Master's Thesis

Low-Complexity Learning for Dynamic Spectrum Access in Multi-User Multi-Channel Networks

Sunjung Kang

Department of Computer Science and Engineering

Graduate School of UNIST

2018

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Low-Complexity Learning for Dynamic Spectrum Access in Multi-User Multi-Channel Networks

Sunjung Kang

Department of Computer Science and Engineering

Graduate School of UNIST

Low-Complexity Learning for Dynamic Spectrum Access in Multi-User Multi-Channel Networks

A thesis
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Master of Science

Sunjung Kang

12. 22. 2017

Approved by

Changhee Joo

Advisor

Changhee Joo

Low-Complexity Learning for Dynamic Spectrum Access in Multi-User Multi-Channel Networks

Sunjung Kang

This certifies that the thesis of Sunjung Kang is approved.

12. 22. 2017

Signature

Changhee Joo

Advisor: Changhee Joo

Signature

Committee Member: Kyunghan Lee

Signature

Jaesik Choi

Committee Member: Jaesik Choi

Abstract

In cognitive radio Networks (CRNs), dynamic spectrum access allows (unlicensed) users to identify and access unused channels opportunistically, thus improves spectrum utility. In this paper, we address the user-channel allocation problem in multi-user multi-channel CRNs without a prior knowledge of channel statistics. A reward of a channel is stochastic with unknown distribution, and statistically different for each user. Each user either explores a channel to learn the channel statistics, or exploits the channel with the highest expected reward based on information collected so far. Further, a channel should be accessed exclusively by one user at a time due to a collision. Using multi-armed bandit framework, we develop provably efficient solutions whose computational complexities are linear to the number of users and channels.

Contents

List of Figures

I Introduction

Since license-based spectrum management has suffered from low spectrum utilization, cognitive radio networks (CRNs) have attracted much attention as a promising solution to current spectrum inefficiency [1]. In CRNs, unlicensed users (or secondary users) can access unused channels that are licensed to primary users. Dynamic spectrum access allows (secondary) users to identify idle channels and use them opportunistically [2, 3].

We consider multi-user multi-channel CRNs where channels are orthogonal and independent of each other. Characteristic of a channel is represented by a reward, i.e., a good channel implies a high expected reward[1]. A reward of a channel is stochastic with unknown distribution, and statistically different for each user. We assume a slotted-time system where each user can access at most one channel at a time slot. Although the channel information is unknown to users, users can learn from their experiences. Every time each user either explores a channel to estimate its expected reward value, or exploits the channel with the highest expected reward based on information so far. Hence, a user faces the well-known exploration-exploitation tradeoff.

This can be formulated as a class of multi-armed bandit (MAB) problems [4–7], which are a framework for sequential decision problems considering the exploration-exploitation tradeoff. In single-user MAB problems, a player (or a user) chooses an arm (i.e., a channel) at each time slot, and receives a reward from the chosen arm. An MAB policy decides which arm to play to get the best (total) reward given observations in previous time slots. The performance metric for evaluating a policy is *regret*, which is the accumulated difference between the highest expected reward and that achieved by the policy. In stochastic MAB problems, the rewards are assumed to be an i.i.d. process with unknown distribution and bounded support. The authors of [4] have shown that the regret of stochastic MAB grows at least logarithmically over time. In [5], the authors have proposed an index-based policy for stochastic MAB using upper confidence bound (UCB) called UCB1, and shown that the expected regret of UCB1 algorithm grows at most logarithmically. On the other hand, adversarial MAB problems consider non-stochastic rewards. In [6], the authors have proposed a policy for adversarial MAB called EXP3 of which regret is sub-linear. In [7], the authors have suggested that decision making problems in CR networks can be formulated using the MAB framework.

In multi-user scenarios where multiple users access channels at the same time, a channel should be accessed exclusively by one user at a time due to a collision. This multi-user scenario can be formulated as combinatorial MAB problems [8–16], where the total reward received by playing multiple arms is either the sum of rewards from played arms (linear rewards) or a function of reward vector (non-linear rewards). In [9], a combinatorial MAB problem with non-linear rewards is studied and applied to several applications such as online advertising. In [10], the authors consider a combinatorial MAB problem with linear rewards and apply it to applications such as maximum weighted matching and shortest path. In this paper, we are interested in a

[1]For an example, a reward can be signal-to-noise rate or the bandwidth of the channel.

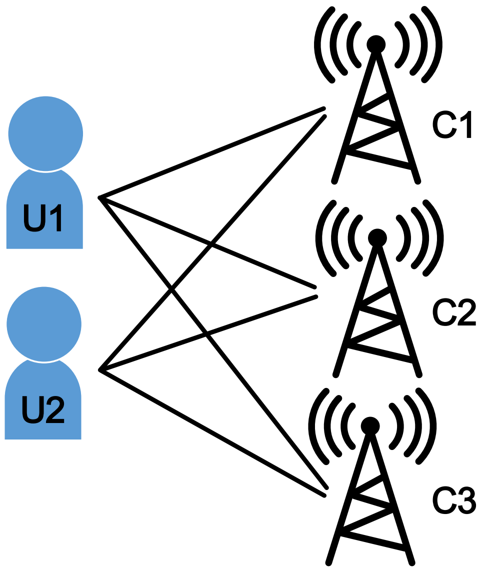
combinatorial (stochastic) MAB problem with linear rewards in multi-user scenarios.

The authors of [11–13] have proposed distributed solutions to MAB problem when the reward from an arm is statistically identical for all the players. They showed that the regret grows logarithmically over time under the proposed policies. For the scenarios where the reward from an arm is statistically different for each player, the problem can be modeled as a weighted bipartite graph with two disjoint sets of players and arms, and the objective is to find an optimal matching (i.e., a maximum weighted matching) with expected reward as weight on edge (player, arm) [14–16]. In this case, the regret of a policy is defined as accumulated total reward achieved by playing an optimal matching minus that achieved by the policy. In [14], the authors have proposed a centralized algorithm, under which a central agent finds a maximum weighted matching with UCB indices at each time using Hungarian algorithm, whose computational complexity is $O(NK(N+K)^3)$ where N and K are the number of players and arms, respectively. In [15], the authors have proposed a decentralized algorithm, under which players participate to the Bertsekas auction algorithm whenever it needs to recompute a matching. It converges to an optimal matching with UCB index with convergence time of $O(N^2 \cdot \max_{i,k} \mu_{i,k}/\epsilon)$, where $\mu_{i,k}$ is the expected reward of arm k for user i and $\epsilon > 0$. Although the policies of [14, 15] achieve logarithmic growth of the regret, they have high-order computational complexity to find the maximum weighted matching. In [16], the authors are interested in finding a stable and orthogonal matching rather than an optimal matching. Although the proposed distributed algorithm has low computational complexity $O(K)$, it does not guarantee the logarithmic growth of the expected regret.

In this paper, we study the multi-user MAB problem where reward statistics are different for each user-channel pair. Each user has no prior knowledge about channel rewards, and estimates the mean reward of each channel by exploring it. A channel can be accessed by at most one user at a time, otherwise a collision occurs and none gets reward for the channel. The procedure of our algorithms are motivated by [17, 18] in that a time slot is divided into a scheduling slot to control collisions and a transmission slot to access the chosen channels. We develop low-complexity learning algorithms for opportunistic spectrum access in multi-user multi-channel cognitive radio networks. To the best of our knowledge, these are the first algorithms that have linear complexity and achieve asymptotic optimality. Our contribution can be summarized as follows.

- We develop linear-complexity solutions to multi-user multi-armed bandit problems.
- We show that our proposed algorithms achieve logarithmic growth of the total expected regret with respect to time t.
- We verify the performance of our algorithms through simulations.

The rest of paper is organized as follows. In Section II, we describe the system model and problem formulation. In Sections III, we propose low-complexity learning algorithm which can



	C1	C2	C3
U1	0.6	0.2	0.1
U2	0.8	0.3	0.5

Figure 1: System model with complete bipartite graph. The maximum weighted matching is marked by circles.

be applied not only to bipartite graph but also to general graph, and evaluate its performance. In Section IV, we propose another low-complexity algorithm which takes the advantage of a property of bipartite graph, thus can further improve the performance. In Section V, we verify our results through simulations, and in Section VI, we conclude our work.

II System Model

We consider a cognitive radio network of N (secondary) users and K orthogonal channels with $K \geq N$. We assume a slotted-time system, where each user can access at most one channel in a time slot. If more than one user accesses the same channel at the same time, then all the conflicting users receive no reward from that channel due to a collision. At time slot t, if user i accesses channel k exclusively, then it receives a reward (e.g., SNR or bandwidth) denoted by $X_{i,k}(t)$, which is a random variable that is i.i.d. across time and has an arbitrary distribution with bounded support. Without loss of generality, we assume that $X_{i,k}(t)$ lies in between 0 and 1 with a mean $\mu_{i,k}$. We assume that each user has no priori knowledge of $X_{i,k}(t)$, and can only observe the returned reward. Let $Z_{i,k}(t)$ denote the actual reward that user i receives from channel k at time t. If user i accesses channel k at time slot t without a collision, then $Z_{i,k}(t) = X_{i,k}(t)$, and otherwise $Z_{i,k}(t) = 0$.

Let $\mathcal{K} = \{1, ..., K\}$ denote the set of channels (or equivalently the set of actions of users), and $x_i(t) \in \mathcal{K}$ denote an action of user i at time slot t, i.e., user i accesses channel $x_i(t)$ at time slot t. We denote its vector $\mathbf{x}(t)$ as *schedule* at time t. Then, the history of users i by time slot t is $\mathcal{H}_i(t) = \{(x_i(1), Z_{i,x_i(1)}(1)), ..., (x_i(t), Z_{i,x_i(t)}(t))\}$ with $\mathcal{H}_i(0) = \emptyset$. A policy $\pi_i = (\pi_i(t))_{t=1}^{\infty}$ for user i is a sequence of maps $\pi_i(t) : \mathcal{H}_i(t-1) \rightarrow \mathcal{K}$ that specifies the channel to access at time slot t given the history seen by the user i. Let $\mathcal{M}$ be the set of feasible schedules such that $\mathcal{M} := \{\mathbf{a} = (a_1, ..., a_N) : a_i \in \mathcal{K}, a_i \neq a_j \text{ for } i \neq j\}$, which is equivalent to the set of all (maximal) matchings in bipartite graph $\mathcal{G} = (\mathcal{N} \cup \mathcal{K}, E)$, where $\mathcal{N}$ and $\mathcal{K}$ are the sets of users and channels, respectively, and E is the set of edges (i, k) for all $i \in \mathcal{N}$ and $k \in \mathcal{K}$. Let $\mathbf{a}^*$ denote an *optimal matching* (i.e., an maximum weighted matching in $\mathcal{G}$) with expected rewards $\mu_{i,k}$ as weights on edges such that

$$\mathbf{a}^* \in \arg\max_{\mathbf{a}\in\mathcal{M}} \sum_{i=1}^{N} \mu_{i,a_i}. \tag{1}$$

Fig. 1 illustrates an example of our model with complete bipartite graph $\mathcal{G} = (\mathcal{N} \cup \mathcal{K}, E)$. There are two users U1 and U2 ($N = 2$), and three channels C1, C2, and C3 ($K = 3$). The matrix shows the expected rewards of each user-channel pair, and the optimal matching is $\mathbf{a}^* = (a_1^*, a_2^*) = (1, 3)$.

Since $\mu_{i,k}$ are unknown parameters, a policy π cannot achieve the optimal performance every time. We consider a regret, which is the difference between the total reward from an optimal matching and that from the non-optimal matching. Let $\mathcal{R}_\pi(T)$ denote the expected total regret by time slot T under policy π:

$$\mathcal{R}_\pi(T) := T \sum_{i=1}^{N} \mu_{i,a_i^*} - \sum_{\mathbf{a}\in\mathcal{M}} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}\left[X_{i,a_i}(t)\mathbb{I}\{\mathbf{x}(t) = \mathbf{a}\}\right], \tag{2}$$

where $\mathbb{I}\{\cdot\}$ is an indicator function which is 1 if the event in $\{\cdot\}$ is true, and 0, otherwise. The objective is to minimize the expected total regret. It is known that the logarithmic growth of expected regret with respect to time is asymptotically optimal [8].

III Uniform Sampling Algorithm

In this section, we develop low-complexity learning algorithm which can be applied to combinatorial MAB with more general graph as well as bipartite graph. We first describe our low-complexity scheme, and then we evaluate the performance of our proposed scheme and show that it is asymptotically optimal.

