Fair Online Learning for Restless Bandits

***Abstract*—The Restless Multi-armed Bandit (RMAB) model is widely used for resource allocation and scheduling in communication networks. In RMAB, a decision maker selects a policy that activates M out of N arms to maximize the overall reward of the system. Traditional RMAB solutions (e.g., Whittle index policy) require known system dynamics, which is often unavailable in practice. Moreover, existing policies often ignore fairness (a key design concept in communication networks) among the resources allocated to all agents: Agents with higher rewards get resources frequently, while agents with low rewards may endure long waits. In this study, we develop an online algorithm for RMABs with unknown system dynamics that maximizes the sum of time-averaged rewards for all agents over finite horizon while satisfying fairness constraints. We propose an Online Fair Maximum Gain First (Online-FMGF) policy using three fundamental tools: (i) Relaxation and Lagrangian decomposition, (ii) Upper Confidence Bound (UCB) approach, and (iii) Two-time scale update method for Lagrange multipliers. Our algorithm achieves sub-linear regret $O(\sqrt{K\log K})$ on the number of episodes K, demonstrates computational efficiency, and does not need to satisfy indexability-a major limitation for Whittle index policies. Numerical results validate that Online-FMGF achieves better performance relative to other baselines.**

I. Introduction

Robust real-time monitoring and control systems are prevalent across a wide array of interconnected domains such as networked control systems, cyber-physical systems, and large-scale Internet of Things (IoT) deployments. In these systems, multiple agents (e.g., robots, UAVs, sensors, surveillance cameras, etc) continuously observe data and send to a decision-making device. For example, in a military communication network, a ground station tracks real-time data from remote agents (UAVs, ships, jets) to maintain situational awareness, make informed decisions, and issue control commands. However, due to communication and energy resource constraints, the decision-maker cannot allocate resources to all agents simultaneously.

Many resource allocation problems can be modeled as RMABs. In this framework, each agent is represented as an arm, and a decision-maker activates M out of N arms to maximize the overall reward or minimize the total cost of the system. Each arm is modeled as a Markov Decision Process (MDP) that evolves stochastically according to whether the arm is selected or not. Solving RMAB problems is significantly challenging and is PSAPCE-hard [1]. In 1988, an efficient method was proposed to solve RMABs, known as the Whittle index policy, which needs to satisfy an indexability condition [2]. Later, this policy was proven to be asymptotically optimal [3]. Since then, many resource allocation and scheduling problems in communication networking, such as network utility maximization, age penalty minimization, remote estimation [4]–[10] have been formulated as RMABs. One major challenge in developing a Whittle index policy is to establish indexability, which is often difficult to achieve in practice. Therefore, non-indexable scheduling policies were also studied in recent years [11]–[17]. An asymptotically optimal Maximum Gain First Policy was proposed in [13]–[17] with known system dynamics. Although most existing studies assume known system dynamics, these are often unknown in practice. Hence, recently online learning methods for RMABs have been developed [18]–[24] that adaptively learn the system dynamics.

However, traditional RMAB often overlook fairness. If fairness is not considered, it may happen that some agents with a higher reward will get resources frequently and an agent with a lower reward may wait for a long time. Motivated by this, we develop online learning algorithms to solve RMABs with fairness constraints and unknown system dynamics by utilizing Lagrangian approach, UCB method, and two time-scale updating for the Lagrange multipliers.

The summary of our contributions are discussed below.

- We develop a UCB based Online-FMGF policy that does not need to satisfy any indexability condition (see Algorithm 1). We leverage relaxation and Lagrangian approach to decouple the problem into multiple independent MDPs. Using a two-time scale update method, we compute the optimal Lagrange multipliers associated with our policy. Unlike [22]–[24], our approach addresses unknown transition probabilities and rewards within a finite horizon system with additional fairness constraints and no requirements on indexability.
- A key advantage of Online-FMGF policy compared to other online learning policies considering fairness [25] is that we can decouple the multi-agent problem. The authors in [25] consider fairness constraints in an infinite horizon system. However, the developed convex optimization-based Extended Linear Program (ELP) could not be decomposed, and hence, the result is difficult to apply in large scale systems. In contrast, our proposed Online-FMGF policy is developed by analyzing the Lagrangian dual problem, which allows the problem to be decoupled.
- We achieve a further simplification of the Online-FMGF policy, where we do not need to estimate any optimistic transition probabilities (Algorithm 2). Algorithm 1 requires solving sequential constrained convex optimization problems in every time-slot of each episode to find the optimistic transition probabilities. To address this difficulty, we derive an upper bound of the optimistic action-value functions (see Lemma 2). Consequently, we

do not need to compute any transition probabilities after this simplification and the developed simplified algorithm is computationally more efficient than Algorithm 1 (see Algorithm 2).

- The Online-FMGF policy (Algorithm 1) and the Simplified Online-FMGF policy (Algorithm 2) both achieve sublinear regret $O(\sqrt{K \log K})$ in the number of episodes K (See Theorem 1 and Theorem 2). These results provide the first sub-linear regret guarantees for RMABs incorporating fairness without employing any LP formulations. Our results extend earlier online learning policies for RMABs [22]–[24] by providing a new online gain index-based policy and by adding fairness constraints, while still achieves similar regret performance as in [22]–[24].
- We compare the performance of our proposed policy with greedy policy, uniform randomized policy, UCB+Lyapunov-based policy, and Extended Linear Program (ELP)-based policy for an interference-aware throughput maximization system, and for a Land Mobile Satellite system. The proposed policy outperforms the other policies in both systems.

II. System Model

We consider an RMAB setting composed of a centralized scheduler and N agents. Each agent is modeled as an arm $n \in \{1, 2, \ldots, N\}$ and each arm n is described as an MDP. All arms share the same state space $\mathcal{S}$ and action space $\mathcal{A} \in \{0, 1\}$. The state transition probability $P_n : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ and reward $r_n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ can be different across arms.

At each time-slot t, the state of the RMAB is denoted as $\mathbf{s}_t = \{s_1(t), s_2(t), \ldots, s_N(t)\} \in \mathcal{S}^N$, where $s_n(t)$ is the state of arm n. The decision maker observes $\mathbf{s}_t$ at each time-slot t and selects a set of arms $\mathbf{a}_t = \{a_1(t), a_2(t), \ldots, a_N(t)\} \in \mathcal{A}^N$ to activate, where $a_n(t)$ is the action of arm n. Each MDP associated with each arm has two actions: active and passive. If arm n is selected for activation at time-slot t, then the action is active, i.e., $a_n(t) = 1$; otherwise, arm n is passive. Due to resource constraints, the decision maker can choose at most M agents at any time-slot t, i.e., $\sum_{n=1}^{N} a_n(t) \leq M$ for all t.

The system operates within a finite time horizon T and the initial state is $\mathbf{s_1} \in \mathcal{S}^{\mathbf{N}}$. Each arm generates a reward $r_n(s_n(t), a_n(t))$ at every time-slot t, where we consider a bounded reward function for all $s_n(t) = s$ and $a_n(t) = a$, i.e., there exists a $0 < D < \infty$ such that $|r_n(s, a)| \leq D$ for all (s, a). The state transition probability at time-slot t is denoted by $P_n(s_n(t+1)|s_n(t), a_n(t)) \in [0, 1]$, where $s_n(t+1)$ is the next state after taking action $a_n(t)$ at time-slot t. Let $\mathbf{P} = \{P_n\}_{n=1}^{N}$ be the set of all transition probabilities and $\mathbf{r} = \{r_n\}_{n=1}^{N}$ be the set of all rewards. In our system, the transition probabilities and the rewards are unknown to the centralized decision maker.

