Fair Online Learning for Restless Bandits

Tasmeen Zaman Ornee*, Arnob Ghosh†, Ananthram Swami‡, and Ness B. Shroff*
*Department of ECE, The Ohio State University, Columbus, OH, USA
†New Jersey Institute of Technology, Newark, NJ, USA
‡DEVCOM Army Research Laboratory, Adelphi, MD, USA
Email: {ornee.1, shroff.11}@osu.edu, arnob.ghosh@njit.edu, ananthram.swami.civ@army.mil

***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 decisions, and issue 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]–[8] have been formulated as RMABs. A major challenge in developing Whittle index policy is to establish indexability, which is difficult to achieve in practice. Therefore, non-indexable scheduling policies also have been studied in recent years [6], [9]–[12]. An asymptotically optimal gain index-based policy was proposed in [6], [9]–[12] with known dynamics. Although most existing studies assume known dynamics, these are often unknown in practice. Hence, online learning methods for RMABs have been developed [13]–[16] that adaptively learn the system dynamics.

Traditional RMAB solutions typically overlook fairness. If the overall goal is to maximize the reward while satisfying the resource constraints, then the agent with a lower reward may never get any resources. To bridge this gap, we incorporate time-average fairness constraints with the RMAB model that ensure a minimum average activation fraction for each arm [17], [18]. The authors in [17] and [18] considered an MAB and an RMAB with fairness constraints for infinite horizon systems, respectively. Specifically, [18] provided an index-based policy based on linear programming, which exhibits polynomial time-complexity with system size. Moreover, translating the infinite horizon problem to a finite horizon requires computing time varying indices. In this study, we develop an online learning framework that solves fairness-constrained RMABs in finite horizon systems with regret guarantee and low-computational complexity.

The summary of our contributions are discussed below.

- We develop a UCB-based Online Fair Maximum Gain First (Online-FMGF) policy with two time-scale Lagrangian update (see Algorithm 1). Unlike Whittle policy, our policy does not need to satisfy any indexability condition. The advantage of the developed policy compared to other online learning policies for RMABs with fairness [18] is its computational scalability and finite-horizon adaptation. Specifically, Algorithm 1 computes the gain index with a complexity of $O(N|\mathcal{S}|^2|\mathcal{A}|)$, where N is the system size, $|\mathcal{S}|$ is the size of the state space, and $|\mathcal{A}|$ is the size of the action space. In contrast, to compute the index from occupancy measures, the convex optimization-based Extended Linear Program (ELP) developed in [18] scales polynomially with the number of total variables

and the complexity is $O((N|\mathcal{S}|^2|\mathcal{A}|)^q)$, where $q > 1$ (see Remark 1 for details). The linear scaling makes our Algorithm more suitable for large-scale systems.

- We achieve a further simplification of the Online-FMGF policy that eliminates the need to estimate optimistic action-value functions (see Algorithm 2). To get the gain indices, Algorithm 1 requires the optimistic action-value functions by optimizing transition probabilities within a ball defined by a confidence radius. Similar to this approach, prior studies also require computation of optimistic transition probabilities to get an online Whittle index policy [16] and optimistic occupancy measures to get an online index-based policy [18]. We reduce these computations by approximating optimistic action-value functions from empirical transition probabilities and confidence radius. This is achieved by deriving an upper bound of the optimistic action-value functions (see Lemma 2). Consequently, Algorithm 2 exhibits $O(N|\mathcal{S}||\mathcal{A}|)$ complexity to compute the gain indices.
- 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). Furthermore, because we utilize a two time-scale update method for Lagrange multipliers, we derive the specific learning rates required to guarantee these theoretical bounds. To the best of our knowledge, these results provide the first sublinear regret guarantees for RMABs incorporating fairness without employing any LP formulations. Our results extend earlier online learning policies for RMABs [15], [16] by providing an online gain index-based policy and by adding fairness constraints [18], while still achieving similar regret performance as in [15], [16], [18].
- We evaluate the performance of our proposed policy with respect to the following baselines: greedy policy, uniform randomized policy, UCB+Lyapunov-based policy [17], and Extended Linear Program (ELP)-based policy [18], where the proposed policy outperforms the other policies.

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 $D \in (0, \infty)$ 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 $s_n(t+1)$ denotes 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 $\beta_n \in (0, 1)$ is the minimum fraction of time that arm n should remain active.

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. IV-B).