3.1 Uniform sampling algorithm

We assume that reward $X_{i,k}(t) \in [0, 1]$ of channel k to user i is a normalized i.i.d. random process. Initially, mean reward $\mu_{i,k}$ is unknown but user i can learn the mean reward of channel k by empirically trying the channel and observing the returned rewards. Let $\hat{\mu}_{i,k}(t)$ denote the empirical mean of returned rewards for (user i, channel k) pair by time slot t, and let $\hat{\tau}_{i,k}(t)$ denote the number of times that user i is successfully matched with channel k by time slot t. If $\hat{\tau}_{i,k}(t) \to \infty$ as $t \to \infty$, then $\hat{\mu}_{i,k} \to \mu_{i,k}$ from the law of large numbers. Let $\mathbf{x}(t) \in \mathcal{M}$ denote the schedule at time slot t, where $x_i(t)$ indicates the channel that is matched with user i. At

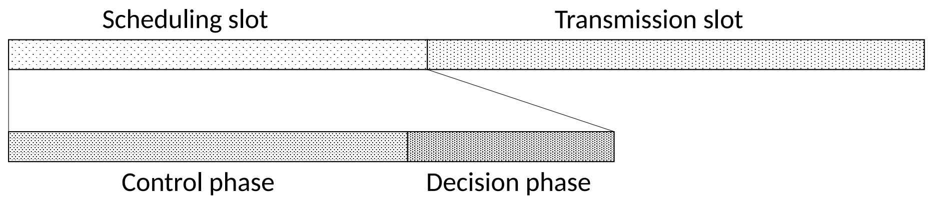


Figure 2: Structure of a time-slot.

the end of time slot t, user i updates $\hat{\mu}_{i,k}(t)$ and $\hat{\tau}_{i,k}(t)$ for channel $k = x_i(t)$ based on returned reward $X_{i,k}(t)$ as

$$\hat{\mu}_{i,k}(t) = \begin{cases} \frac{\hat{\mu}_{i,k}(t-1)\hat{\tau}_{i,k}(t-1)+X_{i,k}(t)}{\hat{\tau}_{i,k}(t-1)+1}, & \text{for } k = x_i(t) \\ \hat{\mu}_{i,k}(t-1), & \text{for } k \neq x_i(t). \end{cases} \tag{3}$$

$$\hat{\tau}_{i,k}(t) = \begin{cases} \hat{\tau}_{i,k}(t-1)+1, & \text{for } k = x_i(t) \\ \hat{\tau}_{i,k}(t-1), & \text{for } k \neq x_i(t). \end{cases} \tag{4}$$

For user i's channel k, we assign an UCB index

$$I_{i,k}(t) := \hat{\mu}_{i,k}(t-1) + \sqrt{\frac{(N+1)\log t}{[\hat{\tau}_{i,k}(t-1)]^+}}, \tag{5}$$

where $[\cdot]^+ = \max\{1, \cdot\}$. It is known that, under single user scenario ($N = 1$), if the user plays the channel with the highest value of UCB index at each time slot, the regret grows logarithmically with respect to time [5]. Under multi-user scenarios, finding the maximum weighted matching with UCB indices at each time slot achieves asymptotic optimality, which, however, has high-order computational complexity [14]. We tackle the problem by developing a linear-complexity algorithm that guarantees the logarithmic growth of the regret.

We start with the description of time structure which is motivated by [17,18]. A time slot is divided into a *scheduling slot* and a *transmission slot*, and the scheduling slot is further divided into a *control phase* and a *decision phase* as shown in Fig. 2. Now we explain our Uniform sampling algorithm (also see Algorithm 1).

- **In the control phase** (lines 1 of Algorithm 1), we select a matching $\mathbf{m}(t) \in \mathcal{M}$ uniformly at random, which is called as a *candidate matching*.

- **In the decision phase** (line 2-3 of Algorithm 1), we compute the total sum of chosen UCB indices from candidate matching $\mathbf{m}(t)$ and that of the previous schedule $\mathbf{x}(t-1)$, and select the one with higher value as new schedule $\mathbf{x}(t)$. Let $V(\mathbf{a}; \mathbf{I}(t)) := \sum_{i=1}^{N} I_{i,a_i}(t)$, which is a value function that evaluates matchings through the index value. Then, the schedule $\mathbf{x}(t)$ can be written as

$$\mathbf{x}(t) \in \underset{\mathbf{a}\in\{\mathbf{m}(t),\mathbf{x}(t-1)\}}{\arg\max} V(\mathbf{a}; \mathbf{I}(t)).$$

- **During the transmission slot,** each user i accesses channel k if $x_i(t) = k$ and gets reward $X_{i,k}(t)$. Then it updates $\hat{\mu}_{i,k}(t)$ and $\hat{\tau}_{i,k}(t)$ according to (3) and (4).

Algorithm 1 Uniform sampling.

At the beginning of each time slot t

Select $m(t) \in \mathcal{M}$ uniformly at random

Calculate $I_{i,m_i(t)}(t)$ and $I_{i,x_i(t-1)}(t)$ for all i

$\mathbf{x}(t) \in \arg\max_{\mathbf{a}\in\{\mathbf{m}(t),\mathbf{x}(t-1)\}} V(\mathbf{a}; \mathbf{I}(t))$

/* make transmissions with schedule $\mathbf{x}(t)$ */

Update $\hat{\mu}_{i,k}(t)$ and $\hat{\tau}_{i,k}(t)$ for all (i, k) with $k = x_i(t)$

While the procedure appears to be similar to that of Q-CSMA [17], we aim to minimize the accumulated regret rather than queue stability, and develop novel techniques to evaluate its performance. The complexity of the algorithm can be obtained as follows. In the control phase, arbitrary matching is selected uniformly at random, which takes $O(1)$ time. In the decision phase, each user calculates UCB indices for at most two channels: one in the candidate matching and/or one in the previous schedule, which can be done in parallel and takes $O(1)$ time. A central agent collects the indicies, which takes $O(N)$ time, and selects schedule $\mathbf{x}(t)$ by comparing $V(\mathbf{m}(t); \mathbf{I}(t))$ and $V(\mathbf{x}(t-1); \mathbf{I}(t))$, which takes $O(N)$ time. After the transmission, an update of $\hat{\mu}_{i,k}(t)$ and $\hat{\tau}_{i,k}(t)$ is necessary at each user i for channel $k = x_i(t)$, which takes $O(1)$ time. Thus, the total computational complexity of Uniform sampling is $O(N)$.

The idea of reducing complexity in uniform sampling can be applied to other MAB problems such as combinatorial MAB with more general graph rather than bipartite graph. Given graph $\mathcal{G} = (V, E)$, where V and E are the sets of nodes and edges, respectively, it is known that playing the maximum weighted matching with UCB index as a weight on each edge at each time slot achieves logarithmic growth of regret with respect to time [10]. Computational complexity of finding a maximum weighted matching in general graph is $O(V^2E)$ [19]. This polynomial complexity can be improved to linear complexity using our uniform sampling.

3.2 Performance evaluation

We now evaluate the performance of uniform sampling algorithm and show that it achieves the logarithmic growth of expected total regret with respect to time t. We first decompose the regret into the maximum non-optimality gap which will be defined later and the expected number of times that non-optimal matchings scheduled, and then show that the expected number of exploration to non-optimal matchings is bounded. The challenge of showing the latter comes from the procedure of selecting a schedule, i.e., sampling a candidate matching and comparing it to the previous schedule. When we select a schedule with the maximum weighted matching, the value of non-optimal matching $\mathbf{a}$ (i.e., $V(\mathbf{a}; \mathbf{I}(t))$) will be compared with the value of optimal matching $\mathbf{a}^*$ (i.e., $V(\mathbf{a}^*; \mathbf{I}(t))$). In contrast, when we 'sample' a candidate matching, there is a positive probability that both the candidate matching and the previous schedule are a non-optimal matching, thus we should take into account the comparison between the values of non-optimal matchings. We start by defining some notations.

Let us define $\Delta^*_{\mathbf{a}} := V(\mathbf{a}^*;\mu) - V(\mathbf{a};\mu)$, which is the expected regret of matching $\mathbf{a}$ and denoted by non-optimality gap of matching $\mathbf{a}$. Let $\Delta^*_{min} := \min_{\mathbf{a}\neq\mathbf{a}^*} \Delta^*_{\mathbf{a}}$ and $\Delta^*_{max} := \max_{\mathbf{a}\neq\mathbf{a}^*} \Delta^*_{\mathbf{a}}$ denote the minimum and maximum non-optimality gap, respectively. Let $\hat{\tau}_{\mathbf{a}}(T)$ denote the number of times that matching $\mathbf{a}$ is scheduled by time T, i,e,. $\hat{\tau}_{\mathbf{a}}(t) = \sum_{s=1}^{t} \mathbb{I}\{\mathbf{x}(s) = \mathbf{a}\}$. We denote the cardinality of a set by $|\cdot|$.

The following lemma provides an upper bound on the regret for any policy [14].

Lemma 1 *For any policy π, the expected total regret defined in (2) is upper-bounded as*

$$\mathcal{R}_\pi(T) \le \Delta^*_{max} \sum_{\mathbf{a}\neq\mathbf{a}^*} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right].$$

Proof: We can rewrite the expected total regret under policy π as

$$\begin{aligned}
\mathcal{R}_\pi(T) &= T\sum_{i=1}^{N} \mu_{i,a_i^*} - \sum_{\mathbf{a}\in\mathcal{M}} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}\left[X_{i,a_i}(t)\mathbb{I}\{\mathbf{x}(t)=\mathbf{a}\}\right] \\
&= \sum_{\mathbf{a}\in\mathcal{M}} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}\left[(\mu_{i,a_i^*} - X_{i,a_i}(t))\mathbb{I}\{\mathbf{x}(t)=\mathbf{a}\}\right] \\
&= \sum_{\mathbf{a}\in\mathcal{M}} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{I}\{\mathbf{x}(t)=\mathbf{a}\}\right](\mu_{i,a_i^*} - \mu_{i,a_i}) \\
&= \sum_{\mathbf{a}\in\mathcal{M}} \left(\mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] \sum_{i=1}^{N} (\mu_{i,a_i^*} - \mu_{i,a_i}) \right).
\end{aligned}$$

Then, we can upper-bound $\mathcal{R}_\pi(T)$ as

$$\mathcal{R}_\pi(T) \le \Delta^*_{max} \sum_{\mathbf{a}\neq\mathbf{a}^*} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right].$$

■

The following proposition is one of our main contributions.

Proposition 1 *Under uniform sampling, the expected number of exploration to non-optimal matchings is upper-bounded as*

$$\sum_{\mathbf{a}\neq\mathbf{a}^*} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] \le (|\mathcal{M}|-1)(|\mathcal{M}|+1)\left(\frac{4N^2(N+1)\log T}{(\Delta^*_{min})^2} + 1\right) + C_1 + C_2,$$

where $C_1 = |\mathcal{M}|(|\mathcal{M}|-2)\left(\frac{(|\mathcal{M}|-1)\cdot(|\mathcal{M}|-2)\cdot N\pi^2}{6|\mathcal{M}|} + 1\right)$ *and* $C_2 = |\mathcal{M}|\left(\frac{(|\mathcal{M}|-1)N\pi^2}{3} + 1\right)$.

Suppose that, for non-optimal matching $\mathbf{a}$, $V(\mathbf{a};\mathbf{I}(t)) \ge V(\mathbf{a}^*;\mathbf{I}(t))$. It implies that at least one of the following events occurs. 1) In $\mathbf{a}^*$, at least one of actual means is underestimated, 2) in $\mathbf{a}$, at least one of actual means is overestimated, and 3) $\mathbf{a}$ needs to be explored. From the

Chernoff-Hoeffding bound [20], the probability that each case of 1) and 2) occurs at time slot t can be bounded by Nt^{-2}. We show that, as non-optimal matching $\mathbf{a}$ is scheduled more (i.e., $\mathbf{a}$ is sufficiently explored), the matching $\mathbf{a}$ satisfies $V(\mathbf{a};\mathbf{I}(t)) < V(\mathbf{a}^*;\mathbf{I}(t))$ with high probability, and the probability that $\mathbf{x}(t) = \mathbf{a}^*$ approaches to 1.

We first show that the probability of underestimation of optimal matching $\mathbf{a}^*$ or overestimation of non-optimal matching $\mathbf{a}$ gets smaller.

Lemma 2 *Suppose that a non-optimal matching* $\mathbf{a}$ *is scheduled more than* $\left\lceil \frac{4N^2(N+1)\log t}{(\Delta^*_{min})^2} \right\rceil$ *times by time slot* t*. Then, the probability that the total sum of UCB indices from* $\mathbf{a}$ *is greater than that from an optimal matching* $\mathbf{a}^*$ *is less than* $2Nt^{-2}$*, i.e.,*

$$\mathbb{P}\left(V(\mathbf{a};\mathbf{I}(t)) \geq V(\mathbf{a}^*;\mathbf{I}(t))\right) \leq 2Nt^{-2}.$$

Proof: We let $\hat{\mu}_{i,k,\tau}$ denote average reward after user i by accessing channel k for τ times, and let $c_{t,s} = \sqrt{\frac{(N+1)\log t}{s}}$ denote the interval of confidence bound at time t. Let $l = \left\lceil \frac{4N^2(N+1)\log t}{(\Delta^*_{min})^2} \right\rceil$. Non-optimal matching $\mathbf{a}$ has been scheduled for $\hat{\tau}_{\mathbf{a}}(t) \geq l$, which implies that each edge (i, a_i) satisfies $\hat{\tau}_{i,a_i}(t) \geq l$ for all i. Comparing with the value function of optimal matching $\mathbf{a}^*$,

$$\begin{aligned}
&\mathbb{I}\{V(\mathbf{a};\mathbf{I}(t)) \geq V(\mathbf{a}^*;\mathbf{I}(t))\} \\
&\overset{(A)}{=} \mathbb{I}\{\sum_{i=1}^{N}(\hat{\mu}_{i,a_i,\hat{\tau}_{i,a_i}(t-1)} + c_{t-1,\hat{\tau}_{i,a_i}(t-1)}) \geq \sum_{i=1}^{N}(\hat{\mu}_{i,a_i^*,\tau_{i,a_i^*}(t-1)} + c_{t-1,\tau_{i,a_i^*}(t-1)})\} \\
&\leq \mathbb{I}\{\max_{l\leq s_1,\dots,s_N<t} \sum_{i=1}^{N}(\hat{\mu}_{i,a_i,s_i} + c_{t-1,s_i}) \geq \min_{0<s'_1,\dots,s'_N<t} \sum_{i=1}^{N}(\hat{\mu}_{i,a_i^*,s'_i} + c_{t-1,s'_i})\} \\
&\overset{(B)}{\leq} \sum_{s_1=l}^{t-1}\cdots\sum_{s_N=l}^{t-1}\sum_{s'_1=1}^{t-1}\cdots\sum_{s'_N=1}^{t-1} \mathbb{I}\{\sum_{i=1}^{N}(\hat{\mu}_{i,a_i,s_i} + c_{t-1,s_i}) \geq \sum_{i=1}^{N}(\hat{\mu}_{i,a_i^*,s'_i} + c_{t-1,s'_i})\} \\
&\leq \sum_{s_1=1}^{t}\cdots\sum_{s_N=1}^{t}\sum_{s'_1=1}^{t}\cdots\sum_{s'_N=1}^{t} \mathbb{I}\{\sum_{i=1}^{N}(\hat{\mu}_{i,a_i,s_i} + c_{t,s_i}) \geq \sum_{i=1}^{N}(\hat{\mu}_{i,a_i^*,s'_i} + c_{t,s'_i})\}
\end{aligned} \tag{6}$$

where equality (A) comes from the definition of $V(\cdot;\mathbf{I}(t))$ and (5), and inequality (B) can be obtained by summing the indicator functions for all $l \leq s_1, ..., s_N \leq t-1$ and $1 \leq s'_1, ..., s'_N \leq t-1$, which can be further extended to the last inequality. Let us denote the event $\sum_{i=1}^{N}(\hat{\mu}_{i,a_i,s_i} + c_{t,s_i}) \geq \sum_{i=1}^{N}(\hat{\mu}_{i,a_i^*,s'_i} + c_{t,s'_i})$ by Z and consider the following $2N+1$ events:

$$\begin{aligned}
A_i &: \hat{\mu}_{i,a_i^*,s'_i} \leq \mu_{i,a_i^*} - c_{t,s'_i}, \quad 1 \leq i \leq N, \\
B_i &: \hat{\mu}_{i,a_i,s_i} \geq \mu_{i,a_i} + c_{t,s_i}, \quad 1 \leq i \leq N, \\
C &: \sum_{i=1}^{N}\mu_{i,a_i^*} < \sum_{i=1}^{N}\mu_{i,a_i} + 2\sum_{i=1}^{N}c_{t,s_i}.
\end{aligned}$$