III. Problem Formulation: Restless Multi-armed Bandit with Fairness Constraints

Let π represent a policy that maps any state from $\mathcal{S}^N$ to an action in $\mathcal{A}^N$. Our goal is to find the policy π that maximizes the sum of the time-averaged expected rewards of the N agents over a finite time-horizon T:

$$\max_{\pi} \frac{1}{T} \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} \sum_{n=1}^{N} r_n(s_n(t), a_n(t)) \right] \tag{1}$$

$$\text{s.t.} \sum_{n=1}^{N} a_n(t) \leq M, t = 1, 2, \ldots, T, \tag{2}$$

$$\text{s.t.} \frac{1}{T} \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} a_n(t) \right] \geq \beta_n, n = 1, 2, \ldots, N, \tag{3}$$

where $0 \leq \beta_n \leq 1$ is the required minimum fraction of time arm n is activated.

The reward function $r_n(s_n(t), a_n(t))$ in (1) is quite general and thus, applicable across various wireless communication network problems. For example, in network resource allocation, $r_n(s_n(t), a_n(t))$ can be throughput, bit rate, SNR, etc. In remote estimation, it can be the autocorrelation function, mutual information, etc. Conversely, the negative of $r_n(s_n(t), a_n(t))$ can serve as a cost function for metrics such as estimation error, age of information, non-linear functions of age of information, etc.

Problem (1)-(3) is an RMAB with fairness constraints. Finding optimal solutions for RMAB problems is significantly challenging [1]. The additional N fairness constraints make problem (1)-(3) even more challenging. One efficient approach to solve RMAB problems is to develop a Whittle index policy, which is asymptotically optimal under indexability [2], [3]. However, indexability is often very difficult to establish in real-world applications. Therefore, we aim to solve the problem (1)-(3) for general RMABs (whether indexable or not). Utilizing Lagrangian dual decomposition, we decompose problem (1)-(3) into N independent per-arm problems and study an FMGF policy for the decoupled problem with known system dynamics (Sec. III-A). Analyzing the FMGF policy for known system dynamics, we are able to find an Online-FMGF policy that is not required to satisfy indexability (Sec. V-A).

A. Relaxation and Lagrangian Decomposition

We first apply Lagrange multipliers to the fairness constraints and obtain the following Lagrangian dual problem associated with Lagrange multipliers $\lambda_n, n = 1, 2, \ldots, N$:

$$\max_{\pi} \frac{1}{T} \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} \sum_{n=1}^{N} r_n(s_n(t), a_n(t)) + \lambda_n(a_n(t) - \beta_n) \right] \tag{4}$$

$$\text{s.t.} \sum_{n=1}^{N} a_n(t) \leq M, t = 1, 2, \ldots, T. \tag{5}$$

Next, we relax constraint (5) as

$$\frac{1}{T} \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} \sum_{n=1}^{N} a_n(t) \right] \leq M, t = 1, 2, \ldots, T. \tag{6}$$

After relaxation, we apply another Lagrange multiplier μ to the relaxed constraint (6). The Lagrangian associated with

Lagrange multipliers $\boldsymbol{\lambda} = \{\lambda_n\}_{n=1}^N$ and μ can be written as

$$L(\pi^*, \mathbf{P}, \mathbf{r}, \mu, \boldsymbol{\lambda}, \mathbf{s}_1) = \max_{\pi} \frac{1}{T} \mathbb{E}_\pi \Bigg[\sum_{t=1}^{T} \sum_{n=1}^{N} r_n(s_n(t), a_n(t)) + (\lambda_n - \mu) a_n(t) - \sum_{n=1}^{N} \lambda_n \beta_n + \mu M \Bigg]. \quad (7)$$

Because the terms $1/T \sum_{t=1}^{T} \sum_{n=1}^{N} \lambda_n \beta_n$ and $1/T \sum_{t=1}^{T} \sum_{n=1}^{N} \mu M$ are constants and do not affect the policy π, they can be removed. Given $\boldsymbol{\lambda}$ and μ, we can decouple problem (7) into N per-arm problems, where the per-arm problem associated with arm n is given by

$$L(\pi_n^*, P_n, r_n, \mu, \lambda_n, s_n(1)) = \max_{\pi_n} \frac{1}{T} \mathbb{E}_{\pi_n} \Bigg[\sum_{t=1}^{T} r_n(s_n(t), a_n(t)) + (\lambda_n - \mu) a_n(t) \Bigg], \quad (8)$$

where $\pi_n : \mathcal{S} \to \mathcal{A}$ represents the policy for arm n.

We first solve the decoupled problem (8) for given λ_n and μ. Following the solution of (8), we will develop a solution for the original multi-agent problem (1)-(3). Then we obtain the optimal Lagrange multipliers from $(\mu^*, \lambda_1^*, \lambda_2^*, \ldots, \lambda_N^*) = \operatorname{argmin}_{\mu, \lambda_1, \lambda_2, \ldots, \lambda_N} L(\pi^*, \mathbf{P}, \mathbf{r}, \mu, \boldsymbol{\lambda}, \mathbf{s}_1)$, where $L(\pi^*, \mathbf{P}, \mathbf{r}, \mu, \boldsymbol{\lambda}, \mathbf{s}_1)$ is the optimal value of (7).

B. Solution to the Decoupled Problem (8)*: Fair Maximum Gain First (FMGF) Policy*

Problem (8) can be solved by dynamic programming [26]. The Bellman optimality equation of the MDP (8) for given state $s_n(t) = s$ and action $a_n(t) = a$ at time-slot t is given by

$$V_{n,t}(s) = \max_{a \in \mathcal{A}} Q_{n,t}(s, a), \quad (9)$$

where $Q_{n,t}(s, a)$ is the action-value function defined as

$$Q_{n,t}(s,a) = (\lambda_n - \mu) a + r_n(s,a) + \sum_{s'} P_n(s'|s,a) V_{n,t+1}(s'). \quad (10)$$

The value function can be obtained by using the backward induction method [26]. WLOG, we assume $V_{n,T+1}(s) = 0$ for all s. The backward induction method for computing the value function $V_{n,t}(s)$ for $t = T, T-1, \ldots, 1$ is given by

$$V_{n,t}(s) = \max_{a} \Bigg\{ r_n(s,a) + (\lambda_n - \mu) a + \sum_{s'} P_n(s'|s,a) V_{n,t+1}(s') \Bigg\}. \quad (11)$$

Arm n is activated if $Q_{n,t}(s, 1) > Q_{n,t}(s, 0)$. An equivalent policy is to utilize the gain index, defined as [13], [14], [16]

$$\alpha_{n,t}(s) = Q_{n,t}(s, 1) - Q_{n,t}(s, 0), \quad (12)$$

which is the difference of two action-value functions. Following the gain index-based policy, arm n is activated if $\alpha_{n,t}(s) > 0$. For problem (8), policies utilizing either the action-value functions or the gain index are equivalent to each other. However, for the multi-agent problem, these policies may not be equivalent.

Because we are solving an RMAB with fairness constraints, we call this policy a Fair Maximum Gain First (FMGF) policy. This policy selects at most M arms with the highest gain indices to activate at every time-slot t.

IV. Online Learning Setting

In our online learning environment, the centralized decision maker interacts with N arms through K episodes. In each episode k, the decision maker utilizes the observations for H time-slots, where H is the length of the horizon in each episode. A key challenge here is that the decision maker has no prior knowledge of the actual transition probabilities $\mathbf{P}$ and rewards $\mathbf{r}$. In every episode k, the decision maker selects a policy π^k starting from the initial state $\mathbf{s}_1$. Subsequently, it utilizes the observations up to H time-slots to iteratively estimate the underlying transition probabilities and rewards.

