP-OFU when $K = 8$, while both algorithms perform comparably when $K = 10$. This difference arises because the feasible set for $\mathbf{p}_t$ is smaller for $K = 8$ compared to $K = 10$. Additionally, in FP-OFU, $\mathbf{g}_t$ is also an optimization variable, which complicates the search for the optimal $\mathbf{p}$. Moreover, both algorithms generally achieve higher minimum link throughput as the K value increases.



(a) $K = 8$ (b) $K = 10$

Fig. 3: CDF of minimum link throughput.

The CDFs of Jain fairness index (JFI) [36] for EFP-MAB and FP-OFU under $K = 8$ and $K = 10$ are illustrated in Fig. 4. A larger JFI indicates better link-level fairness. The results show that both algorithms achieve higher JFI values as K increases. This improvement is attributed to the larger feasible set of $\mathbf{p}_t$ at higher K, which facilitates finding the global optimum. Consistent with the observations in Fig. 3, EFP-MAB slightly outperforms FP-OFU when $K = 8$.

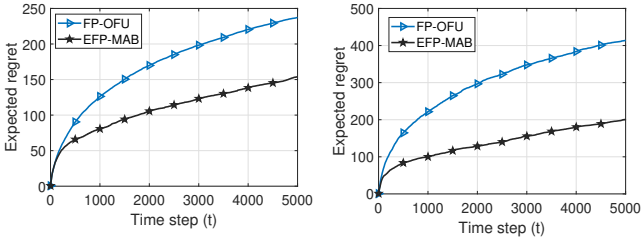


(a) $K = 8$ (b) $K = 10$

Fig. 4: CDF of Jain fairness index (JFI).

C. EFP-MAB with Fairness Constraints

We evaluate the performance of EFP-MAB algorithm with fairness constraints in the following. Fig. 5 presents the expected regret of EFP-MAB and FP-OFU algorithms for $K = 8$ and $K = 10$. We set the minimum selection fraction for each link as 0.3 in the simulation. The results demonstrate that the EFP-MAB algorithm consistently outperforms the baseline FP-OFU in both scenarios. This aligns with theoretical predictions, as the expected regret of EFP-MAB is $O(N\sqrt{KT\log(T)})$, as derived in Equation (44), which is asymptotically lower than the expected regret $O(NK\sqrt{T}\log(T))$ of FP-OFU.



(a) $K = 8$ (b) $K = 10$

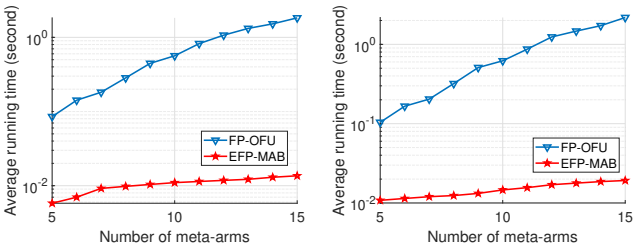
Fig. 5: Expected regret vs time step t.

D. Computational Complexity

We implemented the FP-OFU and EFP-MAB algorithms in MATLAB and conducted simulations on a desktop equipped with an Intel i9-10900K CPU. Fig. 6 displays the average running time per time step for both algorithms under the max-min fairness objective and total reward maximization with fairness constraints. The results show that the average running time of the FP-OFU algorithm increases rapidly with the number of meta-arms. In contrast, the average running time for EFP-MAB grows slowly with the number of meta-arms, consistent with the analysis in Section IV-C. Specifically, the average running time for EFP-MAB remains under 20ms (in the order of 10ms when $K = 15$, whereas it is approximately 2 seconds (in the order of seconds) for FP-OFU. This highlights the significantly better computational efficiency of EFP-MAB compared to FP-OFU. The run time of 20ms can be further

[1]Due to hardware limitation, we were able to complete the FP-OFU simulations within a feasible time frame only for K values up to 10 under 100 LTE/NR/Wi-Fi coexistence topologies.

reduced when the EFP-MAB algorithm is executed on higher-capability hardware in practical scenarios.



(a) Max-min fairness objective (b) Fairness constraints

Fig. 6: Average running time per time step.