Suppose that event Z occurs; $\mathbb{I}\{Z\} = 1$. If $\sum_{i=1}^{N}\mathbb{I}\{A_i\} = 0$, then $\sum_{i=1}^{N}\hat{\mu}_{i,a_i,s_i} + \sum_{i=1}^{N}c_{t,s_i} \geq \sum_{i=1}^{N}\hat{\mu}_{i,a_i^*,s'_i} + \sum_{i=1}^{N}c_{t,s'_i} > \sum_{i=1}^{N}\mu_{i,a_i^*}$, where the first inequality comes from the occurrence of event Z and the second inequality comes from the non-occurrence of events $\{A_i\}$. If $\sum_{i=1}^{N}\mathbb{I}\{B_i\} =$

0, then $\sum_{i=1}^{N}\mu_{i,a_i}+2\sum_{i=1}^{N}c_{t,s_i}>\sum_{i=1}^{N}\hat{\mu}_{i,a_i,s_i}+\sum_{i=1}^{N}c_{t,s_i}$. Thus if none of events A_i and B_i occur, then by combining the two inequalities, we have $\sum_{i=1}^{N}\mu_{i,a_i}+2\sum_{i=1}^{N}c_{t,s_i}>\sum_{i=1}^{N}\mu_{i,a_i^*}$, i.e., $\mathbb{I}\{C\}=1$. Hence, at least one of the above $2N+1$ events must occur. Again, we note that the probability of each event A_i and B_i can be bounded by the Chernoff-Hoeffding bound [20] as

$$\mathbb{P}\left(\hat{\mu}_{i,a_i^*,s_i'}\le\mu_{i,a_i^*}-c_{t,s_i'}\right)\le t^{-2(N+1)},$$
$$\mathbb{P}\left(\hat{\mu}_{i,a_i,s_i}\ge\mu_{i,a_i}+c_{t,s_i}\right)\le t^{-2(N+1)},$$

respectively. Also, the probability of event C equals 0 if $s_i\ge\left\lceil\frac{4N^2(N+1)\log t}{(\Delta^*_{min})^2}\right\rceil$, because

$$\begin{aligned}0&>\sum_{i=1}^{N}\mu_{i,a_i^*}-\sum_{i=1}^{N}\mu_{i,a_i}-2\sum_{i=1}^{N}c_{t,s_i}\\&=\sum_{i=1}^{N}\mu_{i,a_i^*}-\sum_{i=1}^{N}\mu_{i,a_i}-2\sum_{i=1}^{N}\sqrt{\frac{(N+1)\log t}{s_i}}\\&\ge\sum_{i=1}^{N}\mu_{i,a_i^*}-\sum_{i=1}^{N}\mu_{i,a_i}-\Delta^*_{min}\\&\ge 0,\end{aligned}$$

where the last inequality comes from the fact that $\Delta^*_{min}\le\min_{\mathbf{a}\ne\mathbf{a}^*}\sum_{i=1}^{N}(\mu_{i,a_i^*}-\mu_{i,a_i})$. This implies that the probability that event Z occurs is no greater than $\sum_{i=1}^{N}(\mathbb{P}(A_i)+\mathbb{P}(B_i))$. By taking expectation over (6), we can obtain

$$\begin{aligned}&\mathbb{P}\left(V(\mathbf{a};\mathbf{I}(t))\ge V(\mathbf{a}^*;\mathbf{I}(t))\right)\\&\le\sum_{s_1=1}^{t}\cdots\sum_{s_N=1}^{t}\sum_{s_1'=1}^{t}\cdots\sum_{s_N'=1}^{t}\mathbb{P}\left(\sum_{i=1}^{N}(\hat{\mu}_{i,a_i,s_i}+c_{t,s_i})\ge\sum_{i=1}^{N}(\hat{\mu}_{i,a_i^*,s_i'}+c_{t,s_i'})\right)\\&\le\sum_{s_1=1}^{t}\cdots\sum_{s_N=1}^{t}\sum_{s_1'=1}^{t}\cdots\sum_{s_N'=1}^{t}2Nt^{-2(N+1)}\\&\le 2Nt^{-2}.\end{aligned}$$

■

Now we show that the number of exploration to non-optimal matching is bounded and we have the proposition.

Proof of Proposition 1: We classify the case into two exclusive subcases. Let T' is the smallest time slot that satisfies $\hat{\tau}_{\mathbf{a}}(T')\ge\left\lceil\frac{4N^2(N+1)\log T}{(\Delta^*_{min})^2}\right\rceil$ for all $\mathbf{a}\ne\mathbf{a}^*$, which denotes the time when all non-optimal matchings are sufficiently explored. If some non-optimal matching is not sufficiently scheduled, we may have $T'>T$. We divide the set of all matchings $\mathcal{M}$ into the set of non-optimal matchings denoted by $\mathcal{M}^o$ and the set of optimal matchings denoted by $\mathcal{M}^*$, i.e., $\mathcal{M}=\mathcal{M}^o\cup\mathcal{M}^*$.

(1) When $T'\le T$: Let $l=\left\lceil\frac{4N^2(N+1)\log T}{(\Delta^*_{min})^2}\right\rceil$. At time slot t, let $\overline{S}(t)$ denote the set of

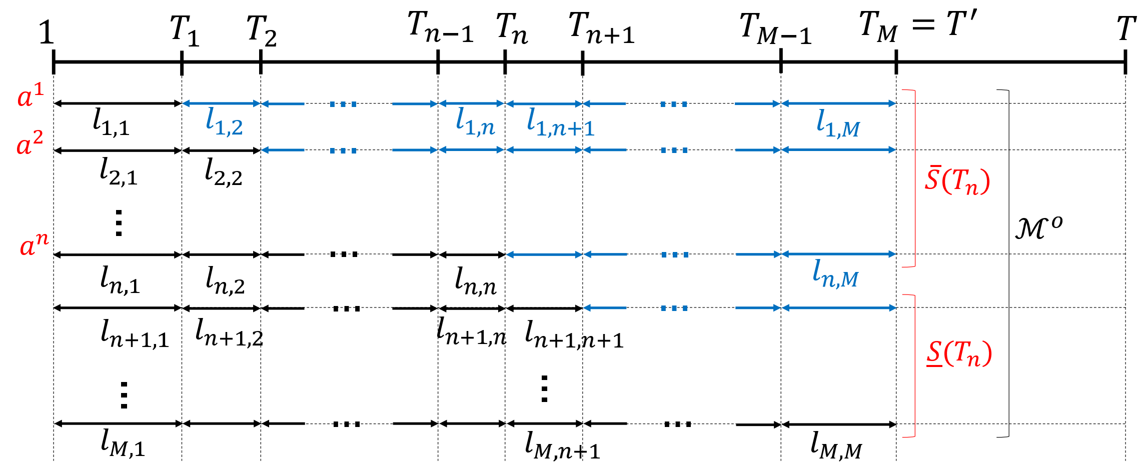


Figure 3: Matching $\mathbf{a}^n$ is scheduled $l_{n,m}$ times during $(T_{m-1}, T_m]$.

non-optimal matchings that are sufficiently scheduled with $\hat{\tau}_{\mathbf{a}}(t) \geq l$, and $\underline{S}(t)$ denote the set of non-optimal matchings that are insufficiently scheduled with $\hat{\tau}_{\mathbf{a}}(t) < l$. Let $M(\leq |\mathcal{M}| - 1)$ denote the number of non-optimal matchings, and let $\mathcal{M}^o = \{\mathbf{a}^1, \mathbf{a}^2, ..., \mathbf{a}^M\}$. Let T_n denote the smallest time at which matching $\mathbf{a}^n$ sufficiently scheduled, i.e., $\hat{\tau}_{\mathbf{a}^n}(T_n) = l$. Without loss of generality, we assume $T_1 < T_2 < \cdots < T_M = T'$. For $\mathbf{a}^n$, let $l_{n,m}$ denote the number of time slots that $\mathbf{a}^n$ is scheduled in $(T_{m-1}, T_m]$, as shown in Fig. 3. Note that $\sum_{m=1}^{n} l_{n,m} = l$ for all n. Then, we have

$$\begin{aligned}\sum_{\mathbf{a}\in\mathcal{M}^o} \hat{\tau}_{\mathbf{a}}(T') &= \sum_{\mathbf{a}\in\mathcal{M}^o} \sum_{t=1}^{T'} \mathbb{I}\{\mathbf{x}(t) = \mathbf{a}\} \\ &= lM + \sum_{n=1}^{M-1} \sum_{t=T_n+1}^{T_{n+1}} \sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{I}\{\mathbf{x}(t) = \mathbf{a}\}. \qquad (7)\end{aligned}$$

In the last equality, the first term denote the total number of schedules for non-optimal matchings up to l, which can be obtained by summing $l_{n,m}$ of black arrows in Fig. 3. The second term denotes the total number of time slots that each non-optimal matching $\mathbf{a}$ is scheduled after it is sufficiently scheduled, denoted by blue arrows in Fig. 3. The second term can be bounded by the maximum number of time slots that matching $\mathbf{a} \in \overline{S}(T_n)$ can be played during $(T_n, T_{n+1}]$. Note that $\overline{S}(T_n) \cup \underline{S}(T_n) \cup \mathcal{M}^* = \mathcal{M}$. We compute the probability of the second term by dividing the event $\mathbf{x}(t) = \mathbf{a}$ into three subcases based on $\mathbf{x}(t-1)$ as

$$
\begin{aligned}
&\sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right) \\
&= \sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a} \mid \mathbf{x}(t-1)\in\mathcal{M}^*\right)\mathbb{P}\left(\mathbf{x}(t-1)\in\mathcal{M}^*\right) && (8)\\
&\quad + \sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a} \mid \mathbf{x}(t-1)\in\underline{S}(T_n)\right)\mathbb{P}\left(\mathbf{x}(t-1)\in\underline{S}(T_n)\right) && (9)\\
&\quad + \sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a} \mid \mathbf{x}(t-1)\in\overline{S}(T_n)\right)\mathbb{P}\left(\mathbf{x}(t-1)\in\overline{S}(T_n)\right). && (10)
\end{aligned}
$$

The first term (8) can be bounded as

$$
\begin{aligned}
&\sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a} \mid \mathbf{x}(t-1)\in\mathcal{M}^*\right)\mathbb{P}\left(\mathbf{x}(t-1)\in\mathcal{M}^*\right) \\
&\le \sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{m}(t)=\mathbf{a}, V(\mathbf{a};\mathbf{I}(t)) \ge V(\mathbf{a}^*;\mathbf{I}(t))\right)\mathbb{P}\left(\mathbf{x}(t-1)\in\mathcal{M}^*\right) \\
&\le \sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(V(\mathbf{a};\mathbf{I}(t)) \ge V(\mathbf{a}^*;\mathbf{I}(t))\right) \\
&\le |\overline{S}(T_n)| \cdot 2Nt^{-2}, && (11)
\end{aligned}
$$

where the last inequality comes from Lemma 5, and the result holds for all $t \in (T_n, T_{n+1}]$. The second term (9) can be bounded by

$$
\begin{aligned}
&\sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a} \mid \mathbf{x}(t-1)\in\underline{S}(T_n)\right)\mathbb{P}\left(\mathbf{x}(t-1)\in\underline{S}(T_n)\right) \\
&\le \mathbb{P}\left(\mathbf{x}(t-1)\in\underline{S}(T_n)\right), && (12)
\end{aligned}
$$

for all $t \in (T_n, T_{n+1}]$. Now we obtain the bound of the third term (10).

$$
\begin{aligned}
&\sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a} \mid \mathbf{x}(t-1)\in\overline{S}(T_n)\right)\cdot\mathbb{P}\left(\mathbf{x}(t-1)\in\overline{S}(T_n)\right) \\
&= \sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)\in\overline{S}(T_n) \mid \mathbf{x}(t-1)=\mathbf{a}\right)\cdot\mathbb{P}\left(\mathbf{x}(t-1)=\mathbf{a}\right).
\end{aligned}
$$

We further divide the conditional probability as, for $\mathbf{a} \in \overline{S}(T_n)$ and some $\mathbf{a}^* \in \mathcal{M}^*$,

$$
\begin{aligned}
&\mathbb{P}\left(\mathbf{x}(t)\in\overline{S}(T_n) \mid \mathbf{x}(t-1)=\mathbf{a}\right) \\
&= \mathbb{P}\left(\mathbf{x}(t)\in\overline{S}(T_n) \mid \mathbf{x}(t-1)=\mathbf{a}, \mathbf{m}(t)=\mathbf{a}^*\right)\cdot\mathbb{P}\left(\mathbf{m}(t)=\mathbf{a}^*\right) \\
&\quad + \mathbb{P}\left(\mathbf{x}(t)\in\overline{S}(T_n) \mid \mathbf{x}(t-1)=\mathbf{a}, \mathbf{m}(t)\neq\mathbf{a}^*\right)\cdot\mathbb{P}\left(\mathbf{m}(t)\neq\mathbf{a}^*\right) \\
&\le \mathbb{P}\left(\mathbf{x}(t)\in\overline{S}(T_n) \mid \mathbf{x}(t-1)=\mathbf{a}, \mathbf{m}(t)=\mathbf{a}^*\right)\cdot\frac{1}{|\mathcal{M}|} + 1\cdot\frac{|\mathcal{M}|-1}{|\mathcal{M}|} \\
&= \mathbb{P}\left(V(\mathbf{a};\mathbf{I}(t)) \ge V(\mathbf{a}^*;\mathbf{I}(t))\right)\cdot\frac{1}{|\mathcal{M}|} + 1\cdot\frac{|\mathcal{M}|-1}{|\mathcal{M}|} \\
&\le \frac{1}{|\mathcal{M}|}\cdot 2Nt^{-2} + 1\cdot\frac{|\mathcal{M}|-1}{|\mathcal{M}|},
\end{aligned}
$$

where the last inequality comes from Lemma 5. Hence, we can obtain an upper bound as

$$
\begin{aligned}
&\sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a} \mid \mathbf{x}(t-1)\in\overline{S}(T_n)\right)\mathbb{P}\left(\mathbf{x}(t-1)\in\overline{S}(T_n)\right) \\
&\leq \sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t-1)=\mathbf{a}\right)\left[\frac{|\mathcal{M}|-1}{|\mathcal{M}|}+\frac{1}{|\mathcal{M}|}\cdot 2Nt^{-2}\right], \qquad (13)
\end{aligned}
$$

for all $t\in(T_n,T_{n+1}]$.