Because finding the optimal policy for RMAB problems is generally intractable [1], it is quite challenging to utilize the optimal policy of RMAB to evaluate the performance of an online learning algorithm. The authors in [23], [24] utilize Lagrangian relaxed problem to evaluate the performance of an online policy. Motivated by [23], [24], we adopt Lagrangian (7) to evaluate the performance of our online policy π^k in episode k. When the system size is large, Lagrangian relaxation provides an asymptotic approximation to the performance of the optimal policy [3], [27]; therefore, we use it as a benchmark for regret. Stronger definition of regret will be considered in future.

Definition 1. Regret. Given the actual transition probabilities $\mathbf{P}$ and actual rewards $\mathbf{r}$, the regret of policy π^k in episode k is given by

$$\text{Reg}(k) = L(\pi^*, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^k, \boldsymbol{\lambda}^k, \mathbf{s}_1), \quad (13)$$

where π^* is the optimal policy for the Lagrangian (7), μ^* and λ^* minimizes the Lagrangian defined in (7), and μ^k and $\boldsymbol{\lambda}^k = (\lambda_n^k)_{n=1}^N$ are the Lagrange multipliers in episode k. We assume that the Lagrange multipliers μ^k and λ_n^k are bounded, such as $\mu^k \leq A$, and $\lambda_n^k \leq G$ for all $n = 1, 2, \ldots, N$, where $0 < A < \infty$ and $0 < G < \infty$ are positive constants. Using (13), we get the total regret as follows:

$$\text{Reg}(K) = \sum_{k=1}^{K} \text{Reg}(k). \quad (14)$$

V. Online Policy Design for Solving (1)-(3)

We solve the online learning problem in two steps: (i) We develop a UCB-based Online-FMGF policy that does not need to satisfy any indexability condition (Section V-A); (ii) Next, we analyze the regret to show that the developed algorithm achieves sublinear regret (Section V-B).

A. Design of Online Policy

We utilize UCB-based online learning to provide an algorithm for the Online-FMGF policy. To compute the gain index (12), we need to know the action-value functions and hence, the transition probabilities $\mathbf{P}$ and rewards $\mathbf{r}$. However,

the transition probabilities $\mathbf{P}$ and rewards $\mathbf{r}$ are unknown in our model. Therefore, we develop an online learning method in Algorithm 1.

1) Confidence Bound for Transition Probabilities and Rewards: For every arm n and every episode k, we maintain variables $B_n^k(s,a,s')$ that represent the number of transitions from state s to state s' under action a within the last k episodes. For a small given constant $\epsilon > 0$, we define the confidence radius as follows:

$$\rho_n^k(s,a) = \sqrt{\frac{2|\mathcal{S}|\log(2|\mathcal{S}||\mathcal{A}|N\frac{k^2}{\epsilon})}{\max(1, B_n^k(s,a))}}, \tag{15}$$

where $B_n^k(s,a) = \sum_{s'\in\mathcal{S}} B_n^k(s,a,s')$ is the number of transitions to state-action pair (s,a) for arm n within episode k.

The empirical transition probability is given by

$$\hat{P}_n^k(s'|s,a) = \frac{B_n^k(s,a,s')}{B_n^k(s,a)} \tag{16}$$

and the empirical reward is given by

$$\hat{r}_n^k(s,a) = \frac{\sum_{k'=1}^{k-1}\sum_{t=1}^{H} r_n^{k'}(s,a)\mathbf{1}(s_n^{k'}(t)=s, a_n^{k'}(t)=a)}{\max(1, B_n^{k-1}(s,a))}.$$

The confidence balls of possible transition probabilities and rewards are given by

$$\mathcal{B}_p^k = \left\{P_n^k \Big| \sum_{s'\in\mathcal{S}} \left|P_n^k(s'|s,a) - \hat{P}_n^k(s'|s,a)\right| \le \rho_n^k(s,a), \forall n, s, a\right\}, \tag{17}$$

$$\mathcal{B}_r^k = \left\{r_n^k | r_n^k(s,a) - \hat{r}_n^k(s,a) = \rho_n^k(s,a), \forall n, s, a\right\}. \tag{18}$$

2) Online Algorithm for FMGF Policy: Using an optimistic approach that implies choosing the most suitable estimates for unknown parameters from a confidence set, we estimate the transition probabilities and rewards. The optimistic reward is given by

$$r_n^k(s,a) = \min\{\hat{r}_n^k(s,a) + \rho_n^k(s,a), D\}. \tag{19}$$

Then we choose the optimistic transition probability P_n^k for each arm n in episode k that maximizes the value function within the confidence radius $\rho_n^k(s,a)$. The optimization problem is given by

$$\max_{P_n^k\in\mathcal{B}_p^k} V_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s), \tag{20}$$

$$\text{s.t. } V_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s) = \max_{a\in\mathcal{A}} Q_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s,a), \tag{21}$$

where $Q_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s,a)$ is the action-value function given as

$$Q_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s,a) = (\lambda_n^k - \mu^k)a + r_n^k(s,a) + \sum_{s'} P_n^k(s'|s,a) V_{n,t+1}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s'). \tag{22}$$

Then we get the gain index for each arm n in episode k as

$$\alpha_{n,t}^{\mu^k, \lambda_n^k}(s) = Q_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s,1) - Q_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s,0). \tag{23}$$

Algorithm 1 Online-FMGF Policy

Input: N arms, constraint M, fairness constraints β_n for all n, episode length H.
Initialize number of visits $B_n^1(s,a,s') = 0$ for all s, a, s'.
Initialize $\mu^{(1)}$ and $\lambda_n^{(1)}$ for all $n = \{1, 2, \ldots, N\}$.
for $k = 1, 2, \ldots$ **do**
  Reset $t = 1$ and $\mathbf{s} = \mathbf{s}_1$.
  Compute optimistic reward for each arm by using (19).
  Compute transition probabilities for each arm by solving (20)-(21).
  for $t = 1, 2, \ldots$ **do**
    Compute gain indices from (23) for each arm.
    Activate M arms with the highest gain index.
    Observe transitions (s, a, s').
    Update visits $B_n^k(s,a,s')$, empirical mean $\hat{\mathbf{P}}^k$, empirical reward $\hat{\mathbf{r}}^k$, confidence regions $\mathcal{B}_p^k$ and $\mathcal{B}_r^k$.
    Every C steps update μ^k from (24).
  end for
  Update λ_n^{k+1} for all $n = 1, 2, \ldots, N$ from (26).
end for

The algorithm to compute the Online-FMGF policy is provided in Algorithm 1. At every episode k, this policy selects M arms with the highest positive gain index (23). A Lagrangian-based online Whittle index policy was studied in [23] for an infinite horizon problem without fairness constraints. The authors in [23] require solving two sequential optimization problems: first to compute the optimistic transition probabilities, next to compute the Whittle indices. Unlike [23], Algorithm 1 requires solving the one optimization problem (20)-(21) to obtain the optimistic transition probabilities from which the gain index (23) is directly computed using optimistic action-value functions in (22). Furthermore, our method significantly reduces complexity compared to the occupancy measure-based Extended Linear Program (ELP) [25]. The ELP in [25] cannot be decomposed, hence, [25] requires finding the occupancy measures without decomposing the ELP into sub-problems. Therefore, the scalability to large multi-agent systems is limited. In contrast, our method permits us to decompose the original problem into per-arm problems for each arm n. Using the solution to the per-arm problem, we get the solution to the multi-agent problem.

The Lagrange multipliers μ^k and λ_n^k in Algorithm 1 are updated at two different time-scales by using the stochastic sub-gradient descent method [28], [29].