VI. Conclusion and Future Work

Spectrum sharing is a key enabler for addressing spectrum shortages caused by the rapid growth in mobile users and data traffic. In this paper, we introduce the efficient fair probabilistic multi-armed bandit (EFP-MAB) algorithm to tackle fairness concerns among distinct networks in the shared spectrum while enhancing the efficiency of resource allocation algorithms. We provide a comprehensive performance analysis, focusing on its expected regret and computational complexity, and conduct extensive simulations in a heterogeneous technology coexistence scenario to evaluate its performance. Both theoretical analysis and simulation results demonstrate that EFP-MAB significantly outperforms the state-of-the-art FP-OFU algorithm in terms of expected regret and efficiency.

As part of our future work, we plan to implement the EFP-MAB algorithm on a real testbed. Additionally, we will explore non-stationary rewards and consider a stochastic number of arms, where an arm may become inactive with a certain probability.

VII. Appendix

A. Proof of Theorem 1

Proof: For a meta-arm $A_i \in \mathcal{F}$ and $a \in \mathcal{N}$, define the clean event $\xi_{i,a} := \{\forall t, \left|\hat{g}(A_i,a,t-1) - g(A_i,a)\right| \leq \sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}\}$ and define $\xi = \cap_{i\in[K],a\in\mathcal{N}}\xi_{i,a}$, representing the intersection of all clean events. By applying the Hoeffding Inequality (as described in Equation (1.5) of [37]), it follows that

$$Pr(\xi_{i,a}) \geq 1 - \frac{2}{T^4}. \tag{9}$$

Using the Union bound, the probability of ξ can be lower bounded as:

$$Pr(\xi) \geq 1 - \sum_{\forall i,\forall a} Pr(\bar{\xi}_{i,a}) \geq 1 - \frac{2NK}{T^4}. \tag{10}$$

For any $\mathbf{p} \in \mathcal{F}_1$ and any arm $a \in \mathcal{N}$, define $h_a(\mathbf{p},\mathbf{g}) = \sum_{i\in[K]}\big(p_i \times g(A_i,a)\big)$, where $\mathbf{g} \in [0,1]^{K\times N}$ is the matrix containing the true mean rewards for all meta-arm and arm combinations. Let $h_a(\mathbf{p},\hat{\mathbf{g}}_{t,a})$ be defined similarly using the empirical means: $h_a(\mathbf{p},\hat{\mathbf{g}}_{t,a}) = \sum_{i\in[K]}\big(p_i \times \hat{g}(A_i,a,t)\big)$. Consequently, the utility functions under true means and empirical means are defined as:

$$U_1(\mathbf{p},\mathbf{g}) = \min_{a\in\mathcal{N}} h_a(\mathbf{p},\mathbf{g}), \tag{11}$$

$$U_1(\mathbf{p},\hat{\mathbf{g}}_t) = \min_{a\in\mathcal{N}} h_a(\mathbf{p},\hat{\mathbf{g}}_{t,a}). \tag{12}$$

The clean event ξ occurs with high probability, as shown in Equation (10). Now, we bound $\left|h_a(\mathbf{p},\mathbf{g}) - h_a(\mathbf{p},\hat{\mathbf{g}}_{t-1,a})\right|$ when the clean event ξ holds,

$$\begin{aligned}
&\left|h_a(\mathbf{p},\mathbf{g}) - h_a(\mathbf{p},\hat{\mathbf{g}}_{t-1,a})\right| \\
&= \left|\sum_{i\in[K]}\big(p_i g(A_i,a)\big) - \sum_{i\in[K]}\big(p_i\hat{g}(A_i,a,t-1)\big)\right| \\
&= \sum_{i\in[K]} p_i\left|g(A_i,a) - \hat{g}(A_i,a,t-1)\right| \\
&\leq \sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}.
\end{aligned} \tag{13}$$

Following Lemma 3 from [27] and setting $X(a) = h_a(\mathbf{p},\hat{\mathbf{g}}_{t-1,a}) - \sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}$ and $Y(a) = h_a(\mathbf{p},\mathbf{g})$, we can conclude that for all $a \in \mathcal{N}$, $X(a) \leq Y(a)$. Therefore, we have:

$$\min_{a\in\mathcal{N}}\{h_a(\mathbf{p},\hat{\mathbf{g}}_{t-1,a}) - \sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}\} \leq \min_{a\in\mathcal{N}}\{h_a(\mathbf{p},\mathbf{g})\}. \tag{14}$$