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 $\lambda_n, n = 1, 2, \ldots, N$:

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

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

Next, we relax constraint (5) as

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

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

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

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

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

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

The optimal Lagrange multipliers are obtained by solving the following dual problem:

$$(\mu^*, \lambda_1^*, \lambda_2^*, \ldots, \lambda_N^*) = \underset{\mu, \lambda_1, \lambda_2, \ldots, \lambda_N}{\operatorname{argmin}} L(\pi^*, \mathbf{P}, \mathbf{r}, \mu, \boldsymbol{\lambda}, \mathbf{s}_1), \quad (9)$$

where $L(\pi^*, \mathbf{P}, \mathbf{r}, \mu, \boldsymbol{\lambda}, \mathbf{s}_1)$ is the optimal value of (7). Following the solution of (8) and (9), we will develop a solution for problem (1)-(3).

B. Solution to the Decoupled Problem (8)

Problem (8) can be solved by dynamic programming [19]. The Bellman optimality equation of MDP (8) for given state $s_n(t) = s$ and action $a_n(t) = a$ at time-slot t is given by

$$V_{n,t}(s) = \max_{a \in \mathcal{A}} Q_{n,t}(s, a), \quad (10)$$

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

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

We obtain the value function by using backward induction [19]. WLOG, we assume $V_{n,T+1}(s) = 0$ for all $s \in \mathcal{S}$.

Bellman's optimality equation (10) yields that arm n is activated if $Q_{n,t}(s, 1) > Q_{n,t}(s, 0)$. An equivalent policy is obtained by utilizing the gain index, defined as [6], [10], [11]

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

which is the difference of two action-value functions. The gain index-based policy activates arm n 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, policies using action-value functions and gain indices are not equivalent. Gain-index based policy is known to be asymptotically optimal for RMAB [9] under certain conditions.

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

In this section, we first describe our learning setting (Sec. IV-A). Then, we develop a UCB-based Online-FMGF policy to solve (1)-(3) (Sec. IV-B). Finally, we analyze the regret to show that the developed algorithm achieves sublinear regret (Sec. IV-C).

A. 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 [16] utilize Lagrangian relaxed problem to evaluate the performance of an online policy. Motivated by [16], we adopt Lagrangian (7) to evaluate the performance of our online policy π^k. When the system size is large, Lagrangian relaxation provides an asymptotic approximation to the performance of the optimal policy [3]; 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 (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 $A \in (0, \infty)$ and

$G \in (0, \infty)$ are positive constants. Using (13), we get the total regret as follows:

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

B. UCB-based Online-FMGF 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, $\mathbf{P}$ and $\mathbf{r}$ 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 the last k episodes.

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

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

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

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 develop our online Algorithm. The optimistic reward is given by

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

The optimistic action-value functions can be determined by using backward induction for $t = H, H-1, \ldots, 1$ as follows

$$Q_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s, a) = (\lambda_n^k - \mu^k)a + r_n^k(s, a) + \max_{P_n^k \in \mathcal{B}_p^k} \sum_{s'} P_n^k(s'|s, a) V_{n,t+1}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s'), \tag{20}$$

Algorithm 1 Online-FMGF Policy

1: Input: N arms, constraint M, fairness constraints β_n for all n, episode length H.
2: Initialize number of visits $B_n^1(s, a, s') = 0$ for all s, a, s' and $n = \{1, 2, \ldots, N\}$.
3: Initialize $\mu^{(1)}$ and $\lambda_n^{(1)}$ for all $n = \{1, 2, \ldots, N\}$.
4: **for** $k = 1, 2, \ldots$ **do**
5: Reset $t = 1$ and $\mathbf{s} = \mathbf{s}_1$.
6: Compute optimistic reward and action-value functions for each arm by using (19), and solving (20), respectively.
7: Compute gain indices from (22) for each arm and for all $t = 1, 2, \ldots, H$.
8: **for** $t = 1, 2, \ldots, H$ **do**
9: Activate M arms with the highest positive gain indices.
10: Observe transitions (s, a, s').
11: 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 $\mathcal{B}_r^k$.
12: Every C steps update μ^k from (23).
13: **end for**
14: Update λ_n^{k+1} for all $n = 1, 2, \ldots, N$ from (25).
15: **end for**

where the optimistic value function is given by

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

WLOG, we assume $V_{n,H+1}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s) = 0$ for all $s \in \mathcal{S}$.

We solve (20) by using an inner maximization solution technique of the extended value iteration described in [20]. To that end, we first sort the set of states $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$ in descending order of their value-functions, such that $V_{n,t+1}(s_i) \geq V_{n,t+1}(s_{i+1})$. Next, we initialize the transition probabilities by setting $P_n^k(s_1|s, a) = \hat{P}_n^k(s_1|s, a) + \rho_n^k(s, a)/2$ for the state with the highest value function, and $P_n^k(s_i|s, a) = \hat{P}_n^k(s_i|s, a)$ for all other states $i > 1$. If the resulting probabilities sum to more than 1, we restore validity by iteratively reducing the probability of the state with the lowest value function. We start with $l = |\mathcal{S}|$ and set $P_n^k(s_l|s, a) = \max\{0, 1 - \sum_{j \neq l} P_n^k(s_j|s, a)\}$. This process is repeated for decreasing l until the transition probabilities sum to 1. The worst case complexity is $O(|\mathcal{S}|)$.