The Lagrange multiplier μ^k is updated every C time-slots within the k-th episode. The update rule is given by

$$\mu^k(j+1) = \max\left\{\mu^k(j) + \frac{\gamma_\mu}{j}\left(\frac{1}{C}\sum_{n=1}^{N}\sum_{t=jC+1}^{(j+1)C} a_{\mu,n}^k(t) - M\right), 0\right\}, \tag{24}$$

where γ_μ is the learning parameter and action $a_{\mu,n}^k(t)$ is

obtained by solving the following decoupled problem:

$$a^k_{\mu,n}(t) = \arg\max_a Q^{P^k_n, r^k_n, \mu^k, \lambda^k_n}_{n,t}(s, a). \tag{25}$$

The Lagrange multipliers $(\lambda^{k+1}_n)^N_{n=1}$ are computed after completing the k-th episode. The update rule is given by

$$\lambda^{k+1}_n = \max\left\{\lambda^k_n - \frac{\gamma_\lambda}{k}\left(\frac{1}{H}\sum_{n=1}^{N}\sum_{t=1}^{H} a^k_n(t) - \beta_n\right), 0\right\}, \tag{26}$$

where γ_λ is the learning parameter and action $a^k_n(t)$ is obtained by solving the multi-agent problem.

The update of μ^k operates at a faster time-scale for given $(\lambda^k_n)^N_{n=1}$. Then, $(\lambda^k_n)^N_{n=1}$ is updated at a slower time-scale. This is because simultaneously converging μ^k and $(\lambda^k_n)^N_{n=1}$ is challenging, and it may happen that none of the multipliers converge to the optimal value. Hence, we first update μ^k for given $(\lambda^k_n)^N_{n=1}$. After μ^k converges, we update $(\lambda^k_n)^N_{n=1}$. This two-time scale does not affect the regret bound, yet it is essential for the convergence of μ^k and λ^k_n in the simulation.

An interesting finding is that the actions $a^k_{\mu,n}(t)$ and $a^k_n(t)$ in (24) and (26) are different and are associated with two different policies. We first solve the decoupled problem to get $a^k_{\mu,n}(t)$ for given $(\lambda_n)^N_{n=1}$. Next, we get $a^k_n(t)$ by solving the multi-agent problem following policy π^k.

B. Regret Bound for Algorithm 1

To bound the regret, we first show that with high probability the true transition $\mathbf{P}$ lies within the confidence ball $\mathcal{B}^k_p$ in (17) and the true reward $\mathbf{r}$ lies within the confidence ball $\mathcal{B}^k_r$ in (18).

Proposition 1. *Given $\epsilon > 0$ and $k \geq 1$, we get that* $\Pr(\mathbf{P} \in \mathcal{B}^k_p) \geq 1 - \frac{\epsilon}{k^2}$ *and* $\Pr(\mathbf{r} \in \mathcal{B}^k_r) \geq 1 - \frac{2}{(|\mathcal{S}||\mathcal{A}|N)^{2|\mathcal{S}|-1}}(\frac{\epsilon}{k^2})^{2|\mathcal{S}|}$.

Proof sketch. Following [30, Theorem 2.1], we bound the L_1-deviation of the actual transition and the empirical transition. Moreover, by using the Chernoff-Hoeffding inequality [31], we bound the actual reward and the empirical reward. □

Proposition 1 implies that for each episode k, two confidence bounds can be obtained within which the actual transition and actual reward lie with high probability. Next, we utilize Proposition 1 to bound the regret.

Lemma 1. *Given initial state $(s_1(1), s_2(1), \ldots, s_N(1))$, the following holds with probability $1 - \epsilon$:*

$$\mathrm{Reg}(k) \leq \sum_{n=1}^{N} V^{P^k_n, r^k_n, \mu^k, \lambda^k_n}_{n,1}(s_n(1)) - V^{P_n, r_n, \mu^k, \lambda^k_n}_{n,1}(s_n(1)), \tag{27}$$

where P^k_n and r^k_n are the transition probability and reward in episode k, and P_n and r_n are the actual transition probability and reward of arm n.

Proof sketch. When the confidence bound holds, we can show that

$$\begin{aligned}
\mathrm{Reg}(k) =& L(\pi^*, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^k, \boldsymbol{\lambda}^k, \mathbf{s}_1) \\
=& \sum_{n=1}^{N} V^{P_n, r_n, \mu^*, \lambda^*_n}_{n,1}(s_n(1)) - V^{P_n, r_n, \mu^k, \lambda^k_n}_{n,1}(s_n(1)) \\
\leq& \sum_{n=1}^{N} V^{P_n, r_n, \mu^k \lambda^k_n}_{n,1}(s_n(1)) - V^{P_n, r_n, \mu^k \lambda^k_n}_{n,1}(s_n(1)) \quad (28) \\
\leq& \sum_{n=1}^{N} V^{P^k_n, r^k_n, \mu^k \lambda^k_n}_{n,1}(s_n(1)) - V^{P_n, r_n, \mu^k \lambda^k_n}_{n,1}(s_n(1)), \quad (29)
\end{aligned}$$

where (28) holds since μ^* and λ^* minimize the Lagrangian L in (7). To prove (29), we first show by backward induction that for any $t = H, H-1, \ldots, 1$ the following holds

$$V^{P_n, r_n, \mu^k, \lambda^k_n}_{n,t}(s) \leq V^{P^k_n, r^k_n, \mu^k, \lambda^k_n}_{n,t}(s). \tag{30}$$

Then, (29) holds due to (20)-(21) and the fact that $\mathbf{P} \in \mathcal{B}^k_p$, $\mathbf{r} \in \mathcal{B}^k_r$. □

Using Lemma 1, we bound the regret for K episodes, to obtain the following theorem.

Theorem 1. *With probability $1 - \epsilon$, the cumulative regret of Algorithm 1 in all K episodes is given by*

$$\mathrm{Reg}(K) \leq O\left(V_{\max}|\mathcal{S}|\sqrt{|\mathcal{A}|}NH\sqrt{K\log K}\right), \tag{31}$$

where $V_{\max}$ is the maximum value-function, $|\mathcal{S}|$ is the size of the state space, $|\mathcal{A}|$ is the size of the action space, N is the number of agents, H is the time-horizon length in each episode, and K is the total number of episodes.

Proof. See Appendix A. □

As shown by Theorem 1, we achieve sub-linear regret in the number of episodes K. This regret bound is similar to the existing online learning algorithms with [20], [21], [23], [24], [32]–[34] or without fairness constraints [25], where [25] solves the problem with LP formulation, which is difficult to implement in large-scale systems.

VI. Simplification

A. Simplified Online-FMGF policy

Because the value function (20) depends on time, Algorithm 1 requires solving a sequential constrained convex optimization for all $t = H, H-1, \ldots, 1$. Therefore, it is very challenging to solve (20)-(21) to get the optimistic transition probabilities. In this sequel, we find an upper bound of the optimistic action-value function associated with the empirical transition probability $\hat{P}^k_n(\cdot|s, a)$.