Since the term $\sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}$ has no dependence on arm a, Equation (14) can be rewritten as:

$$\min_{a\in\mathcal{N}}\{h_a(\mathbf{p},\hat{\mathbf{g}}_{t-1,a})\} - \min_{a\in\mathcal{N}}\{h_a(\mathbf{p},\mathbf{g})\} \leq \sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}, \tag{15}$$

This can be equivalently expressed as:

$$U_1(\mathbf{p},\hat{\mathbf{g}}_{t-1}) - U_1(\mathbf{p},\mathbf{g}) \leq \sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}, \tag{16}$$

where we recall the expressions for $U_1(\mathbf{p},\hat{\mathbf{g}}_{t-1})$ and $U_1(\mathbf{p},\mathbf{g})$ from Equations (11) and (12).

Similarly, applying Lemma 3 from [27] with $X(a) = h_a(\mathbf{p},\mathbf{g})$ and $Y(a) = h_a(\mathbf{p},\hat{\mathbf{g}}_{t-1,a}) + \sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}$, we obtain that for all $a \in \mathcal{N}$, $X(a) \leq Y(a)$. Following the same reasoning, we can derive the following bound:

$$U_1(\mathbf{p},\mathbf{g}) - U_1(\mathbf{p},\hat{\mathbf{g}}_{t-1}) \leq \sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}. \tag{17}$$

Combining Equation (16) and Equation (17), we obtain

$$\left|U_1(\mathbf{p},\mathbf{g}) - U_1(\mathbf{p},\hat{\mathbf{g}}_{t-1})\right| \leq \sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}. \tag{18}$$

Next, we proceed to bound $\left|h_a(\mathbf{p},\bar{\mathbf{g}}_{\mathbf{t},\mathbf{a}}) - h_a(\mathbf{p},\hat{\mathbf{g}}_{t-1,a})\right|$:

$$
\begin{aligned}
&\left|h_a(\mathbf{p}, \bar{\mathbf{g}}_{\mathbf{t}}, \mathbf{a}) - h_a(\mathbf{p}, \hat{\mathbf{g}}_{t-1,a})\right| \\
=&\left|\sum_{i\in[K]} \big(p_i \bar{g}(A_i, a, t)\big) - \sum_{i\in[K]} \big(p_i \hat{g}(A_i, a, t-1)\big)\right| \\
=&\sum_{i\in[K]} p_i \left|\bar{g}(A_i, a, t) - \hat{g}(A_i, a, t-1)\right| \\
\overset{(a)}{\leq}&\sum_{i\in[K]} p_i \sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}},
\end{aligned} \tag{19}
$$

where (a) is due to the relationship between $\bar{g}(A_i, a, t)$ and $\hat{g}(A_i, a, t-1)$ as shown in line 3 of Algorithm 1.

Following the same reasoning from Equation (13) to Equation (17), we obtain

$$
\left|U_1(\mathbf{p}, \bar{\mathbf{g}}_t) - U_1(\mathbf{p}, \hat{\mathbf{g}}_{t-1})\right| \leq \sum_{i\in[K]} p_i \sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}. \tag{20}
$$

Combining Equation (18) and Equation (20), we derive:

$$
\left|U_1(\mathbf{p}, \mathbf{g}) - U_1(\mathbf{p}, \bar{\mathbf{g}}_t)\right| \leq 2\sum_{i\in[K]} p_i \sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}} \tag{21}
$$

Let us define $R_T^{U_1} = \sum_{t=1}^{T}(U_1(\mathbf{p}^*, \mathbf{g}) - U_1(\mathbf{p}_t, \mathbf{g}))$. When the clean event ξ happens, the difference $U_1(\mathbf{p}^*, \mathbf{g}) - U_1(\mathbf{p}_t, \mathbf{g})$ can be bounded as:

$$
\begin{aligned}
&U_1(\mathbf{p}^*, \mathbf{g}) - U_1(\mathbf{p}_t, \mathbf{g}) \\
\overset{(b)}{\leq}&U_1(\mathbf{p}^*, \bar{\mathbf{g}}_t) - U_1(\mathbf{p}_t, \mathbf{g}) \\
\overset{(c)}{\leq}&U_1(\mathbf{p}_t, \bar{\mathbf{g}}_t) - U_1(\mathbf{p}_t, \mathbf{g}) \\
\overset{(d)}{\leq}&2\sum_{i\in[K]} p_{t,i} \sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}},
\end{aligned} \tag{22}
$$