Letting $\alpha := \frac{|\mathcal{M}|-1}{|\mathcal{M}|}$ and combining (11), (12), and (13), we have

$$
\begin{aligned}
&\sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right) \\
&\leq \mathbb{P}\left(\mathbf{x}(t-1)\in\underline{S}(T_n)\right)+\left(|\overline{S}(T_n)|+\frac{1}{|\mathcal{M}|}\right)\cdot 2Nt^{-2}+\alpha\sum_{\mathbf{a}\in\overline{S}(T_n)}\mathbb{P}\left(\mathbf{x}(t-1)=\mathbf{a}\right).
\end{aligned}
$$

Letting $A := \left(|\overline{S}(T_n)|+\frac{1}{|\mathcal{M}|}\right)\cdot 2N$, the inequality can be extended in a recursive manner for $t\in(T_n,T_{n+1}]$ as

$$
\begin{aligned}
\sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right) &\leq \mathbb{P}\left(\mathbf{x}(t-1)\in\underline{S}(T_n)\right)+At^{-2} \\
&\quad+\alpha\left(\mathbb{P}\left(\mathbf{x}(t-2)\in\underline{S}(T_n)\right)+A(t-1)^{-2}+\alpha\sum_{\mathbf{a}\in\overline{S}(T_n)}\mathbb{P}\left(\mathbf{x}(t-2)=\mathbf{a}\right)\right).
\end{aligned}
$$

By extending it down to T_n, we can obtain

$$
\begin{aligned}
&\sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right) \\
&\leq \mathbb{P}\left(\mathbf{x}(t-1)\in\underline{S}(T_n)\right)+\alpha\mathbb{P}\left(\mathbf{x}(t-2)\in\underline{S}(T_n)\right)+\cdots+\alpha^{t-T_n-1}\mathbb{P}\left(\mathbf{x}(T_n+1)\in\underline{S}(T_n)\right) \qquad (14)\\
&\quad+A\left(t^{-2}+\alpha(t-1)^{-2}+\cdots+\alpha^{t-T_n-1}(T_n+1)^{-2}\right) \qquad (15)\\
&\quad+\alpha^{t-T_n}\sum_{\mathbf{a}\in\overline{S}(T_n)}\mathbb{P}\left(\mathbf{x}(T_n)=\mathbf{a}\right). \qquad (16)
\end{aligned}
$$

Now we compute $\sum_{t=T_n+1}^{T_{n+1}}\sum_{\mathbf{a}\in\overline{S}(T_n)}\mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right)$. By summing up (14) over $t\in(T_n,T_{n+1}]$, we have

$$
\begin{aligned}
&\sum_{t=T_n+1}^{T_{n+1}}\left(\mathbb{P}\left(\mathbf{x}(t-1)\in\underline{S}(T_n)\right)+\alpha\mathbb{P}\left(\mathbf{x}(t-2)\in\underline{S}(T_n)\right)+\cdots+\alpha^{t-T_n-1}\mathbb{P}\left(\mathbf{x}(T_n+1)\in\underline{S}(T_n)\right)\right) \\
&=\sum_{t=T_n+1}^{T_{n+1}}\sum_{s=0}^{T_{n+1}-t}\alpha^s\cdot\mathbb{E}\left[\mathbb{I}\{\mathbf{x}(t)\in\underline{S}(T_n)\}\right] \\
&\leq\frac{1}{1-\alpha}\mathbb{E}\left[\sum_{t=T_n+1}^{T_{n+1}}\mathbb{I}\{\mathbf{x}(t)\in\underline{S}(T_n)\}\right] \\
&=\frac{1}{1-\alpha}\mathbb{E}\left[\sum_{s=n+1}^{M}l_{s,n+1}\right], \qquad (17)
\end{aligned}
$$

where $\sum_{s=n+1}^{|\mathcal{M}|-1} l_{s,n+1}$ is shown as black arrows in Fig. 3. Similarly, we take the sum of (15) over $(T_n, T_{n+1}]$, as

$$\begin{aligned}
&\sum_{t=T_n+1}^{T_{n+1}} A\left(t^{-2}+\alpha(t-1)^{-2}+\cdots+\alpha^{t-T_n-1}(T_n+1)^{-2}\right)\\
&= A\sum_{t=T_n+1}^{T_{n+1}}\sum_{s=0}^{T_{n+1}-t}\alpha^s\cdot t^{-2}\\
&\le A\cdot\frac{1}{1-\alpha}\cdot\frac{\pi^2}{6}. \qquad (18)
\end{aligned}$$

Also, the last term (16) can be summed as

$$\begin{aligned}
&\sum_{t=T_n+1}^{T_{n+1}}\alpha^{t-T_n}\sum_{\mathbf{a}\in\overline{S}(T_n)}\mathbb{P}\left(\mathbf{x}(T_n)=\mathbf{a}\right)\\
&\le \sum_{t=T_n+1}^{T_{n+1}}\alpha^{t-T_n}\\
&\le \frac{1}{1-\alpha}, \qquad (19)
\end{aligned}$$

since $\sum_{\mathbf{a}\in\overline{S}(T_n)}\mathbb{P}\left(\mathbf{x}(T_n)=\mathbf{a}\right)\le 1$.

Combining (17), (18), and (19), and from $\alpha=\frac{|\mathcal{M}|-1}{|\mathcal{M}|}$, we have

$$\begin{aligned}
&\sum_{t=T_n+1}^{T_{n+1}}\sum_{\mathbf{a}\in\overline{S}(T_n)}\mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right)\\
&\le |\mathcal{M}|\cdot\left[\left(|\overline{S}(T_n)|+\frac{1}{|\mathcal{M}|}\right)\cdot\frac{N\pi^2}{3}+\mathbb{E}\left[\sum_{s=n+1}^{M} l_{s,n+1}\right]+1\right].
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\sum_{\mathbf{a}\in\mathcal{M}^o}\mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T')\right]\\
&=\sum_{\mathbf{a}\in\mathcal{M}^o}\sum_{t=1}^{T'}\mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right)\\
&= lM+\sum_{n=1}^{M-1}\sum_{t=T_n+1}^{T_{n+1}}\sum_{\mathbf{a}\in\overline{S}(T_n)}\mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right)\\
&\le lM+|\mathcal{M}|\left[\frac{N\pi^2}{3}\sum_{n=1}^{M-1}|\overline{S}(T_n)|+\mathbb{E}\left[\sum_{n=1}^{M-1}\sum_{s=n+1}^{M} l_{s,n+1}\right]+(M-1)\left(\frac{1}{|\mathcal{M}|}\cdot\frac{N\pi^2}{3}+1\right)\right]\\
&\le lM+|\mathcal{M}|(M-1)\left(\frac{(M-2)N\pi^2}{6}+l+\frac{1}{|\mathcal{M}|}\cdot\frac{N\pi^2}{3}+1\right),
\end{aligned}$$

where the last inequality comes from the following facts.

$$\text{(A)} \sum_{n=1}^{M-1} |\overline{S}(T_n)| = \sum_{n=1}^{M-1} n = \frac{(M-1)(M-2)}{2}, \tag{20}$$

$$\text{(B)} \sum_{n=1}^{M-1} \sum_{s=n+1}^{M} l_{s,n+1} = \sum_{n=1}^{M-1} \sum_{s=2}^{n} l_{n,s} \leq \sum_{n=1}^{M-1} l = l(M-1). \tag{21}$$

Therefore, we have

$$\sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T')\right] \leq (|\mathcal{M}|-1)(|\mathcal{M}|+1)\left(\frac{4N^2(N+1)\log T}{(\Delta^*_{min})^2}+1\right)+C_1, \tag{22}$$

where $C_1 = |\mathcal{M}|(|\mathcal{M}|-2)\left(\frac{(|\mathcal{M}|-1)\cdot(|\mathcal{M}|-2)\cdot N\pi^2}{6|\mathcal{M}|}+1\right)$.

Further, we have $\sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T) - \hat{\tau}_{\mathbf{a}}(T')\right] = \sum_{t=T'+1}^{T} \sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right)$, and divided $\mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right)$ into three subcases depending on the previous schedule and the candidate matching that can yield a non-optimal matching as

$$\begin{aligned}
&\sum_{t=T'+1}^{T} \sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right)\\
&= \sum_{t=T'+1}^{T} \Bigg(\sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{P}\left(\mathbf{x}(t-1)=\mathbf{a}\right)\mathbb{P}\left(\mathbf{m}(t)\in\mathcal{M}^*\right)\mathbb{P}\left(V(\mathbf{a},\mathbf{I}(t)) \geq V(\mathbf{a}^*,\mathbf{I}(t))\right)\\
&\qquad + \sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{P}\left(\mathbf{x}(t-1)=\mathbf{a}^*\right)\mathbb{P}\left(\mathbf{m}(t)=\mathbf{a}\right)\mathbb{P}\left(V(\mathbf{a},\mathbf{I}(t)) \geq V(\mathbf{a}^*,\mathbf{I}(t))\right)\\
&\qquad + \sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{P}\left(\mathbf{x}(t-1)=\mathbf{a}\right)\mathbb{P}\left(\mathbf{m}(t)\in\mathcal{M}^o\right)\Bigg)\\
&\leq \sum_{t=T'+1}^{T} \left(\sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{P}\left(V(\mathbf{a},\mathbf{I}(t)) \geq V(\mathbf{a}^*,\mathbf{I}(t))\right) + \mathbb{P}\left(\mathbf{m}(t)\in\mathcal{M}^o\right) \sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{P}\left(\mathbf{x}(t-1)=\mathbf{a}\right)\right).
\end{aligned} \tag{23}$$

From the Lemma 5, we have $\mathbb{P}\left(V(\mathbf{a},\mathbf{I}(t)) \geq V(\mathbf{a}^*,\mathbf{I}(t))\right) \leq 2Nt^{-2}$ for all $\mathbf{a}\in\mathcal{M}^o$, and $\mathbb{P}\left(\mathbf{m}(t)\in\mathcal{M}^o\right) \leq \frac{|\mathcal{M}|-1}{|\mathcal{M}|}$. Let $\alpha = \frac{|\mathcal{M}|-1}{|\mathcal{M}|}$, we have

$$\begin{aligned}
&\sum_{t=T'+1}^{T} \sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right)\\
&\leq \sum_{t=T'+1}^{T} \left((|\mathcal{M}|-1)\cdot 2Nt^{-2} + \alpha \sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{P}\left(\mathbf{x}(t-1)=\mathbf{a}\right)\right) && (24)\\
&\overset{(A)}{\leq} \sum_{t=T'+1}^{T} \left((|\mathcal{M}|-1)\cdot 2N \sum_{s=0}^{T-t} \alpha^s t^{-2} + \alpha^{t-T'} \sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{P}\left(\mathbf{x}(T')=\mathbf{a}\right)\right)\\
&\leq |\mathcal{M}|\left(\frac{(|\mathcal{M}|-1)N\pi^2}{3}+1\right), && (25)
\end{aligned}$$

where inequality (A) can be obtained by extending (24) in a recursive manner.

Therefore, combining (22) and (25) together, we have

$$\sum_{\mathbf{a}\neq\mathbf{a}^*} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T')\right] \leq (|\mathcal{M}|-1)(|\mathcal{M}|+1)\left(\frac{4N^2(N+1)\log T}{(\Delta^*_{min})^2}+1\right)+C_1+C_2, \tag{26}$$

where $C_1 = |\mathcal{M}|(|\mathcal{M}|-2)\left(\frac{(|\mathcal{M}|-1)\cdot(|\mathcal{M}|-2)\cdot N\pi^2}{6|\mathcal{M}|}+1\right)$ and $C_2 = |\mathcal{M}|\left(\frac{(|\mathcal{M}|-1)N\pi^2}{3}+1\right)$.