Lemma 2. *It holds that*

$$\begin{aligned}
& Q^{P^k_n, r^k_n, \mu^k, \lambda^k_n}_{n,t}(s, a) - Q^{\hat{P}^k_n, r^k_n, \mu^k, \lambda^k_n}_{n,t}(s, a) \\
\leq& \rho^k_n(s, a) V_{max}, \qquad (32)
\end{aligned}$$

where the action-value function $Q_{n,t}(s, a)$ for any P_n, r_n, μ, and λ_n is defined in (10)*.*

Algorithm 2 Simplified Online-FMGF Policy

Input N arms, constraint M, fairness constraints β_n for all n, episode length H.
Initialize number of visits $B_n^k(s,a,s') = 0$ for all s, a, s'.
Initialize $\mu^{(1)}$ and $\lambda_n^{(1)}$ for all $n = \{1, 2, \ldots, N\}$.
for $k = 1, 2, \ldots$ **do**
    Reset $t = 1$ and $\mathbf{s} = \mathbf{s}_1$
    for $t = 1, 2, \ldots$ **do**
        Compute gain indices for each arm using (23) and (32).
        Activate M arms with the highest gain index.
        Observe transitions (s, a, s').
        Update visits $B_n^k(s,a,s')$, empirical mean $\hat{\mathbf{P}}^k$, empirical reward $\hat{\mathbf{r}}^k$, confidence regions $\mathcal{B}_p^k$, and $\mathcal{B}_r^k$.
        Every C steps update μ^k from (24).
    end for
    Update λ_n^{k+1} for all $n = 1, 2, \ldots, N$ form (26).
end for

Proof sketch. For notational simplicity, we omit the superscript $r_n^k, \mu^k, \lambda_n^k$ in this section. Let

$$Q_{n,t}^k(s,a) = Q_{n,t}^{\hat{P}_n^k}(s,a) + \rho_n^k(s,a) V_{\max}. \tag{33}$$

We first show that

$$\begin{aligned} &Q_{n,t}^k(s,a) - Q_{n,t}^{P_n^k}(s,a) \\ &\geq \sum_{s'} P_n^k(s'|s,a)(V_{n,t+1}^k(s') - V_{n,t+1}^{P_n^k}(s')), \end{aligned} \tag{34}$$

which holds because [35]

$$\sum_{s'} (P_n^k(s'|s,a) - \hat{P}_n^k(s'|s,a)) V_{n,t+1}^k(s') \leq \rho_n^k(s,a) V_{\max}.$$

Next, we show by backward induction that $V_{n,t}^k(s) \geq V_{n,t}^{P_n^k}(s)$ for any $t = H, H-1, \ldots, 1$. It is trivially true for $t = H+1$ as the value functions are 0 at $t = H+1$. Next, we assume that it is true for $t = h+1$, and show that it is true for $t = h$, which holds because of (34). Thus, we get $Q_{n,t}^k(s,a) \geq Q_{n,t}^{P_n^k}(s,a)$. Since, $V_{n,t}^k(s) = \max_a Q_{n,t}^k(s,a)$, we get $V_{n,t}^k(s) \geq V_{n,t}^{P_n^k}(s)$. This yields Lemma 2. □

Lemma 2 provides an upper bound of the optimistic action-value function $Q_{n,t}^{P_n^k}(s,a)$. Hence, we can use the upper bound of $Q_{n,t}^{P_n^k}(s,a)$ to directly compute the gain index and no longer need to solve any optimization problem (20)-(21). The approximation gap will be smaller if the confidence radius $\rho_n^k(s,a)$ is small. Specifically, as $k \to \infty$, the number of visits $B_n^k(s,a)$ will grow to ∞ and $\rho_n^k(s,a)$ becomes close to zero. To compute the gain index, we use the upper bound of $Q_{n,t}^{P_n^k}(s,a)$ from (32).

The simplified algorithm is presented in Algorithm 2. The benefit over earlier studies [22], [23], [25] is that we do not need to compute the optimistic transition probabilities in Algorithm 2. Instead, we directly compute the gain index.

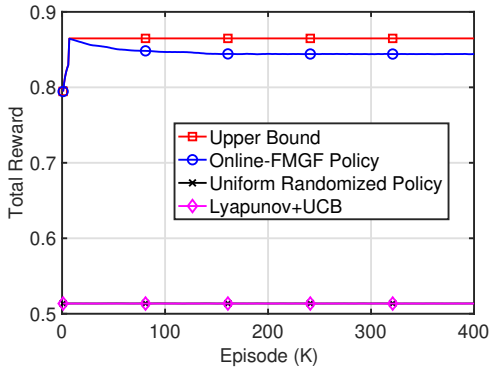

Fig. 1: Total reward vs the number of # episodes (K), with # agents $N = 5$ in interference-aware throughput maximization system.

B. Regret Bound for Algorithm 2

Next, we analyze the regret of Algorithm 2. In this sequel, we have the following theorem.

Theorem 2. *With probability $1 - \epsilon$, the cumulative regret in K episodes is given by*

$$\text{Reg}(K) \leq O\left(V_{\max} |\mathcal{S}| \sqrt{|\mathcal{A}|} N H \sqrt{K \log K}\right), \tag{35}$$

where $V_{\max}$ is the maximum value function, $|\mathcal{S}|$ is the size of the state space, $|\mathcal{A}|$ is the size of the action space, N is the number of agents, H is the time-horizon length in each episode, and K is the total number of episodes.

Proof sketch. From Lemma 1, we get

$$\text{Reg}(k) \leq \sum_{n=1}^{N} V_{n,1}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s_n(1)) - V_{n,1}^{P_n, r_n, \mu^k, \lambda_n^k}(s_n(1)). \tag{36}$$

We show that $V_{n,t}^k(s) \geq V_{n,t}^{P_n^k}(s)$ in the proof sketch of Lemma 2, where $V_{n,t}^k(s) = \max_a Q_{n,t}^k(s,a)$, and $Q_{n,t}^k(s,a)$ is defined in (33). Using this in (36) implies

$$\begin{aligned} \text{Reg}(k) \leq &\sum_{n=1}^{N} V_{n,1}^{\hat{P}_n^k, r_n^k, \mu^k, \lambda_n^k}(s_n(1)) - V_{n,1}^{P_n, r_n, \mu^k, \lambda_n^k}(s_n(1)) + \\ &\mathbb{E}_{\mathcal{F}_k, \pi^k} \rho_n^k(s,a) V_{\max}. \end{aligned} \tag{37}$$

Next, applying the similar argument as presented in Appendix A with the additional constant term in (37), (35) follows. In particular, adapting the above with respect to the filtration $\mathcal{F}_k$ till episode k, we can apply the standard concentration lemma. □

Theorem 2 implies that Algorithm 2 achieves sub-linear regret in the number of episodes K. One major benefit is that even after simplification, the performance of Algorithm 2 aligns with Algorithm 1. Initially, the performance of Algorithm 2 may not be as good as Algorithm 1; however, with increasing number of episodes, it will be comparable with Algorithm 1.

VII. Simulation Results

We evaluate the performance of the following policies:

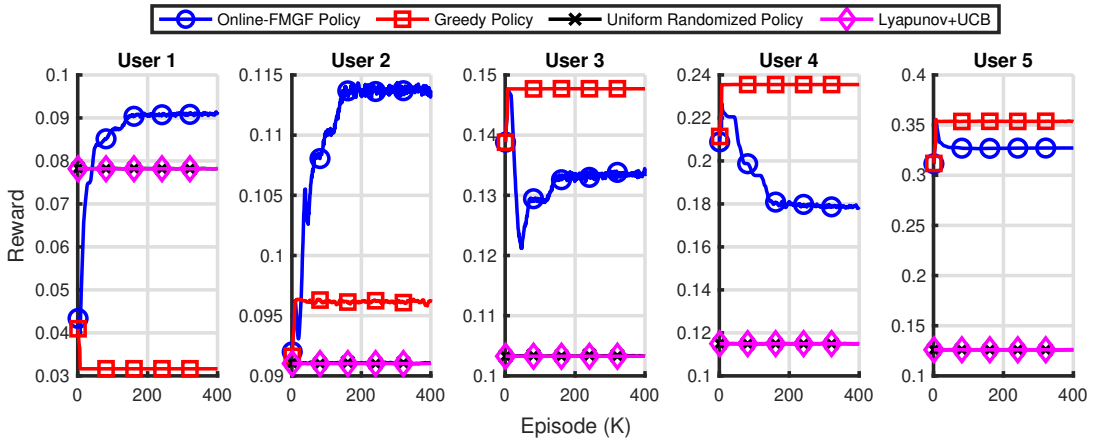

Fig. 2: Individual reward vs the number of # episodes (K), with #agents $N = 5$ in interference-aware throughput maximization system.