where (b) holds since given $\mathbf{p}^*$, $U_1(\mathbf{p}^*, \mathbf{g}) = \min_{a\in\mathcal{N}} h_a(\mathbf{p}^*, \mathbf{g})$, and $\forall a$, $h_a(\mathbf{p}^*, \mathbf{g})$ is monotonically increasing with respect to $\mathbf{g}$, each element in $\mathbf{g}$ is less than or equal to $\bar{\mathbf{g}}_t$ due to the clean event ξ happening and line 3 of Algorithm 1; (c) holds because $\mathbf{p}_t$ is the optimal solution of Oracle$_1(\bar{\mathbf{g}}_t)$; (d) is due to Equation (21).

Next, we relate R_T to $R_T^{U_1}$ by bounding the two terms in Equation (2). For the first term, applying the Hoeffding's inequality, we obtain that $\forall a \in \mathcal{N}$, with probability at least $1 - \frac{1}{T^2}$,

$$
\left|\sum_{t=1}^{T} r(b_t, a, t) - T\sum_{i\in[K]} p_i^* g(A_i, a)\right| \leq \sqrt{T\log(2T)}. \tag{23}
$$

By applying the union bound, we obtain that with probability at least $1 - \frac{N}{T^2}$,

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

For the second term of Equation (2), applying Azuma's Inequality[2], we obtain that $\forall a \in \mathcal{N}$, with probability at least $1 - \frac{1}{T^2}$,

$$
\left|\sum_{t=1}^{T} r(A_{c_t}, a, t) - \sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i, a)\right| \leq 4\sqrt{T\log(2T)}. \tag{25}
$$

Using Union bound, we obtain that with probability at least $1 - \frac{N}{T^2}$,

$$
\left|\min_{a\in\mathcal{N}}\sum_{t=1}^{T} r(A_{c_t}, a, t) - \min_{a\in\mathcal{N}}\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i, a)\right| \leq 4\sqrt{T\log(2T)}. \tag{26}
$$

Define the event $\eta :=$ {Equation (24) and Equation (26) hold}. The complement of η is denoted as $\bar{\eta}$. Applying the union bound, we obtain that event η occurs with probability at least $1 - \frac{2N}{T^2}$. Recall that event ξ happens with high probability. When the events ξ and η both occur, the regret R_T can be bounded as:

$$
\begin{aligned}
R_T =& \min_{a\in\mathcal{N}}\sum_{t=1}^{T} r(b_t, a, t) - \min_{a\in\mathcal{N}}\sum_{t=1}^{T} r(c_t, a, t) \\
\leq& T\min_{a\in\mathcal{N}}\sum_{i\in[K]} p_i^* g(A_i, a) - \min_{a\in\mathcal{N}}\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i, a) \\
&+ \sqrt{T\log(2T)} + 4\sqrt{T\log(2T)} \\
\overset{(e)}{\leq}& Tf(\mathbf{p}^*) - f(\sum_{t=1}^{T}\mathbf{p_t}) + O(\sqrt{T\log(T)}) \\
\overset{(j)}{\leq}& Tf(\mathbf{p}^*) - \sum_{t=1}^{T} f(\mathbf{p_t}) + O(\sqrt{T\log(T)}) \\
=& R_T^{U_1} + O(\sqrt{T\log(T)}),
\end{aligned} \tag{27}
$$

where $f(\mathbf{p}) = \min_{a\in\mathcal{N}}\{\sum_{i\in[K]}\big(p_i \times g(A_i, a)\big)\}$ in (e), (j) follows from Lemma 4 of [27].