(2) When $T' > T$: Let $l = \left\lceil \frac{4N^2(N+1)\log T}{(\Delta^*_{min})^2} \right\rceil$. Let $\overline{S}(t)$ denote the set of matchings $\mathbf{a}$ with $\hat{\tau}_{\mathbf{a}}(t) \geq l$, and $\underline{S}(t)$ denote the set of matchings with $\hat{\tau}_{\mathbf{a}}(t) < l$. Let $|\overline{S}|$ and $|\underline{S}|$ denote the size of the set $\overline{S}(T)$ and $\underline{S}(T)$, respectively. Let $\{\mathbf{a}^1, \mathbf{a}^2, ..., \mathbf{a}^{|\overline{S}|}\}$ denote the set of non-optimal matchings which are sufficiently scheduled with $\hat{\tau}_{\mathbf{a}^n}(T) \geq l$, and let T_n denote the time at which matching $\mathbf{a}^n$ sufficiently scheduled, $\hat{\tau}_{\mathbf{a}^n}(T_n) = l$. Without loss of generality, we assume $T_1 < T_2 < ... < T_{|\overline{S}|}$. By time slot T, $\underline{S}(T)$ is non-empty. It is clear that $\sum_{\mathbf{a}\in\underline{S}(T)} \hat{\tau}_{\mathbf{a}}(T) \leq l|\underline{S}|$. Thus, we can write

$$
\begin{aligned}
&\sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] \\
&= \sum_{\mathbf{a}\in\underline{S}(T)} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] + \sum_{\mathbf{a}\in\overline{S}(T)} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] \\
&\leq l|\underline{S}| + \sum_{t=1}^{T}\sum_{\mathbf{a}\in\overline{S}} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right) \\
&\overset{(A)}{\leq} l|\underline{S}| + l|\overline{S}| + |\mathcal{M}||\overline{S}|\left[\left(\frac{(|\underline{S}|-1)N\pi^2}{6} + l + \frac{1}{|\mathcal{M}|}\cdot\frac{N\pi^2}{3}+1\right)\right] \\
&= l(|\mathcal{M}|-1+|\mathcal{M}||\overline{S}|) + |\mathcal{M}||\overline{S}|\left[\frac{(2+|\mathcal{M}|(|\overline{S}|-1))\cdot N\pi^2}{6|\mathcal{M}|}+1\right],
\end{aligned}
\tag{27}
$$

where inequality (A) can be obtained as the proof of the case when $T' \leq T$.

From (26) and (27), we have

$$
\sum_{\mathbf{a}\neq\mathbf{a}^*} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T')\right] \leq (|\mathcal{M}|-1)(|\mathcal{M}|+1)\left(\frac{4N^2(N+1)\log T}{(\Delta^*_{min})^2}+1\right) + C_1 + C_2,
$$

where $C_1 = |\mathcal{M}|(|\mathcal{M}|-2)\left(\frac{(|\mathcal{M}|-1)\cdot(|\mathcal{M}|-2)\cdot N\pi^2}{6|\mathcal{M}|}+1\right)$ and $C_2 = |\mathcal{M}|\left(\frac{(|\mathcal{M}|-1)N\pi^2}{3}+1\right)$. ■

Lemma 1 and Proposition 1 lead to the following result.

Theorem 1 *Under uniform sampling, the expected total regret $\mathcal{R}_U(T)$ by time T is upper bounded as*

$$
\mathcal{R}_U(T) \leq \Delta^*_{max}\left((|\mathcal{M}|-1)(|\mathcal{M}|+1)\cdot\left(\frac{4N^2(N+1)\log T}{(\Delta^*_{min})^2}+1\right)+C_1+C_2\right), \tag{28}
$$

where $C_1 = |\mathcal{M}|(|\mathcal{M}|-2)\left(\frac{(|\mathcal{M}|-1)\cdot(|\mathcal{M}|-2)\cdot N\pi^2}{6|\mathcal{M}|}+1\right)$ *and* $C_2 = |\mathcal{M}|\left(\frac{(|\mathcal{M}|-1)N\pi^2}{3}+1\right)$.

The theorem shows that regret $\mathcal{R}_{\mathrm{U}}(T)$ of uniform sampling is upper bounded by $O(\log T)$, which is asymptotically optimal [8]. We highlight that it is the first scheme that achieves $O(\log T)$ regret with linear complexity.

IV Greedy Algorithm in Randomized Orders

In this section, we first introduce a greedy algorithm that maps an order to a matching, and then describe our another low-complexity scheme that users the structure of bipartite graph and improves the regret performance of uniform sampling. We evaluate the performance of our proposed scheme, and show that it also achieves asymptotic optimality with respect to time.

4.1 Greedy algorithm

We consider the orders that can be mapped to a matching through a greedy algorithm, which will be used later in our scheme. We define an order $\mathbf{o}$ as a sequence of users $(o_1, ..., o_N)$ such that $o_i \in \{1, ..., N\}, o_i \neq o_j$ for any $i \neq j$ (i.e., a permutation of $\{1, ..., N\}$), where $o_j = i$ implies that user i is j-th in the order. Let $\mathcal{O}$ denote the set of all orders (permutations) of N users.

We now consider a greedy matching $greedy^{\mathbf{Y}}(\mathbf{o})$ that maps each order $\mathbf{o}$ to a matching $\mathbf{a}$ under some weight $\mathbf{Y}$. Given weight $\mathbf{Y}$ and order $\mathbf{o}$, it allows user o_1 to select channel $a_{o_1} = \arg\max_{k \in \mathcal{K}(\mathcal{G})} Y_{o_i,k}$, where $\mathcal{K}(\mathcal{G})$ denotes the set of channels in $\mathcal{G}$. A tie is broken in a predefined deterministic manner. Then we consider an induced graph $\bar{\mathcal{G}}_2^{\mathbf{Y}}$ by removing user o_1, a_{o_1}, and all edges connected to o_1 and a_{o_1} from $\bar{\mathcal{G}}_1^{\mathbf{Y}} = \mathcal{G}$. The next user o_2 selects channel a_{o_2} with the maximum weight in the induced graph $\bar{\mathcal{G}}_2^{\mathbf{Y}}$, and yields $\bar{\mathcal{G}}_3^{\mathbf{Y}}$ by removing o_2, a_{o_2}, and their associated edges. The procedure repeats following the order $\mathbf{o}$ as shown in Algorithm 2.

Algorithm 2 Greedy matching algorithm $greedy^{\mathbf{Y}}(\mathbf{o})$.

Input: $\mathcal{G} = (\mathcal{N} \cup \mathcal{K}, E)$, weight $\mathbf{Y}$, order $\mathbf{o}$

```
1: Ḡ_1^Y ← G
2: for j = 1 to N do
3:     i ← o_j
4:     a_i ← argmax_{k∈K(Ḡ_j^Y)} Y_{i,k}
5:     Ḡ_{j+1}^Y obtained by removing i, a_i, and all edges connected to i and a_i from Ḡ_j^Y
6: end for
7: return a;
```

We consider the greedy matchings with $\mathbf{Y} = \{\mu_{i,k}\}$, from which an order $\mathbf{o}$ is mapped to a set of channels $\mathbf{a}^o := greedy^{\mu}(\mathbf{o})$. Note that different orders may yield the same greedy matching. Let $\mathcal{M}_G := \{\mathbf{a}^{\mathbf{o}} : \mathbf{o} \in \mathcal{O}\}$ denote the set of all possible greedy matchings with weight of actual means $\{\mu_{i,k}\}$. Let us define value function $V(\mathbf{a}; \mu) := \sum_{i=1}^{N} \mu_{i,a_i}$, which can be used to evaluate a matching. An optimal matching $\mathbf{a}^*$ can be written as $\mathbf{a}^* \in \arg\max_{\mathbf{a}} V(\mathbf{a}; \mu)$. We show the following lemma.

Lemma 3 *The set $\mathcal{M}_G$ of all greedy matchings includes an optimal matching, i.e., $\mathbf{a}^* \in \mathcal{M}_G$.*

Lemma 3 implies that there exists an order $\mathbf{o}$ such that $V(greedy^{\mu}(\mathbf{o}); \mu) = V(\mathbf{a}^*; \mu)$. We prove it by exploiting the fact that in the optimal matching $\mathbf{a}^*$, at least one user must play its best channel (i.e., the channel with the highest actual mean). Let S_1 denote the set of users who are associated with their best channel in the optimal matching $\mathbf{a}^*$. We show the following lemma.

Lemma 4 *In bipartite graph $\mathcal{G}$, S_1 is not empty.*

Proof: We assume without loss of generality that the graph $\mathcal{G}$ is symmetric complete bipartite graph with $N = K$. For non-symmetric or incomplete bipartite graphs, we can construct such a graph by adding additional users and by setting zero weight to originally non-existing edges. Suppose that S_1 is empty, i.e., no user plays its best channel in the optimal matching of $\mathcal{G}$. Let k_i^* denote user i's best channel. Given $\mathcal{G} = (\mathcal{N} \cup \mathcal{K}, E)$, we consider subgraph $\mathcal{G}' = (\mathcal{N} \cup \mathcal{K}, E')$, where E' consists of edges (i, a_i^*) with its (non-negative) weight and edges (i, k_i^*) with their weight multiplied by -1 (i.e., non-positive weight) for all i. Since no user plays its best channel, each user has exactly two edges (one with non-negative weight and another with non-positive weight). Then, graph $\mathcal{G}'$ has the same number of vertices and edges of $2N$, and there exist at least one alternating cycle C [21]. The cycle should have a negative weight sum since the sum of incoming and outgoing edges of each user is always less than or equal to zero. This implies that we can improve the weight sum by replacing all edges $\{(i, a_i^*)\} \cap C$ with edges of $C \setminus \{(i, a_i^*)\}$. This contradicts that $\mathbf{a}^*$ is an optimal matching. ■

Proof of Lemma 3: We prove the lemma by constructing an order $\mathbf{o}^*$ given an optimal matching $\mathbf{a}^*$. By Lemma 4, S_1 is not empty, and we let the users in S_1 to have the earliest order. Note that the order within S_1 is not important under our greedy algorithm since each user will choose a different channel in $\mathbf{a}^*$. Let $\mathcal{G}_s$ denote (bipartite) subgraph obtained by excluding all users in S_1 and all assigned channels and corresponding edges. Let S_1' denote the set of users playing their best channel in subgraph $\mathcal{G}_s$. Note that the induced matching $\mathbf{a}^*|_{\mathcal{G}_s}$ is also an optimal matching in subgraph $\mathcal{G}_s$ (otherwise, we can easily show that $\mathbf{a}^*$ is not optimal in $\mathcal{G}$), and from Lemma 4, we can find that S_1' is also not empty. We let the users in S_1' to be in the group with the second earliest order. Repeating the procedure, we can obtain an order $\mathbf{o}^*$ that yields $\mathbf{a}^*$ through greedy algorithm. ■

Using Lemma 3, we can find an optimal matching through an exhaustive search over the orders. By finding the order $\mathbf{o}^*$ with the maximum value function $\mathbf{o}^* \in \arg\max_{\mathbf{o} \in \mathcal{O}} V(greedy^{\mu}(\mathbf{o}); \mu)$, we can obtain an optimal matching $\mathbf{a}^* = greedy^{\mu}(\mathbf{o}^*)$. However, it requires searching over all $N!$ permutations. In the following, we develop a search algorithm with lower complexity.

4.2 Greedy algorithm in randomized orders

The results of Section 4.1 cannot be directly used since the channel statistics are unknown a priori. Thus, we apply the same approach in Section III, i.e., storing the history as (3) and (4), and scheduling with UCB indices as (5) instead of actual means.

Now we explain our GreedY in Randomized Orders (*GYRO*), which is shown in Algorithm 3. *GYRO* has the same time structure and procedure as uniform sampling except for the way to select candidate matching in the control phase. It selects an order $\mathbf{o}(t) \in \mathcal{O}$ uniformly at random, and then maps $\mathbf{o}(t)$ to matching $\mathbf{m}(t)$ by using greedy matching algorithm with weight $\mathbf{Y} = \mathbf{I}(t)$. Note that while uniform sampling selects $\mathbf{m}(t)$ uniformly at random in $\mathcal{M}$, *GYRO* selects $\mathbf{m}(t)$

from $\mathcal{M}_G$ asymptotically. This will result in a noticeable performance improvement. We show that *GYRO* outperforms uniform sampling, in particular with larger network size, through simulations in Section V.

Algorithm 3 GreedY in Randomized Orders (GYRO).

At the beginning of each time slot t

Select $\mathbf{o}(t) \in \mathcal{O}$ uniformly at random
Calculate $I_{i,k}(t)$ for all (i,k)
$\mathbf{m}(t) \leftarrow greedy^{\mathbf{I}(t)}(\mathbf{o})$
$\mathbf{x}(t) \in \arg\max_{\mathbf{a}\in\{\mathbf{m}(t),\mathbf{x}(t-1)\}} V(\mathbf{a};\mathbf{I}(t))$
/* make transmissions with schedule $\mathbf{x}(t)$ */
Update $\hat{\mu}_{i,k}(t)$ and $\hat{\tau}_{i,k}(t)$ for all (i,k) with $k = x_i(t)$

The computational complexity of *GYRO* can be obtained as follows. In control phase, each user calculates UCB indices for all channels in parallel, which takes $O(K)$ time. A central agent collects the indices, which takes $O(N)$ times, and selects an order uniformly at random. Given the indices and the order, the agent determines candidate matching $\mathbf{m}(t)$, which takes $O(N)$ time. In decision phase, the agent selects schedule $\mathbf{x}(t)$ by comparing $V(\mathbf{m}(t);\mathbf{I}(t))$ and $V(\mathbf{x}(t-1);\mathbf{I}(t))$, which can be done in $O(N)$ time. The final schedule $\mathbf{x}(t)$ is distributed to each user in $O(N)$ time. After the transmission, an update of $\hat{\mu}_{i,k}(t)$ and $\hat{\tau}_{i,k}(t)$ is necessary at each user i for channel $k = x_i(t)$, which takes $O(1)$ time. Thus, the total computational complexity of *GYRO* is $O(K)$ for $K \geq N$.

4.3 Performance evaluation

We now show that it achieves the logarithmic growth of expected total regret with respect to time t, and improves the performance of uniform sampling. We start with some notations.