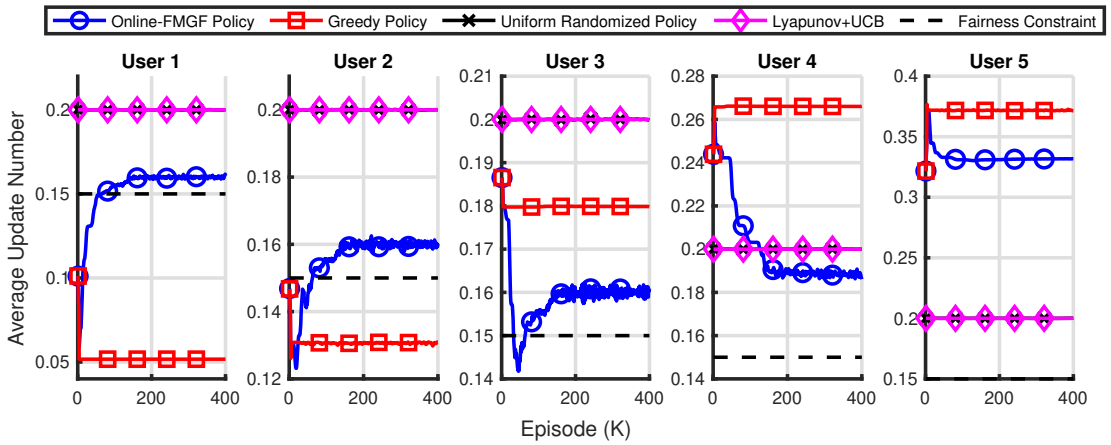

Fig. 3: Update Number vs the number of # episodes (K), with # agents $N = 5$ in interference-aware throughput maximization system.

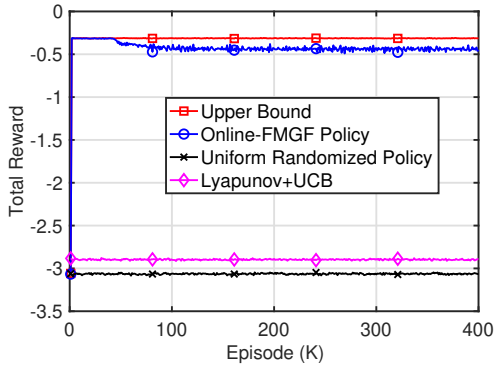

Fig. 4: Total reward vs the number of # episodes (K), with # agents $N = 4$ in land mobile satellite system.

- Uniform Randomized Policy: This policy randomly selects users following a uniform distribution.
- Online-FMGF Policy in Algorithm 2.
- Extended Linear Program (ELP)-based Policy in [25]
- Lyapunov+UCB-based policy in [36]
- Greedy Policy (Upper bound): This policy solves the following problem:

$$\max_{\pi} \frac{1}{T}\mathbb{E}_{\pi}\left[\sum_{t=1}^{T}\sum_{n=1}^{N} r_n(s_n(t), a_n(t))\right]$$
$$\text{s.t.} \sum_{n=1}^{N} a_n(t) \leq M, \forall t. \quad (38)$$

This policy does not consider any fairness constraints and the result provides an upper bound of the solution to problem (1)-(3).

We consider the following three systems to evaluate the performance of the proposed Online-FMGF policy.

A. Interference-aware Throughput Maximization

Consider a throughput maximization system with 5 classes of arms of the Restless bandits. Each class contains one agent. The system state X_k follows a 5-state Markov chain with 2 actions. If the action for one agent is 0, then the reward is 0. If the action is 1, then the reward is $\log(1 + X_k)$. The resource constraint is $M = 1$, and the fairness constraints are $\beta_n = 0.15$ for all n. The simulation is run for $T = 100000$ time-slots over 400 episodes.

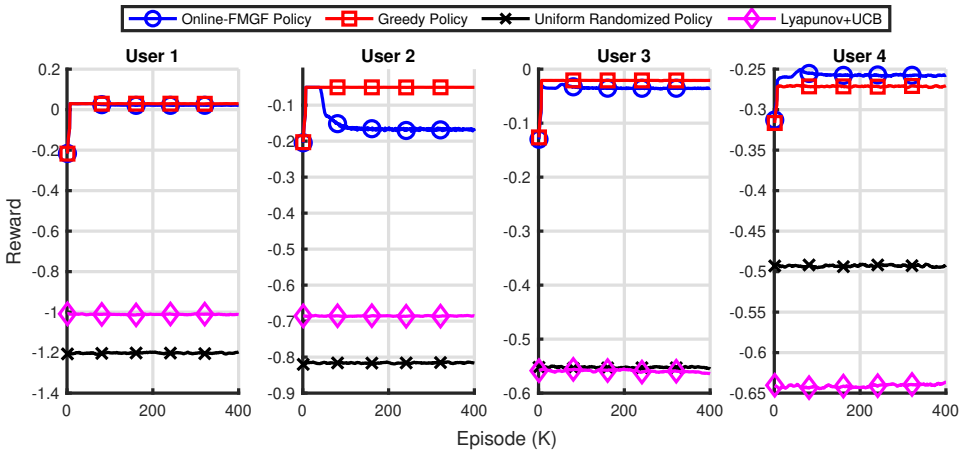

Fig. 5: Individual reward vs the number of # episodes (K), with # agents $N = 4$ in land mobile satellite system.

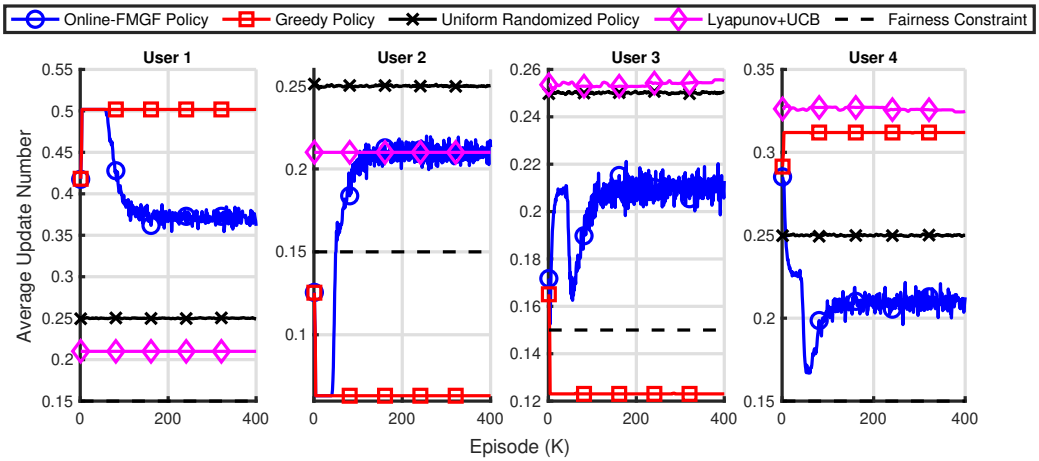

Fig. 6: Update Number vs the number of # episodes (K), with # agents $N = 4$ in land mobile satellite system.

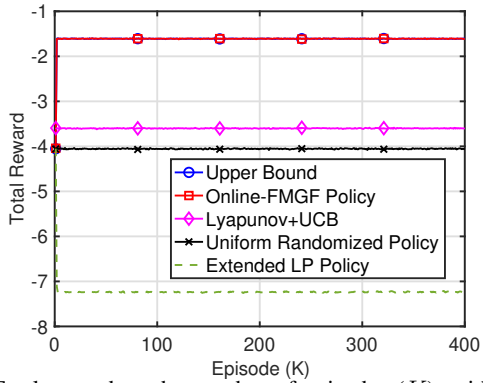

Fig. 7: Total reward vs the number of episodes (K), with # agents $N = 2$ in land mobile satellite system.