Next, we bound $\mathbb{E}[R_T^{U_1}\mathbb{I}(\xi \cap \eta)]$. From Equation (22) and the definition of $R_T^{U_1}$, we have

$$
\begin{aligned}
\mathbb{E}[R_T^{U_1}\mathbb{I}(\xi\cap\eta)] &= \mathbb{E}[\mathbb{I}(\xi\cap\eta)\sum_{t=1}^{T}(U_1(\mathbf{p}^*, \mathbf{g}) - U_1(\mathbf{p}_t, \mathbf{g}))] \\
&\leq \mathbb{E}[\sum_{t=1}^{T}(U_1(\mathbf{p}^*, \mathbf{g}) - U_1(\mathbf{p}_t, \mathbf{g}))] \\
&\leq 2\mathbb{E}[\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i}\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}].
\end{aligned} \tag{28}
$$

We now show how to bound the term $\mathbb{E}[\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i}\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}]$. Note that $n_t(A_i) - n_{t-1}(A_i) = \mathbb{I}(c_t = A_i)$. Also, $p_{t,i}$ represents the probability of selecting meta-arm A_i at time step t. Therefore, we have

$$
\mathbb{E}[\mathbb{I}(c_t = A_i)|\mathcal{H}_{t-1}] = p_{t,i}, \tag{29}
$$

where $\mathcal{H}_{t-1} = \{\mathbf{p}_{t'}, c_{t'}, (r(A_{c_{t'}}, a, t'))_{a\in\mathcal{N}}\}_{t'=1}^{t-1}$.

[2]The application of Azuma's inequality to the second term of Equation (2) is justified by the reasoning provided in the Proof of Theorem 4 in Ref. [27]

Plugging Equation (29) into Equation (28), we have

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i}\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}] \\
&\leq \mathbb{E}[\sum_{t=1}^{T}\sum_{i\in[K]} \mathbb{E}[\mathbb{I}(c_t=A_i)|\mathcal{H}_{t-1}]\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}] \\
&= \sum_{t=1}^{T}\sum_{i\in[K]} \mathbb{E}\big[\mathbb{E}[\mathbb{I}(c_t=A_i)|\mathcal{H}_{t-1}]\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}] \\
&\overset{(l)}{=} \sum_{t=1}^{T}\sum_{i\in[K]} \mathbb{E}\big[\mathbb{E}[\mathbb{I}(c_t=A_i)\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}|\mathcal{H}_{t-1}]\big] \\
&= \sum_{t=1}^{T}\sum_{i\in[K]} \mathbb{E}[\mathbb{I}(c_t=A_i)\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}] \\
&= \sum_{i\in[K]}\sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(c_t=A_i)\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}] \\
&= \mathbb{E}[\sum_{i\in[K]}\sum_{s=0}^{n_T(A_i)}\sqrt{\frac{2\log(T)}{s+1}}] \\
&\leq 2\mathbb{E}[\sqrt{2\log(T)}\sum_{i\in[K]}\sqrt{n_T(A_i)}] \\
&\leq \mathbb{E}[2\sqrt{2\log(T)}\sqrt{KT}] = O(\sqrt{\log(T)KT}),
\end{aligned} \tag{30}$$

where (l) is because that given $\mathcal{H}_{t-1}$, $\forall i \in [K], n_{t-1}(A_i)$ is fixed and not a random variable.

Therefore, from Equations (28) and (30), we can conclude that $\mathbb{E}[R_T^{U_1}\mathbb{I}(\xi\cap\eta)] \leq O(\sqrt{\log(T)KT})$.

Using Equation (27), we get:

$$\begin{aligned}
\mathbb{E}\left[R_T\mathbb{I}(\xi\cap\eta)\right] &\leq \mathbb{E}\left[R_T^{U_1}\mathbb{I}(\xi\cap\eta)\right] + O(\sqrt{T\log(T)}) \\
&\leq O(\sqrt{\log(T)KT}) + O(\sqrt{T\log(T)}) \\
&\leq O(\sqrt{\log(T)KT}).
\end{aligned} \tag{31}$$

Applying the Union bound, we have $Pr(\xi\cap\eta) \geq 1 - \frac{2N}{T^2} - \frac{2NK}{T^4}$. By averaging over all events, $\mathbb{E}[R_T]$ can be upper bounded as:

$$\begin{aligned}
\mathbb{E}\left[R_T\right] &\leq \mathbb{E}[R_T\mathbb{I}(\xi\cap\eta)] + \mathbb{E}[R_T\mathbb{I}(\overline{\xi\cap\eta})] \\
&\leq O(\sqrt{KT\log(T)}) + T\cdot(\frac{2N}{T^2}+\frac{2NK}{T^4}) \\
&\leq O(\sqrt{KT\log(T)}).
\end{aligned} \tag{32}$$

B. Proof of Theorem 2

Proof: The proof of Theorem 2 follows a structure similar to that of Theorem 1. To avoid redundancy, we focus on highlighting the differences.