Let us define $\delta_{\mathbf{a}}^{\mathbf{o}} := \min_{i,\mu_{i,a_i^{\mathbf{o}}}>\mu_{i,a_i}}\{\mu_{i,a_i^{\mathbf{o}}} - \mu_{i,a_i}\}$, which is the minimum mean gap among users such that $\mu_{i,a_i^{\mathbf{o}}} > \mu_{i,a_i}$ in a matching $\mathbf{a}$ given an order $\mathbf{o}$. Let $\delta_{min}^{\mathbf{o}} = \min_{\mathbf{a}\neq\mathbf{a}^{\mathbf{o}}} \delta_{\mathbf{a}}^{\mathbf{o}}$, and $\Delta_{min} = \min\{\Delta_{min}^*,\ \min_{\mathbf{o}\in\mathcal{O}} \delta_{min}^{\mathbf{o}}\}$. Let $\mathcal{O}^*$ denote the set of orders such that $greedy^{\mu}(\mathbf{o}) = \mathbf{a}^*$, which is not empty by Lemma 3.

Proposition 2 *Under GYRO, the expected number of exploration to non-optimal matchings is upper-bounded as*

$$\sum_{\mathbf{a}\neq\mathbf{a}^*} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] \leq (|\mathcal{M}|-1)(N!+1)\cdot\left(\frac{4N^2(N+1)\log T}{\Delta_{min}^2}+1\right)+C_1+C_2,$$

where $C_1 = N!(|\mathcal{M}|-2)\left(\frac{(|\mathcal{M}|-3)N\pi^2}{6}\left(1+\frac{1}{N!}\right)+1\right)$, *and* $C_2 = \frac{N!}{|\mathcal{O}^*|}\left(\frac{(|\mathcal{M}|-1)N\pi^2}{3}\left(1+\frac{|\mathcal{O}^*|}{N!}\right)+1\right)$.

Suppose that in control phase, given $\mathbf{o}(t) = \mathbf{o}$, non-greedy matching $\mathbf{a} \neq greedy^{\mu}(\mathbf{o})$ is picked as candidate matching $\mathbf{m}(t)$. It implies that at least one of the following events occurs. 1)

In $\mathbf{a^o} = greedy^{\mu}(\mathbf{o})$, at least one of actual means is underestimated where $\mathbf{a}$ wins $\mathbf{a^o}$ in the greedy comparison, 2) in $\mathbf{a}$, at least one of actual means is overestimated, and 3) a non-greedy matching needs to be explored (i.e., some index excessively increases). From the Chernoff-Hoeffding bound [20], the probability that each case of 1) and 2) occurs at time slot t can be bounded by Nt^{-2}. Further, if a non-greedy matching is played for a sufficient number of times, then the matching does not need to be explored with high probability. This implies that after non-optimal matchings are scheduled sufficiently, $\mathbf{m}(t) = \mathbf{a}^*$ with positive probability, and the probability that $V(\mathbf{a}; \mathbf{I}(t)) < V(\mathbf{a}^*; \mathbf{I}(t))$ is close to 1. It implies that there is a positive probability that an optimal matching is scheduled and then it remains scheduled with high probability, which provides the bound.

We first show some lemmas which is used to prove Proposition 2.

Lemma 5 *Suppose that a non-optimal matching* $\mathbf{a}$ *is scheduled more than* $\left\lceil \frac{4N^2(N+1)\log t}{\Delta_{min}^2} \right\rceil$ *times by time slot* t*. Then, the probability that the total sum of UCB indices from* $\mathbf{a}$ *is greater than that from an optimal matching* $\mathbf{a}^*$ *is less than* $2Nt^{-2}$*, i.e.,*

$$\mathbb{P}\left(V(\mathbf{a}; \mathbf{I}(t)) \geq V(\mathbf{a}^*; \mathbf{I}(t))\right) \leq 2Nt^{-2}.$$

We omit its proof since it can be shown similarly as the proof of Lemma 5. Note that $\Delta_{min} \leq \Delta^*_{min}$

Lemma 6 *Let* $\bar{\mathbf{a}}_t = \arg\max_{\mathbf{a}} V(\mathbf{a}; \mathbf{I}(t))$ *denote a matching with highest UCB index at time slot* t*. Then, there exists an order* $\bar{\mathbf{o}}_t \in \mathcal{O}$ *which results in* $\bar{\mathbf{a}}_t$*, i.e.,* $\bar{\mathbf{o}}_t \in \arg\max_{\mathbf{o}} greedy^{\mathbf{I}(t)}(\mathbf{o})$*.*

We omit its proof since it can be shown similarly as the proof of Lemma 3.

Lemma 7 *Consider matching* $\mathbf{a}$ *that has been scheduled sufficiently* $\hat{\tau}_{\mathbf{a}}(t) \geq \left\lceil \frac{4N^2(N+1)\log t}{\Delta_{min}^2} \right\rceil$*. If an order* $\mathbf{o}$ *such that* $\mathbf{a} \neq greedy^{\mu}(\mathbf{o})$ *is chosen in the control phase at time slot* t*, then the probability that* $\mathbf{a}$ *is picked as candidate matching* $\mathbf{m}(t)$ *is less than* $2Nt^{-2}$*, i.e.,*

$$\mathbb{P}\left(\mathbf{m}(t) = \mathbf{a} \mid \mathbf{o}(t) = \mathbf{o}, \mathbf{a} \neq greedy^{\mu}(\mathbf{o})\right) \leq 2Nt^{-2}.$$

Further, if $\hat{\tau}_{\mathbf{a}}(t) \geq \left\lceil \frac{4N^2(N+1)\log t}{\Delta_{min}^2} \right\rceil$ *for all matchings* $\mathbf{a} \neq \mathbf{a}^*$*, then*

$$\mathbb{P}\left(\mathbf{m}(t) \neq \mathbf{a}^*\right) \leq \frac{N! - |\mathcal{O}^*|}{N!} + \frac{|\mathcal{O}^*|}{N!} 2(|\mathcal{M}| - 1)Nt^{-2}.$$

Proof: For given $\mathbf{o}$ and $\mathbf{a} \neq \mathbf{a^o} = greedy^{\mu}(\mathbf{o})$, since $\mathbf{m}(t) = greedy^{\mathbf{I}(t)}(\mathbf{o})$, if $\mathbf{m}(t) = \mathbf{a}$, there exist at least one user i such that $\mu_{i,a_i} \leq \mu_{i,a_i^{\mathbf{o}}}$ and $I_{i,a_i}(t) \geq I_{i,a_i^{\mathbf{o}}}(t)$. Let $\hat{\mu}_{i,k,\tau}$ denote average reward for user i by playing channel k for τ times, and let $c_{t,s} = \sqrt{\frac{(N+1)\log t}{s}}$ denote the confidence bound at time t. Further, let $l = \left\lceil \frac{4N^2(N+1)\log t}{\Delta_{min}^2} \right\rceil$. Since matching $\mathbf{a}$ has been scheduled for $\hat{\tau}_{\mathbf{a}}(t) \geq l$, each edge (i, a_i) should satisfy $\hat{\tau}_{i,a_i}(t) \geq l$ for all i. Then, for $\mathbf{a} \neq \mathbf{a^o}$, we have

$$
\begin{aligned}
&\mathbb{I}\{greedy^{\mathbf{I}(t)}(\mathbf{o}) = \mathbf{a}\} \\
&\leq \mathbb{I}\{I_{i,a_i}(t) \geq I_{i,a_i^\circ}(t) \text{ and } \mu_{i,a_i} \leq \mu_{i,a_i^\circ} \text{ for some } i\} \\
&\leq \sum_{i:\mu_{i,a_i} \leq \mu_{i,a_i^\circ}} \mathbb{I}\{I_{i,a_i}(t) \geq I_{i,a_i^\circ}(t)\} \\
&\overset{(A)}{=} \sum_{i:\mu_{i,a_i} \leq \mu_{i,a_i^\circ}} \mathbb{I}\{\hat{\mu}_{i,a_i,\hat{\tau}_{i,a_i}(t-1)} + c_{t-1,\hat{\tau}_{i,a_i}(t-1)} \geq \hat{\mu}_{i,a_i^\circ,\hat{\tau}_{i,a_i^\circ}(t-1)} + c_{t-1,\hat{\tau}_{i,a_i^\circ}(t-1)}\} \\
&\leq \sum_{i:\mu_{i,a_i} \leq \mu_{i,a_i^\circ}} \mathbb{I}\{\max_{l \leq s_i < t}(\hat{\mu}_{i,a_i,s_i} + c_{t-1,s_i}) \geq \min_{0 < s_i' < t}(\hat{\mu}_{i,a_i^\circ,s_i'} + c_{t-1,s_i'})\} \\
&\overset{(B)}{\leq} \sum_{i:\mu_{i,a_i} \leq \mu_{i,a_i^\circ}} \sum_{s_i=l}^{t-1} \sum_{s_i'=1}^{t-1} \mathbb{I}\{\hat{\mu}_{i,a_i,s_i} + c_{t-1,s_i} \geq \hat{\mu}_{i,a_i^\circ,s_i'} + c_{t-1,s_i'}\} \\
&\leq \sum_{i:\mu_{i,a_i} \leq \mu_{i,a_i^\circ}} \sum_{s_i=1}^{t} \sum_{s_i'=1}^{t} \mathbb{I}\{\hat{\mu}_{i,a_i,s_i} + c_{t,s_i} \geq \hat{\mu}_{i,a_i^\circ,s_i'} + c_{t,s_i'}\},
\end{aligned} \tag{29}
$$

where equality (A) comes from (5), and inequality (B) can be obtained by summing the indicator functions for all $l \leq s_i \leq t-1$ and $1 \leq s_i' \leq t-1$, which can be further extended to the last inequality.

We pay attention to the event $\hat{\mu}_{i,a_i,s_i} + c_{t,s_i} \geq \hat{\mu}_{i,a_i^\circ,s_i'} + c_{t,s_i'}$ for users i such that $\mu_{i,a_i^\circ} \geq \mu_{i,a_i}$. For those i, at least one of the following three events must occur.

$$
\begin{aligned}
&A_i : \hat{\mu}_{i,a_i^\circ,s_i'} \leq \mu_{i,a_i^\circ} - c_{t,s_i'}, \\
&B_i : \hat{\mu}_{i,a_i,s_i} \geq \mu_{i,a_i} + c_{t,s_i}, \\
&C_i : \mu_{i,a_i^\circ} < \mu_{i,a_i} + 2c_{t,s_i}.
\end{aligned}
$$

If event A_i does not occur, then $\hat{\mu}_{i,a_i,s_i} + c_{t,s_i} \geq \hat{\mu}_{i,a_i^\circ,s_i'} + c_{t,s_i'} > \mu_{i,a_i^\circ}$. If event B_i does not occur, then $\mu_{i,a_i} + 2c_{t,s_i} > \hat{\mu}_{i,a_i,s_i} + c_{t,s_i}$. Thus if both events A_i and B_i do not occur, then by combining these two inequalities, we have $\mu_{i,a_i} + 2c_{t,s_i} > \mu_{i,a_i^\circ}$, which implies event C_i. Hence, at least one of the above events must occur. Note that the probability of events A_i and B_i can be bounded by the Chernoff-Hoeffding bound [20] as,

$$
\begin{aligned}
&\mathbb{P}\left(\hat{\mu}_{i,a_i^\circ,s_i'} \leq \mu_{i,a_i^\circ} - c_{t,s_i'}\right) \leq t^{-2(N+1)}, \\
&\mathbb{P}\left(\hat{\mu}_{i,a_i,s_i} \geq \mu_{i,a_i} + c_{t,s_i}\right) \leq t^{-2(N+1)},
\end{aligned}
$$

respectively. Also, the probability of event C_i equals 0 if $s_i \geq \left\lceil \frac{4N^2(N+1)\log t}{\Delta_{min}^2} \right\rceil$, because

$$
\begin{aligned}
0 &> \mu_{i,a_i^\circ} - \mu_{i,a_i} - 2c_{t,s_i} \\
&= \mu_{i,a_i^\circ} - \mu_{i,a_i} - 2\sqrt{\frac{(N+1)\log t}{s_i}} \\
&\geq \mu_{i,a_i^\circ} - \mu_{i,a_i} - \frac{\Delta_{min}}{N} \\
&\geq 0,
\end{aligned}
$$

where the last inequality comes from the fact that $\Delta_{min} \leq \min_{i,\mu_{i,a_i^\circ} > \mu_{i,a_i}} \{\mu_{i,a_i^\circ} - \mu_{i,a_i}\}$. This implies that for user i with $\mu_{i,a_i^\circ} \geq \mu_{i,a_i}$, the probability that the event $\hat{\mu}_{i,a_i,s_i} + c_{t,s_i} \geq \hat{\mu}_{i,a_i^\circ,s_i'} + c_{t,s_i'}$ occurs is no greater than $\mathbb{P}(A_i) + \mathbb{P}(B_i)$. By taking conditional expectation over (29), we can obtain

$$\begin{aligned}
&\mathbb{P}(\mathbf{m}(t) = \mathbf{a} \mid \mathbf{o}(t) = \mathbf{o}, \mathbf{a} \neq \mathit{greedy}^{\mu}(\mathbf{o})) \\
&\leq \sum_{i:\mu_{i,a_i} \leq \mu_{i,a_i^\circ}} \sum_{s_i=1}^{t} \sum_{s_i'=1}^{t} \mathbb{P}\left(\hat{\mu}_{i,a_i,s_i} + c_{t,s_i} \geq \hat{\mu}_{i,a_i^\circ,s_i'} + c_{t,s_i'}\right) \\
&\leq \sum_{i:\mu_{i,a_i} \leq \mu_{i,a_i^\circ}} \sum_{s_i=1}^{t} \sum_{s_i'=1}^{t} 2t^{-2(N+1)} \\
&\leq 2Nt^{-2}.
\end{aligned}$$