Figure 1 illustrates the total reward vs the number of episodes, where the greedy policy provides an upper bound of the total reward. The performance of the Online-FMGF policy is very close to the upper bound and better than those of the uniform randomized policy and Lyapunov+UCB-based policy in all episodes.

Figure 2 and Figure 3 illustrate that whenever the greedy policy does not satisfy the fairness constraints, it performs worse, specifically for agents 1 and 2. Hence, agents 1 and 2 will always be neglected in the greedy policy, as the goal is to only maximize the total reward. Therefore, the individual rewards for both agents 1 and 2 are worse. The other three policies satisfy the fairness constraints. However, the individual reward performance of the Online-FMGF policy is better.

B. Channel State-aware Throughput Maximization in Land Mobile Satellite System

Consider a land mobile satellite system as presented in [37]. We consider 4 agents with four different elevation angles $(40°, 60°, 70°, 80°)$ of the antenna in an urban area [37, Table III]. The system state is the channel state (*Good* or *Bad*), which is a two-state Markov chain. We consider the state transition matrix in [37]. The reward is the average direct signal mean when the chosen action is 1; otherwise, the reward is 0. The resource constraint is $M = 1$, and the fairness constraints are $\beta_n = 0.15$ for all n. The simulation is run for $T = 100000$ time-slots over 400 episodes.

In Figure 4, the greedy policy provides an upper bound of the total reward, and the total reward of our proposed policy is very close to the upper bound. Also, in Figures 5 and 6, the greedy policy does not satisfy the fairness constraints for agents 2 and 3. Hence, the individual rewards for agents 2 and 3 for greedy policy is worse than with uniform randomized, Online-FMGF, and Lyapunov+UCB-based policies.

We also simulate the ELP policy for two agents. Due to the high complexity of the ELP policy in [25], we consider a simple scenario with 2 agents and a two-state Markov chain which requires satisfying 41 constraint equations as decribed in [25]. Therefore, we cannot compare the performance with more than two agent systems. Figure 7 illustrates that the Online-FMGF policy outperforms all the other policies including the ELP and is very close to the upper bound.

VIII. Conclusion

We studied online learning algorithms for resource allocation in multi-agent systems. We formulate and solve an RMAB problem with fairness constraints and developed an Online Fair Maximum Gain First (Online-FMGF) policy that does not need to satisfy indexability. Our proposed algorithm does not need to compute the optimistic transition probabilities and hence can be applicable to large-scale systems. For future work, we will study regret bound and comparison of our policy with the best gain index-based policy with known system dynamics. The characterization of the regret for the large-state-space using the function approximation constitutes an important future research direction.

Appendix A
Proof of Theorem 1

When the actual transition probability and reward lies within the confidence ball, i.e., $\mathbf{P} \in \mathcal{B}_p^k$ and $\mathbf{r} \in \mathcal{B}_r^k$, applying (29), we can write

$$\begin{aligned}\text{Reg}(k) \leq& \sum_{n=1}^{N} V_{n,1}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s_n(1)) - V_{n,1}^{P_n, r_n, \mu^k, \lambda_n^k}(s_n(1)) \\ =& \sum_{n=1}^{N} \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H} r_n^k(s_n(t), a_n(t)) - r_n(s_n(t), a_n(t)) \\ & \Bigg| s_n(t)\Bigg] + \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H}\sum_{s'}\Big(P_n^k(s'|s_n(t), a_n(t)) \\ & - P_n(s'|s_n(t), a_n(t))\Big) V_{n,t+1}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s')\Bigg], \end{aligned} \tag{39}$$

where (39) holds from [38, Lemma 10]. Continuing from (39), we can write

$$\begin{aligned}\text{Reg}(k) \leq& \sum_{n=1}^{N} \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H} \rho_n^k(s_n(t), a_n(t))\Bigg] + \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H}\sum_{s' \in \mathcal{S}} \\ & \Big|P_n^k(s'|s_n(t), a_n(t)) - P_n(s'|s_n(t), a_n(t))\Big| V_{\max}\Bigg], \end{aligned} \tag{40}$$

where $V_{\max}$ is bounded because the optimistic reward $r_n^k(s,a)$ is bounded by D from (19) and the Lagrange multipliers μ^k and λ_n^k are bounded by A, and G, respectively. Then, we can write

$$\begin{aligned}\text{Reg}(k) \leq& \sum_{n=1}^{N}(1+V_{\max})\mathbb{E}_{P_n,\pi_n^k}\Bigg[\sum_{t=1}^{H}\rho_n^k(s_n(t),a_n(t))\Bigg] \\ =& \sum_{n=1}^{N}(1+V_{\max})\mathbb{E}_{P_n,\pi_n^k}\Bigg[\sum_{t=1}^{H}\sqrt{\frac{2|\mathcal{S}|\log(2|\mathcal{S}||\mathcal{A}|N\frac{k^2}{\epsilon})}{\max(1, B_n^k(s_n(t),a_n(t)))}}\Bigg] \\ =& \sum_{n=1}^{N}(1+V_{\max})\sqrt{2|\mathcal{S}|\log(2|\mathcal{S}||\mathcal{A}|N\frac{k^2}{\epsilon})} \\ & \mathbb{E}_{P_n,\pi_n^k}\Bigg[\sum_{t=1}^{H}\sqrt{\frac{1}{\max(1, B_n^k(s_n(t),a_n(t)))}}\Bigg] \\ \leq& \sum_{n=1}^{N}\frac{1}{\epsilon}(1+V_{\max})\sqrt{2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|NK)} \\ & \mathbb{E}_{P_n,\pi_n^k}\Bigg[\sum_{s\in S, a\in A}\frac{\gamma_n^k(s,a)}{\max(1, B_n^k(s_n(t),a_n(t)))}\Bigg], \end{aligned} \tag{41}$$

where $\gamma_n^k(s,a)$ is a random variable that denotes the number of visits to $(s,a) \in \mathcal{S}\times\mathcal{A}$ in episode k for arm n. Furthermore, applying [23, Lemma E.3] to (41) yields

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\gamma_n^k(s,a)}{\sqrt{\max(1, B_n^k(s,a))}} \\ \leq& \Big(\sqrt{H+1}+1\Big)\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{B_n^k(s,a)} \\ \leq& \Big(\sqrt{H+1}+1\Big)\sqrt{|\mathcal{S}||\mathcal{A}|KH} \end{aligned} \tag{42}$$

Next, by taking summation over all K episodes, we get the total regret

$$\begin{aligned}&\sum_{n=1}^{N}\frac{1}{\epsilon}(1+V_{\max})\sqrt{2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|NK)}(\sqrt{H+1}+1) \\ &\qquad \sqrt{|\mathcal{S}||\mathcal{A}|KH} \\ &\leq O\Big(V_{\max}|\mathcal{S}|\sqrt{|\mathcal{A}|}NH\sqrt{K\log K}\Big) \end{aligned} \tag{43}$$

The regret outside the confidence ball vanishes with probability

$$\begin{aligned}&\Pr(\mathbf{P}\in\mathcal{B}_p^k) \text{ and } \mathbf{r}\in\mathcal{B}_r^k, \forall\sqrt{K}\leq k\leq K) \\ =&1-\Pr(\mathbf{P}\in\mathcal{B}_p^k)-\Pr(\mathbf{r}\in\mathcal{B}_r^k) \\ \geq&1-\sum_{\sqrt{K}\leq k\leq K}\frac{\epsilon}{k^2}-\sum_{\sqrt{K}\leq k\leq K}\frac{m\epsilon^{2|\mathcal{S}|}}{k^{4|\mathcal{S}|}}=1-\int_{\sqrt{K}}^{K}\frac{\epsilon}{k^2}-\int_{\sqrt{K}}^{K}\frac{m\epsilon^{2|\mathcal{S}|}}{k^{4|\mathcal{S}|}} \\ =&1-\epsilon\Big(\frac{1}{\sqrt{K}}-\frac{1}{K}\Big)-m\epsilon^{2|\mathcal{S}|}\Big(\frac{1}{K^{2|\mathcal{S}|-(1/2)}}-\frac{1}{K^{4|\mathcal{S}|-1}}\Big), \end{aligned} \tag{44}$$

where $m = \frac{2}{(|\mathcal{S}||\mathcal{A}|N)^{2|\mathcal{S}|-1}}$. Applying union bound for all possible $K \in \mathbb{N}$, it holds with high probability $1 - O(\epsilon)$.