For all $i \in [K]$ and $a \in \mathcal{N}$, we define the clean event $\xi_{i,a}$, ξ, and function $h_a(\mathbf{p},\mathbf{g})$ analogously to Theorem 1. Equation (10) remains valid. Given the difference in the objective function in Theorem 1, we define the following utility functions tailored for total reward maximization under fairness constraints:

$$U_2(\mathbf{p},\mathbf{g}) = \sum_{a\in\mathcal{N}} h_a(\mathbf{p},\mathbf{g}), \mathbf{p}\in\mathcal{F}_2, \tag{33}$$

$$U_2(\mathbf{p},\hat{\mathbf{g}}_t) = \sum_{a\in\mathcal{N}} h_a(\mathbf{p},\hat{\mathbf{g}}_{t,a}), \mathbf{p}\in\mathcal{F}_2. \tag{34}$$

Since $h_a(\mathbf{p},\mathbf{g})$ is linear with respect to $\mathbf{p}$ and $\mathbf{g}$, Equation (13) holds. For any $\mathbf{p} \in \mathcal{F}_2$, and noting that the term $\sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}$ is independent of arm a in Equation (13), we derive the following bound:

$$\left|U_2(\mathbf{p},\mathbf{g}) - U_2(\mathbf{p},\hat{\mathbf{g}}_{t-1})\right| \leq N\sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}. \tag{35}$$

Similarly, Equation (19) still applies, leading to:

$$\left|U_2(\mathbf{p},\bar{\mathbf{g}}_t) - U_2(\mathbf{p},\hat{\mathbf{g}}_{t-1})\right| \leq N\sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}. \tag{36}$$

By combining Equation (35) and Equation (36), we obtain:

$$\left|U_2(\mathbf{p},\mathbf{g}) - U_2(\mathbf{p},\bar{\mathbf{g}}_t)\right| \leq 2N\sum_{i\in[K]} p_i\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}. \tag{37}$$

Using similar reasoning to that in Equation (22), when the clean event ξ happens, we can further derive

$$U_2(\mathbf{p}^*,\mathbf{g}) - U_2(\mathbf{p}_t,\mathbf{g}) \leq 2N\sum_{i\in[K]} p_{t,i}\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}. \tag{38}$$

Next, we proceed to bound the two terms in Equation (5). For the first term, applying the Hoeffding's inequality, we find that $\forall a \in \mathcal{N}$, with probability at least $1-\frac{1}{T^2}$, Equation (23) holds. By applying the Union bound, we obtain that at least with probability $1-\frac{N}{T^2}$,

$$\left|\sum_{a\in\mathcal{N}}\sum_{t=1}^{T} r(b_t,a,t) - T\sum_{a\in\mathcal{N}}\sum_{i\in[K]} p_i^* g(A_i,a)\right| \leq N\sqrt{T\log(2T)}. \tag{39}$$

For the second term of Equation (5), applying Azuma's Inequality, we obtain that $\forall a \in \mathcal{N}$, with probability at least $1-\frac{1}{T^2}$, Equation (25) holds. Using Union bound, we further obtain that with probability at least $1-\frac{N}{T^2}$,

$$\begin{aligned}
&\left|\sum_{a\in\mathcal{N}}\sum_{t=1}^{T} r(A_{c_t},a,t) - \sum_{a\in\mathcal{N}}\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i,a)\right| \\
&\leq 4N\sqrt{T\log(2T)}.
\end{aligned} \tag{40}$$