Further, if $\hat{\tau}_{\mathbf{a}}(t) \geq \left\lceil \frac{4N^2(N+1)\log t}{\Delta_{min}^2} \right\rceil$ for all matchings $\mathbf{a} \neq \mathbf{a}^*$, using the above result, we can obtain

$$\begin{aligned}
\mathbb{P}(\mathbf{m}(t) \neq \mathbf{a}^*) &= \sum_{\mathbf{o} \in \mathcal{O}} \mathbb{P}(\mathbf{m}(t) \neq \mathbf{a}^* \mid \mathbf{o}(t) = \mathbf{o}) \, \mathbb{P}(\mathbf{o}(t) = \mathbf{o}) \\
&\leq \sum_{\mathbf{o} \notin \mathcal{O}^*} \mathbb{P}(\mathbf{o}(t) = \mathbf{o}) + \sum_{\mathbf{o} \in \mathcal{O}^*} \sum_{\mathbf{a} \neq \mathbf{a}^*} \mathbb{P}(\mathbf{m}(t) = \mathbf{a} \mid \mathbf{o}(t) = \mathbf{o}, \mathbf{a} \neq \mathit{greedy}^{\mu}(\mathbf{o})) \cdot \mathbb{P}(\mathbf{o}(t) = \mathbf{o}) \\
&\leq \frac{N! - |\mathcal{O}^*|}{N!} + \frac{|\mathcal{O}^*|}{N!} \cdot (|\mathcal{M}| - 1) \cdot 2Nt^{-2},
\end{aligned}$$

where the first inequality holds, since, for $\mathbf{o} \in \mathcal{O}^*$ and $\mathbf{a} \neq \mathbf{a}^*$, we have $\mathit{greedy}^{\mu}(\mathbf{o}) = \mathbf{a}^*$ and thus $\mathbb{P}(\mathbf{m}(t) = \mathbf{a} \mid \mathbf{o}(t) = \mathbf{o}) = \mathbb{P}(\mathbf{m}(t) = \mathbf{a} \mid \mathbf{o}(t) = \mathbf{o}, \mathbf{a} \neq \mathit{greedy}^{\mu}(\mathbf{o}))$, and the last inequality holds since the order is chosen uniformly at random from $N!$ permutations (i.e., $\mathbb{P}(\mathbf{o}(t) = \mathbf{o}) = \frac{1}{N!}$) and the number of non-optimal matchings is no greater than $|\mathcal{M}| - 1$. ■

Proof of Proposition 2: The procedure of showing the proposition is the same as the proof of Proposition 1 except for obtaining the bound of (10). This difference comes from the way to select a candidate matching under uniform sampling and *GYRO*. Therefore, we omit the description of notations used in the proof, and begin with obtaining the bound of (10) when $T' \leq T$.

(1) When $T' \leq T$:

$$\begin{aligned}
&\sum_{\mathbf{a} \in \overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t) = \mathbf{a} \mid \mathbf{x}(t-1) \in \overline{S}(T_n)\right) \cdot \mathbb{P}\left(\mathbf{x}(t-1) \in \overline{S}(T_n)\right) \\
&= \sum_{\mathbf{a} \in \overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t) \in \overline{S}(T_n) \mid \mathbf{x}(t-1) = \mathbf{a}\right) \cdot \mathbb{P}\left(\mathbf{x}(t-1) = \mathbf{a}\right).
\end{aligned}$$

Note that for $\bar{\mathbf{a}}_t = \arg\max_{\mathbf{a} \in \mathcal{M}} V(\mathbf{a}; \mathbf{I}(t))$, there exists $\bar{\mathbf{o}}_t \in \mathcal{O}$ such that $\mathit{greedy}^{\mathbf{I}(t)}(\bar{\mathbf{o}}_t) = \bar{\mathbf{a}}_t$

from Lemma 6. We further divide the conditional probability using $\bar{\mathbf{o}}_t$ as

$$\begin{aligned}
&\mathbb{P}\left(\mathbf{x}(t) \in \overline{S}(T_n) \mid \mathbf{x}(t-1) = \mathbf{a}\right) \\
&= \mathbb{P}\left(\mathbf{x}(t) \in \overline{S}(T_n) \mid \mathbf{x}(t-1) = \mathbf{a}, \mathbf{o}(t) = \bar{\mathbf{o}}_t\right) \cdot \mathbb{P}\left(\mathbf{o}(t) = \bar{\mathbf{o}}_t\right) \\
&\quad + \mathbb{P}\left(\mathbf{x}(t) \in \overline{S}(T_n) \mid \mathbf{x}(t-1) = \mathbf{a}, \mathbf{o}(t) \neq \bar{\mathbf{o}}_t\right) \cdot \mathbb{P}\left(\mathbf{o}(t) \neq \bar{\mathbf{o}}_t\right) \\
&\leq \mathbb{P}\left(\mathbf{x}(t) \in \overline{S}(T_n) \mid \mathbf{x}(t-1) = \mathbf{a}, \mathbf{o}(t) = \bar{\mathbf{o}}_t\right) \cdot \frac{1}{N!} + 1 \cdot \frac{N!-1}{N!}.
\end{aligned}$$

Note that $\bar{\mathbf{o}}_t$ leads to $\bar{\mathbf{a}}_t$ and $V(\bar{\mathbf{a}}_t; \mathbf{I}(t)) \geq V(\mathbf{a}; \mathbf{I}(t))$ for all $\mathbf{a}$ at time t, which implies that $\mathbf{x}(t) = \bar{\mathbf{a}}_t$ regardless of $\mathbf{x}(t-1)$. Hence, $\mathbb{P}\left(\mathbf{x}(t) \in \overline{S}(T_n) \mid \mathbf{x}(t-1) = \mathbf{a}, \mathbf{o}(t) = \bar{\mathbf{o}}_t\right) = \mathbb{P}\left(\bar{\mathbf{a}}_t \in \overline{S}(T_n)\right)$, where

$$\begin{aligned}
\mathbb{P}\left(\bar{\mathbf{a}}_t \in \overline{S}(T_n)\right) &\leq \sum_{\mathbf{a} \in \overline{S}(T_n)} \mathbb{P}\left(\mathbf{a} \in \arg\max_{\mathbf{a}' \in \mathcal{M}} V(\mathbf{a}'; \mathbf{I}(t))\right) \\
&\leq \sum_{\mathbf{a} \in \overline{S}(T_n)} \mathbb{P}\left(V(\mathbf{a}; \mathbf{I}(t)) \geq V(\mathbf{a}^*; \mathbf{I}(t))\right) \\
&\leq |\overline{S}(T_n)| \cdot 2Nt^{-2},
\end{aligned}$$

where the last inequality holds since the matchings in $\overline{S}(T_n)$ are sufficiently scheduled (Lemma 5). Hence, we can obtain an upper bound as

$$\begin{aligned}
&\sum_{\mathbf{a} \in \overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t) = \mathbf{a} \mid \mathbf{x}(t-1) \in \overline{S}(T_n)\right) \mathbb{P}\left(\mathbf{x}(t-1) \in \overline{S}(T_n)\right) \\
&\leq \sum_{\mathbf{a} \in \overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t-1) = \mathbf{a}\right) \left[\frac{N!-1}{N!} + \frac{1}{N!} \cdot \mathbb{P}\left(\mathbf{x}(t) \in \overline{S}(T_n) \mid \mathbf{x}(t-1) = \mathbf{a}, \mathbf{o}(t) = \bar{\mathbf{o}}_t\right)\right] \\
&\leq \sum_{\mathbf{a} \in \overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t-1) = \mathbf{a}\right) \left[\frac{N!-1}{N!} + \frac{1}{N!} \cdot |\overline{S}(T_n)| \cdot 2Nt^{-2}\right],
\end{aligned} \tag{30}$$

for all $t \in (T_n, T_{n+1}]$. Letting $\alpha := \frac{N!-1}{N!}$ and combining (11), (12), and (30), we have

$$\begin{aligned}
&\sum_{\mathbf{a} \in \overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t) = \mathbf{a}\right) \\
&\leq \mathbb{P}\left(\mathbf{x}(t-1) \in \underline{S}(T_n)\right) + \left(1 + \frac{1}{N!}\right) |\overline{S}(T_n)| \cdot 2Nt^{-2} + \alpha \sum_{\mathbf{a} \in \overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t-1) = \mathbf{a}\right).
\end{aligned}$$

The inequality can be extended in a recursive manner as in the proof of Proposition 1 except for letting $A := \left(1 + \frac{1}{N!}\right) |\overline{S}(T_n)| \cdot 2N$, and we have, from (17), (18), and (19),

$$\begin{aligned}
&\sum_{t=T_n+1}^{T_{n+1}} \sum_{\mathbf{a} \in \overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t) = \mathbf{a}\right) \\
&\leq \frac{1}{1-\alpha} \mathbb{E}\left[\sum_{s=n+1}^{|\mathcal{M}|-1} l_{s,n+1}\right] + A \cdot \frac{1}{1-\alpha} \cdot \frac{\pi^2}{6} + \frac{1}{1-\alpha} \\
&= N! \left[\left(1 + \frac{1}{N!}\right) \cdot \frac{N\pi^2}{3} \cdot |\overline{S}(T_n)| + \mathbb{E}\left[\sum_{s=n+1}^{M} l_{s,n+1}\right] + 1\right].
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T')\right] \\
&= \sum_{\mathbf{a}\in\mathcal{M}^o} \sum_{t=1}^{T'} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right) \\
&= lM + \sum_{n=1}^{M-1} \sum_{t=T_n+1}^{T_{n+1}} \sum_{\mathbf{a}\in\overline{S}(T_n)} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right) \\
&\leq lM + N!\left[\left(1+\frac{1}{N!}\right)\frac{N\pi^2}{3}\sum_{n=1}^{M-1}|\overline{S}(T_n)| + \mathbb{E}\left[\sum_{n=1}^{M-1}\sum_{s=n+1}^{M} l_{s,n+1}\right] + (M-1)\right] \\
&\leq lM + N!(M-1)\left(\left(1+\frac{1}{N!}\right)\frac{(M-2)N\pi^2}{6} + l + 1\right),
\end{aligned}$$

where the last inequality comes from (20) and (21).

Therefore, with $M = |\mathcal{M}|-1$, we have

$$\sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T')\right] \leq (|\mathcal{M}|-1)(N!+1)\left(\frac{4N^2(N+1)\log T}{\Delta_{min}^2}+1\right) + C_1, \tag{31}$$

where $C_1 = N!(|\mathcal{M}|-2)\left(\left(1+\frac{1}{N!}\right)\frac{(|\mathcal{M}|-3)N\pi^2}{6}+1\right)$.

Further, we have $\sum_{\mathbf{a}\in\mathcal{M}^o}\mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)-\hat{\tau}_{\mathbf{a}}(T')\right] = \sum_{t=T'+1}^{T}\sum_{\mathbf{a}\in\mathcal{M}^o}\mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right)$, which can be upper bounded as (23). From the Lemma 5, we have $\mathbb{P}\left(V(\mathbf{a},\mathbf{I}(t))\geq V(\mathbf{a}^*,\mathbf{I}(t))\right)\leq 2Nt^{-2}$ for all $\mathbf{a}\in\mathcal{M}^o$, and from the Lemma 7, we have $\mathbb{P}\left(\mathbf{m}(t)\in\mathcal{M}^o\right)\leq \frac{N!-|\mathcal{O}^*|}{N!}+\frac{|\mathcal{O}^*|}{N!}2(|\mathcal{M}|-1)Nt^{-2}$. Note that $|\mathcal{O}^*|>0$ from the Lemma 3. Let $\alpha=\frac{N!-|\mathcal{O}^*|}{N!}$, we have

$$\begin{aligned}
&\sum_{t=T'+1}^{T}\sum_{\mathbf{a}\neq\mathbf{a}^*}\mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right) \\
&\leq \sum_{t=T'+1}^{T}\left(\left(1+\frac{|\mathcal{O}^*|}{N!}\right)\cdot(|\mathcal{M}|-1)\cdot 2Nt^{-2} + \alpha\sum_{\mathbf{a}\neq\mathbf{a}^*}\mathbb{P}\left(\mathbf{x}(t-1)=\mathbf{a}\right)\right) &(32)\\
&\overset{(A)}{\leq} \sum_{t=T'+1}^{T}\left(\left(1+\frac{|\mathcal{O}^*|}{N!}\right)\cdot(|\mathcal{M}|-1)\cdot 2N\sum_{s=0}^{T-t}\alpha^s t^{-2} + \alpha^{t-T'}\sum_{\mathbf{a}\neq\mathbf{a}^*}\mathbb{P}\left(\mathbf{x}(T')=\mathbf{a}\right)\right) \\
&\leq \frac{N!}{|\mathcal{O}^*|}\left(\left(1+\frac{|\mathcal{O}^*|}{N!}\right)\frac{(|\mathcal{M}|-1)N\pi^2}{3}+1\right), &(33)
\end{aligned}$$

where equality (A) can be obtained by extending (32) in a recursive manner.

Therefore, combining (31) and (33) together, we have

$$\sum_{\mathbf{a}\neq\mathbf{a}^*}\mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] \leq (|\mathcal{M}|-1)(N!+1)\left(\frac{4N^2(N+1)\log T}{\Delta_{min}^2}+1\right)+C_1+C_2, \tag{34}$$

where $C_1 = N!(|\mathcal{M}|-2)\left(\left(1+\frac{1}{N!}\right)\frac{(|\mathcal{M}|-3)N\pi^2}{6}+1\right)$ and $C_2=\frac{N!}{|\mathcal{O}^*|}\left(\left(1+\frac{|\mathcal{O}^*|}{N!}\right)\frac{(|\mathcal{M}|-1)N\pi^2}{3}+1\right)$.