REFERENCES

[1] C. Papadimitriou and J. Tsitsiklis, "The complexity of optimal queueing network control," in *IEEE CCC*, 1994, pp. 318–322.

[2] P. Whittle, "Restless bandits: activity allocation in a changing world," *Journal of Applied Probability*, vol. 25A, pp. 287–298, 1988.

[3] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of applied probability*, vol. 27, no. 3, pp. 637–648, 1990.

[4] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[5] W. Ouyang, A. Eryilmaz, and N. B. Shroff, "Low-complexity optimal scheduling over time-correlated fading channels with ARQ feedback," *IEEE Trans. Mob. Comput.*, vol. 15, no. 9, pp. 2275–2289, 2016.

[6] I. Kadota, A. Sinha, and E. Modiano, "Optimizing age of information in wireless networks with throughput constraints," in *IEEE INFOCOM*, 2018, pp. 1844–1852.

[7] V. Tripathi and E. Modiano, "A Whittle index approach to minimizing functions of age of information," in *IEEE Allerton*, 2019, pp. 1160–1167.

[8] G. Chen, S. C. Liew, and Y. Shao, "Uncertainty-of-information scheduling: A restless multiarmed bandit framework," *IEEE Trans. Inf. Theory*, vol. 68, no. 9, pp. 6151–6173, 2022.

[9] M. K. C. Shisher, Y. Sun, and I.-H. Hou, "Timely communications for remote inference," *IEEE/ACM Transactions on Networking*, vol. 32, no. 5, pp. 3824–3839, 2024.

[10] T. Z. Ornee and Y. Sun, "A Whittle index policy for the remote estimation of multiple continuous Gauss-Markov processes over parallel channels," in *ACM MobiHoc*, 2023, p. 91–100.

[11] G. Xiong, X. Qin, B. Li, R. Singh, and J. Li, "Index-aware reinforcement learning for adaptive video streaming at the wireless edge," in *ACM MobiHoc*, 2022, pp. 81–90.

[12] Y. Zou, K. T. Kim, X. Lin, and M. Chiang, "Minimizing age-of-information in heterogeneous multi-channel systems: A new partial-index approach," in *ACM MobiHoc*, 2021, pp. 11–20.

[13] G. Chen and S. C. Liew, "An index policy for minimizing the uncertainty-of-information of Markov sources," *arXiv preprint arXiv:2212.02752*, 2022.

[14] M. K. C. Shisher, B. Ji, I.-H. Hou, and Y. Sun, "Learning and communications co-design for remote inference systems: Feature length selection and transmission scheduling," *IEEE J. Sel. Areas Inf. Theory*, vol. 4, pp. 524–538, 2023.

[15] T. Z. Ornee, M. K. C. Shisher, C. Kam, and Y. Sun, "Context-aware status updating: Wireless scheduling for maximizing situational awareness in safety-critical systems," in *IEEE MILCOM*, 2023, pp. 194–200.

[16] ——, "Remote safety monitoring: Significance-aware status updating for situational awareness," 2025. [Online]. Available: https://arxiv.org/abs/2507.09833

[17] M. K. C. Shisher, A. Piaseczny, Y. Sun, and C. G. Brinton, "Computation and communication co-scheduling for timely multi-task inference at the wireless edge," *IEEE INFOCOM*, 2025.

[18] G. Neu and G. Bartók, "An efficient algorithm for learning with semi-bandit feedback," in *International Conference on Algorithmic Learning Theory*. Springer, 2013, pp. 234–248.

[19] D. Foster and A. Rakhlin, "Beyond ucb: Optimal and efficient contextual bandits with regression oracles," in *International conference on machine learning*. PMLR, 2020, pp. 3199–3210.

[20] P. Auer and R. Ortner, "Logarithmic online regret bounds for undiscounted reinforcement learning," *Advances in neural information processing systems*, vol. 19, 2006.

[21] R. Ortner, D. Ryabko, P. Auer, and R. Munos, "Regret bounds for restless markov bandits," in *International conference on algorithmic learning theory*. Springer, 2012, pp. 214–228.

[22] N. Akbarzadeh and A. Mahajan, "On learning whittle index policy for restless bandits with scalable regret," *IEEE Transactions on Control of Network Systems*, vol. 11, no. 3, pp. 1190–1202, 2023.

[23] K. Wang, L. Xu, A. Taneja, and M. Tambe, "Optimistic Whittle index policy: Online learning for restless bandits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 10 131–10 139.

[24] M. K. C. Shisher, V. Tripathi, M. Chiang, and C. G. Brinton, "Online learning of Whittle indices for restless bandits with non-stationary transition kernels," 2025. [Online]. Available: https://arxiv.org/abs/2506.18186

[25] S. Wang, G. Xiong, and J. Li, "Online restless multi-armed bandits with long-term fairness constraints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2024, pp. 15 616–15 624.

[26] D. P. Bertsekas *et al.*, "Dynamic programming and optimal control, 4th edition," *Belmont, MA: Athena Scientific*, vol. 1, 2011.

[27] I. M. Verloop, "Asymptotically optimal priority policies for indexable and nonindexable restless bandits," 2016.

[28] A. Nedic and A. Ozdaglar, "Subgradient methods in network resource allocation: Rate analysis," in *IEEE CISS*, 2008, pp. 1189–1194.

[29] K. E. Avrachenkov and V. S. Borkar, "Whittle index based Q-learning for restless bandits with average reward," *Automatica*, vol. 139, p. 110186, 2022.

[30] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, "Inequalities for the L_1 deviation of the empirical distribution," *Hewlett-Packard Labs, Tech. Rep*, p. 125, 2003.

[31] D. P. Dubhashi and A. Panconesi, *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[32] Y. H. Jung and A. Tewari, "Regret bounds for thompson sampling in episodic restless bandit problems," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[33] S. Wang, L. Huang, and J. Lui, "Restless-UCB, an efficient and low-complexity algorithm for online restless bandits," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 878–11 889, 2020.

[34] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," *Advances in neural information processing systems*, vol. 21, 2008.

[35] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, "Reinforcement learning: Theory and algorithms," *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, vol. 32, p. 96, 2019.

[36] 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.

[37] R. Prieto-Cerdeira, F. Perez-Fontan, P. Burzigotti, A. Bolea-Alamañac, and I. Sanchez-Lago, "Versatile two-state land mobile satellite channel model with first application to dvb-sh analysis," *International Journal of Satellite Communications and Networking*, vol. 28, no. 5-6, pp. 291–315, 2010.

[38] P. Ju, A. Ghosh, and N. B. Shroff, "Achieving fairness in multi-agent markov decision processes using reinforcement learning," *arXiv preprint arXiv:2306.00324*, 2023.