Define the event $\eta :=$ {Equation (39) and Equation (40) hold}. The complement of η is denoted as $\bar{\eta}$. Applying the union bound, we conclude that event η occurs with probability at least $1-\frac{2N}{T^2}$. Recall that event ξ happens with high probability (at least $1-\frac{2NK}{T^4}$). When the events ξ and η both occur, similary to Equation (27), the regret R_T in Equation (5) can be bounded as:

$$\begin{aligned}
R_T &= \sum_{a\in\mathcal{N}}\sum_{t=1}^{T} r(b_t,a,t) - \sum_{a\in\mathcal{N}}\sum_{t=1}^{T} r(c_t,a,t) \\
&\leq T\sum_{a\in\mathcal{N}}\sum_{i\in[K]} p_i^* g(A_i,a) - \sum_{a\in\mathcal{N}}\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i} g(A_i,a) \\
&+ N\sqrt{T\log(2T)} + 4N\sqrt{T\log(2T)} \\
&\leq R_T^{U_2} + O(N\sqrt{T\log(T)}),
\end{aligned} \tag{41}$$

where $R_T^{U_2} = \sum_{t=1}^{T}(U_2(\mathbf{p}^*,\mathbf{g}) - U_2(\mathbf{p}_t,\mathbf{g}))$.

Next, we bound $\mathbb{E}[R_T^{U_2}\mathbb{I}(\xi \cap \eta)]$. From Equation (38) and the definition of $R_T^{U_2}$, we have

$$\begin{aligned}\mathbb{E}[R_T^{U_2}\mathbb{I}(\xi \cap \eta)] &= \mathbb{E}[\mathbb{I}(\xi \cap \eta)\sum_{t=1}^{T}(U_2(\mathbf{p}^*, \mathbf{g}) - U_2(\mathbf{p}_t, \mathbf{g}))] \\ &\leq \mathbb{E}[\sum_{t=1}^{T}(U_2(\mathbf{p}^*, \mathbf{g}) - U_2(\mathbf{p}_t, \mathbf{g}))] \\ &\leq 2N\mathbb{E}[\sum_{t=1}^{T}\sum_{i\in[K]} p_{t,i}\sqrt{\frac{2\log(T)}{n_{t-1}(A_i)+1}}].\end{aligned} \tag{42}$$

By Equation (30), it follows that:

$$\mathbb{E}[R_T^{U_2}\mathbb{I}(\xi \cap \eta)] \leq O(N\sqrt{\log(T)KT}). \tag{43}$$

Following a similar reasoning in Equation (31) and Equation (32), $\mathbb{E}[R_T]$ in Equation (6) can be upper bounded as:

$$\mathbb{E}[R_T] \leq O(N\sqrt{\log(T)KT}). \tag{44}$$

REFERENCES

[1] 3GPP, "Scenarios and requirements for small cell enhancements for e-utra and e-utran (release 12)," TR 36.932 (V12.1.0), 2013.

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

[3] 3GPP, "Study on NR-based access to unlicensed spectrum," Standard (TR) 36.889, V16.0.0, 2018.

[4] B. Qian, H. Zhou, T. Ma, K. Yu, Q. Yu, and X. Shen, "Multi-operator spectrum sharing for massive IoT coexisting in 5G/B5G wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 881–895, 2020.

[5] H. Pirayesh, P. K. Sangdeh, and H. Zeng, "Coexistence of Wi-Fi and IoT communications in wlans," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7495–7505, 2020.

[6] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, "A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 668–695, 2021.

[7] 3GPP, "Technical specification group radio access network; study on new radio access technology: Radio access architecture and interfaces," TR 38.801 version V14.0.0,, 2022.

[8] L. Zhang, M. Xiao, G. Wu, M. Alam, Y.-C. Liang, and S. Li, "A survey of advanced techniques for spectrum sharing in 5G networks," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 44–51, 2017.

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

[10] Y. Jian, C.-F. Shih, B. Krishnaswamy, and R. Sivakumar, "Coexistence of Wi-Fi and LAA-LTE: Experimental evaluation, analysis and insights," in *IEEE international conference on communication workshop (ICCW)*, 2015, pp. 2325–2331.

[11] A. M. Cavalcante, E. Almeida, R. D. Vieira, S. Choudhury, E. Tuomaala, K. Doppler, F. Chaves, R. C. Paiva, and F. Abinader, "Performance evaluation of LTE and Wi-Fi coexistence in unlicensed bands," in *IEEE 77th Vehicular Technology Conference (VTC Spring)*, 2013, pp. 1–6.

[12] S. Yun and L. Qiu, "Supporting WiFi and LTE co-existence," in *IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 810–818.

[13] Z. Guo, M. Li, and M. Krunz, "Exploiting successive interference cancellation for spectrum sharing over unlicensed bands," *IEEE Transactions on Mobile Computing*, vol. 23, no. 3, pp. 2438–2455, 2023.