(2) When $T' > T$: We omit the description of notations used in the proof, which is the same as in the proof of Proposition 1. We can write

$$\begin{aligned}
&\sum_{\mathbf{a}\in\mathcal{M}^o} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] \\
&= \sum_{\mathbf{a}\in\underline{S}(T)} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] + \sum_{\mathbf{a}\in\overline{S}(T)} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] \\
&\le l|\underline{S}| + \sum_{t=1}^{T}\sum_{\mathbf{a}\in\overline{S}} \mathbb{P}\left(\mathbf{x}(t)=\mathbf{a}\right) \\
&\overset{(A)}{\le} l|\underline{S}| + l|\overline{S}| + N!|\overline{S}| \left[\left(1+\frac{1}{N!}\right)\frac{(|\overline{S}|-1)N\pi^2}{6} + l + 1\right] \\
&= l(|\mathcal{M}|-1+N!|\overline{S}|) + N!|\overline{S}|\left[\left(1+\frac{1}{N!}\right)\frac{(|\overline{S}|-1)N\pi^2}{6}+1\right], \qquad (35)
\end{aligned}$$

where inequality (A) can be obtained as the proof of the case when $T' \le T$. From (35) and (34), we have

$$\sum_{\mathbf{a}\neq\mathbf{a}^*} \mathbb{E}\left[\hat{\tau}_{\mathbf{a}}(T)\right] \le (|\mathcal{M}|-1)(N!+1)\left(\frac{4N^2(N+1)\log T}{\Delta_{min}^2}+1\right) + C_1 + C_2,$$

where $C_1 = N!(|\mathcal{M}|-2)\left(\left(1+\frac{1}{N!}\right)\frac{(|\mathcal{M}|-3)N\pi^2}{6}+1\right)$ and $C_2 = \frac{N!}{|\mathcal{O}^*|}\left(\left(1+\frac{|\mathcal{O}^*|}{N!}\right)\frac{(|\mathcal{M}|-1)N\pi^2}{3}+1\right)$. ■

Lemma 1 and Proposition 2 lead to the following result.

Theorem 2 *Under GYRO, the expected total regret $\mathcal{R}_{GYRO}(T)$ by time T is upper bounded as*

$$\mathcal{R}_{GYRO}(T) \le \Delta_{max}^*\left((|\mathcal{M}|-1)(N!+1)\cdot\left(\frac{4N^2(N+1)\log T}{\Delta_{min}^2}+1\right)+C_1+C_2\right), \qquad (36)$$

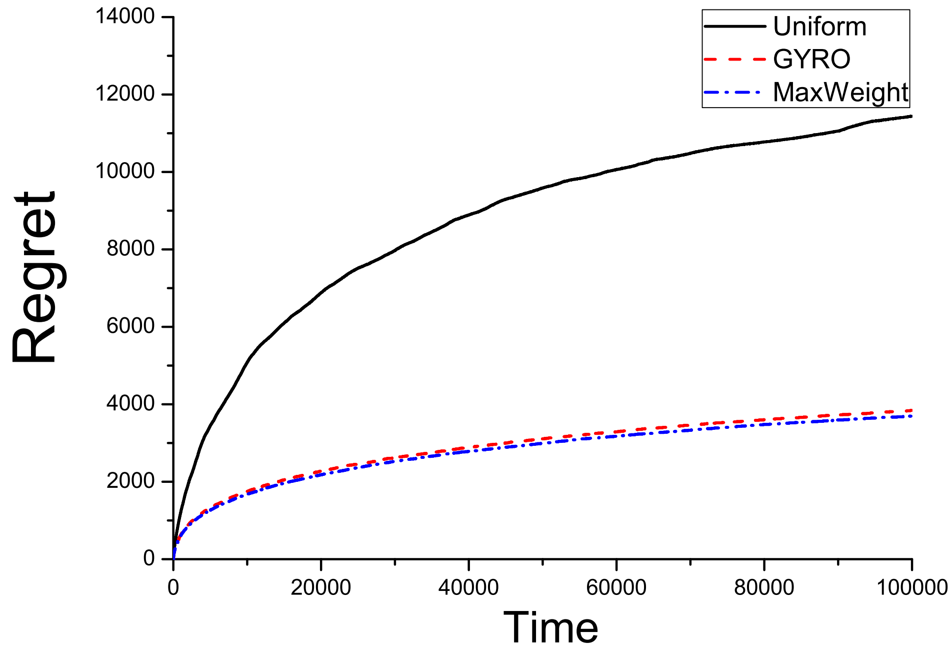
where $C_1 = N!(|\mathcal{M}|-2)\left(\frac{(|\mathcal{M}|-3)N\pi^2}{6}\left(1+\frac{1}{N!}\right)+1\right)$, *and* $C_2 = \frac{N!}{|\mathcal{O}^*|}\left(\frac{(|\mathcal{M}|-1)N\pi^2}{3}\left(1+\frac{|\mathcal{O}^*|}{N!}\right)+1\right)$.

The theorem shows that regret $\mathcal{R}_{GYRO}(T)$ of *GYRO* is upper bounded by $O(\log T)$, which is asymptotically optimal.

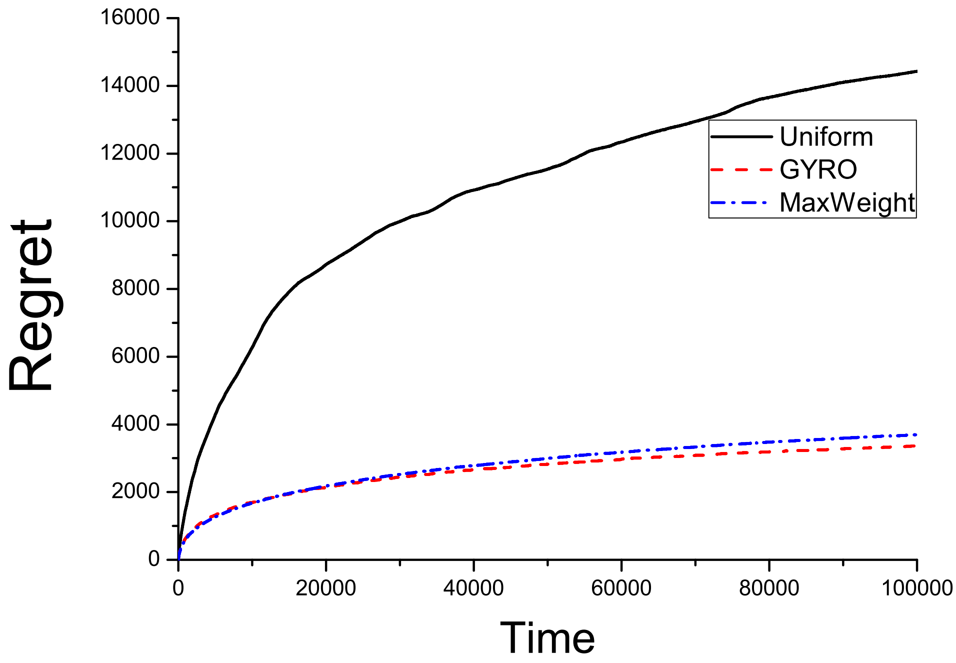
V Simulation results

We have shown that under our algorithms of uniform sampling and *GYRO*, the expected regret grows logarithmically with respect to time. In this section, we demonstrate the performance of our algorithms through simulations. We consider $N = 5$ users and $K = 10$ channels. If user-channels pair (i, k) is played, then user i receives a binary reward drawn from Bernoulli distribution with mean $\mu_{i,k}$ which is drawn uniformly at random between [0,1]. Simulation runs for $T = 10^5$ time slots, and results are averaged over 20 repetitions.

We compare our algorithms (i.e., uniform sampling labeled as Uniform in Fig. 4 and *GYRO*) with a well-known MaxWeight that solves the maximum weighted bipartite matching problem



(a) Complete bipartite graph.



(b) Incomplete bipartite graph.

Figure 4: Average of total regrets with respect to time slots.

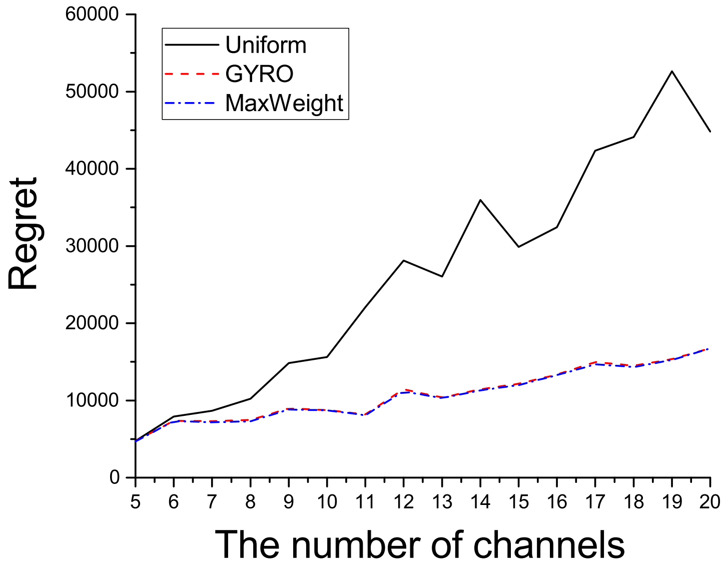


Figure 5: Average of total regrets at $T = 10^5$ with respect to the number of channels.

at each time slot, i.e., $\mathbf{x}(t) \in \arg\max_{\mathbf{a}\in\mathcal{M}} V(\mathbf{a}; \mathbf{I}(t))$. MaxWeight can be implemented using brute-force search or Hungarian algorithm [22] whose computational complexities are $O(K^N)$ and $O((N+K)^3)$, respectively. Note that the complexities of uniform sampling algorithm and *GYRO* are $O(N)$ and $O(K)$, respectively.

We consider two bipartite graph: one complete bipartite graph as shown in Fig. 1 (i.e., there are NK edges with $\mu_{i,k} > 0$), and one incomplete bipartite graph where each user i has 6 channels with $\mu_{i,k} > 0$ out of 10 channels. Fig. 4 shows the total regrets of three algorithms over time. In Fig. 4(a), the results from the complete graph are shown, and in Fig. 4(b), the results from the incomplete graph are shown. As expected, the regret grows logarithmically over time. Further, in both cases, the regret of uniform sampling is distinctly worse than *GYRO*. Interestingly, in some cases, *GYRO* outperforms MaxWeight, in which case MaxWeight explores non-optimal matchings more frequently than *GYRO*.

Now we show that the performance gaps between uniform sampling and *GYRO* enlarge with respect to the size of network. The number of users is fixed with $N = 5$, and the number of channels varies from $K = 5$ to 20. Simulation runs for $T = 10^5$ time slots, and the regret is captured at $T = 10^5$. The result is averaged over 20 repetitions. As seen in Fig. 5, *GYRO* distinctly outperforms uniform sampling, and it is comparable with MaxWeight. Note that since the regret is affected by both channel statistics (i.e., $\mu_{i,k}$) and the number of channels, the optimal regret does not linearly increase with the number of channels.

VI Conclusion

In this paper, we develop low-complexity learning algorithms for opportunistic spectrum access in multi-user multi-channel cognitive radio networks, and show that they achieves the expected total regret growing at most logarithmically with respect to time. Through numerical simulations, we verify our results, and compare the performance with the well-known maximum weighted matching algorithm at each time slot. The idea of reducing complexity can be applied to other learning problems and remains as a future work.

References

[1] P. Kolodzy and I. Avoidance, "Spectrum policy task force," *Federal Commun. Comm., Washington, DC, Rep. ET Docket*, vol. 40, no. 4, pp. 147–158, 2002.

[2] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[3] S. Kang, C. Joo, J. Lee, and N. B. Shroff, "Pricing for Past Channel State Information in Multi-Channel Cognitive Radio Networks," *IEEE Transactions on Mobile Computing (TMC)*, to appear.

[4] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM journal on computing*, vol. 32, no. 1, pp. 48–77, 2002.

[7] W. Jouini, D. Ernst, C. Moy, and J. Palicot, "Multi-armed bandit based policies for cognitive radio's decision making issues," in *International Conference on Signals, Circuits and Systems (SCS)*, 2009.

[8] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 1987.

[9] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *International Conference on Machine Learning*, 2013, pp. 151–159.

[10] 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 (TON)*, vol. 20, no. 5, pp. 1466–1478, 2012.

[11] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[12] Y. Gai and B. Krishnamachari, "Decentralized online learning algorithms for opportunistic spectrum access," in *GLOBECOM*, 2011.

[13] 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.

[14] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *IEEE Symposium on New Frontiers in Dynamic Spectrum*, 2010.

[15] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[16] O. Avner and S. Mannor, "Multi-user lax communications: a multi-armed bandit approach," in *Computer Communications, IEEE INFOCOM 2016*.

[17] J. Ni, B. Tan, and R. Srikant, "Q-csma: Queue-length-based csma/ca algorithms for achieving maximum throughput and low delay in wireless networks," *IEEE/ACM Transactions on Networking (ToN)*, vol. 20, no. 3, pp. 825–836, 2012.

[18] J. Ryu, C. Joo, N. B. Shroff, Y. Choi, *et al.*, "Dss: Distributed sinr-based scheduling algorithm for multihop wireless networks," *IEEE Transactions on Mobile Computing (TMC)*, vol. 12, no. 6, pp. 1120–1132, 2013.

[19] J. Edmonds, "Paths, trees, and flowers," *Canadian Journal of mathematics*, vol. 17, no. 3, pp. 449–467, 1965.

[20] D. Pollard, "Convergence of stochastic processes," 2012.

[21] J. A. Bondy and U. S. R. Murty, *Graph theory with applications.* Elsevier North-Holland, 1976.

[22] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 2, no. 1-2, pp. 83–97, 1955.

Acknowledgements

I would like to express my gratitude to my advisor, Prof. Changhee Joo, for his valuable advice and continuous support for my master's thesis. I would never have been able to finish my thesis without his guidance and help.

I would like to thank my parents, Yunho Kang and Jinbong Lee, for their endless love and support. They always stand by me and encourage me.

I would like to thank my elder sister, Jiseon. She always listens to me and understands me with a sympathetic ear. I would also like to thank my younger brother, Jaemin.

I would like to thank my colleague from Network Research Laboratory, Minho Kim, Seunghyun Lee, Fatih Akturk, Jihyeon Yun, Abylay Satybaldy, and Taewon Min.

I would like to thank Myeungun Kim, who is my good friend. It would have been a lonely campus without her.