[14] N. Shlezinger, R. Fu, and Y. C. Eldar, "DeepSIC: Deep soft interference cancellation for multiuser MIMO detection," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1349–1362, 2020.

[15] L. Valentini, M. Chiani, and E. Paolini, "Interference cancellation algorithms for grant-free multiple access with massive mimo," *IEEE Transactions on Communications*, vol. 71, no. 8, pp. 4665–4677, 2023.

[16] Y. Yan, P. Yang, X.-Y. Li, Y. Zhang, J. Lu, L. You, J. Wang, J. Han, and Y. Xiong, "Wizbee: Wise zigbee coexistence via interference cancellation with single antenna," *IEEE Transactions on Mobile Computing*, vol. 14, no. 12, pp. 2590–2603, 2014.

[17] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wen, and M. Tao, "Resource allocation in spectrum-sharing OFDMA femtocells with heterogeneous services," *IEEE Transactions on Communications*, vol. 62, no. 7, pp. 2366–2377, 2014.

[18] M. Han, S. Khairy, L. X. Cai, Y. Cheng, and R. Zhang, "Reinforcement learning for efficient and fair coexistence between LTE-LAA and Wi-Fi," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8764–8776, 2020.

[19] M. A. Raza, M. Abolhasan, J. Lipman, N. Shariati, W. Ni, and A. Jamalipour, "Multi-agent multi-armed bandit learning for grant-free access in ultra-dense iot networks," *IEEE Transactions on Cognitive Communications and Networking*, 2024.

[20] Y. Huang, S. Li, C. Li, Y. T. Hou, and W. Lou, "A deep-reinforcement-learning-based approach to dynamic eMBB/URLLC multiplexing in 5G NR," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6439–6456, 2020.

[21] A. Kumar, A. Roy, and R. Bhattacharjee, "Actively adaptive multi-armed bandit based beam tracking for mmwave MIMO systems," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2024, pp. 1–6.

[22] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4585–4600, 2021.

[23] 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.

[24] 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.

[25] Z. Guo, C. Zhang, M. Li, and M. Krunz, "Fair coexistence of heterogeneous networks: A novel probabilistic multi-armed bandit approach," in *IEEE 21st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2023, pp. 1–8.

[26] R. Bajracharya, R. Shrestha, and H. Jung, "Bandit approach for fair and efficient coexistence of NR-U in unlicensed bands," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 4, pp. 5208–5223, 2022.

[27] Z. Guo, C. Zhang, M. Li, and M. Krunz, "Fair probabilistic multi-armed bandit with applications to network optimization," *IEEE Transactions on Machine Learning in Communications and Networking*, 2024.

[28] L. Cheng, B. Henty, F. Bai, and D. D. Stancil, "Doppler spread and coherence time of rural and highway vehicle-to-vehicle channels at 5.9 ghz," in *IEEE Global Telecommunications Conference*, 2008, pp. 1–6.

[29] A. Pratap, R. Misra, and S. K. Das, "Maximizing fairness for resource allocation in heterogeneous 5G networks," *IEEE transactions on mobile computing*, vol. 20, no. 2, pp. 603–619, 2019.

[30] F. Tang, Y. Zhou, and N. Kato, "Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5G HetNet," *IEEE Journal on selected areas in communications*, vol. 38, no. 12, pp. 2773–2782, 2020.

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

[32] 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.

[33] W. Chen, Y. Wang, Y. Yuan, and Q. Wang, "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms," *Journal of Machine Learning Research*, vol. 17, no. 50, pp. 1–33, 2016.

[34] Y. Kim, J. Jang, and H. J. Yang, "Distributed resource allocation and user association for max-min fairness in hetnets," *IEEE Transactions on Vehicular Technology*, 2023.

[35] C. Xu, S. Chen, J. Xu, W. Shen, X. Zhang, G. Wang, and Z. Dong, "P-mmf: Provider max-min fairness re-ranking in recommender system," in *Proceedings of the ACM Web Conference*, 2023, pp. 3701–3711.

[36] Y. He, X. Gang, and Y. Gao, "Intelligent decentralized multiple access via multi-agent deep reinforcement learning," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2024, pp. 1–6.

[37] A. Slivkins *et al.*, "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.