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

$$\alpha_{n,t}^{\mu^k, \lambda_n^k}(s) = Q_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s, 1) - Q_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s, 0). \tag{22}$$

The Online-FMGF policy is provided in Algorithm 1. At every episode k, this policy selects M arms with the highest positive gain indices (22).

Remark 1. A Lagrangian-based online Whittle index policy was studied in [16] for an infinite horizon problem without fairness constraints. The approach in [16] requires solving two sequential optimization problems: first to compute the optimistic transition probabilities, next, to compute the Whittle indices. Unlike [16], Algorithm 1 requires solving one opti-

Algorithm 2 Simplified Online-FMGF Policy

Input and initialization similar to Algorithm 1
for $k = 1, 2, \ldots$ **do**
  Reset $t = 1$ and $\mathbf{s} = \mathbf{s}_1$
  Compute gain indices for each arm using (22) and (28).
  for $t = 1, 2, \ldots, H$ **do**
    Activate M arms with the highest positive gain indices.
    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 (23).
  end for
  Update λ_n^{k+1} for all $n = 1, 2, \ldots, N$ form (25).
end for

mization problem (20) to obtain the optimistic action-value functions from which the gain indices (22) are directly computed. Furthermore, our method significantly reduces computational complexity compared to the occupancy measure-based Extended Linear Program (ELP) proposed in [18]. The ELP in [18] cannot be decomposed; hence, [18] requires finding the occupancy measures without decomposing the ELP into sub-problems. Therefore, scalability to large multi-agent systems is limited. In contrast, our method permits us to decompose the original problem into N per-arm problems. Moreover, the complexity of the ELP in [18] scales polynomially with number of variables of the ELP. Specifically, the complexity is $O((N|\mathcal{S}|^2|\mathcal{A}|)^q)$ with $q > 1$ in [18], whereas the computation of gain indices in our study for all states and all arms yields $O(N|\mathcal{S}|^2|\mathcal{A}|)$ complexity.

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 [21], [22].

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

$$\mu^k(j+1) = \max\left\{\mu^k(j) + \gamma_\mu\left(\frac{1}{C}\sum_{n=1}^{N}\sum_{t=jC+1}^{(j+1)C}\mathbb{E}[a_{\mu,n}^k(t)] - M\right), 0\right\}, \quad (23)$$

where $\mu^k(j)$ is the j-th update within episode k starting from $j = 1$, γ_μ is the learning parameter, and action $a_{\mu,n}^k(t)$ satisfies:

$$a_{\mu,n}^k(t) = \arg\max_a Q_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s, a). \quad (24)$$

We set $\mu^{k+1}(1) = \mu^k(\lfloor H/C \rfloor + 1)$. The Lagrange multipliers $(\lambda_n^{k+1})_{n=1}^N$ are computed after completing the k-th episode. The update rule is given by

$$\lambda_n^{k+1} = \max\left\{\lambda_n^k - \gamma_\lambda\left(\frac{1}{H}\sum_{t=1}^{H}\mathbb{E}[a_n^k(t)] - \beta_n\right), 0\right\}, \quad (25)$$

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

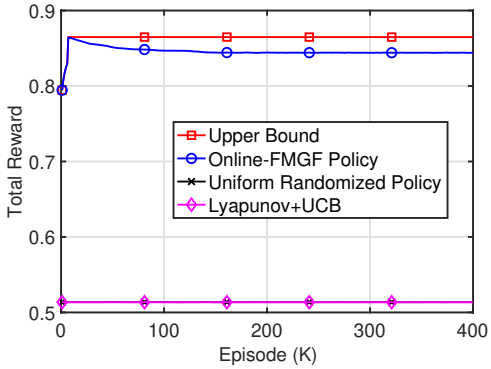

Fig. 1: Total reward vs # episodes (K), with # agents $N = 5$ in Throughput maximization with Markov SNR Model.

The update of μ^k operates at a faster time-scale for given $(\lambda_n^k)_{n=1}^N$. Then, $(\lambda_n^k)_{n=1}^N$ is updated at a slower time-scale. This is because simultaneously converging μ^k and $(\lambda_n^k)_{n=1}^N$ is challenging, and it may happen that none of the multipliers converge to the optimal value. Hence, we first update μ^k for given $(\lambda_n^k)_{n=1}^N$. After μ^k converges, we update $(\lambda_n^k)_{n=1}^N$.

C. Regret Bound for Algorithm 1

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

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

Proof sketch. Following [23, Theorem 2.1], we bound the L_1-deviation of the actual transition and the empirical transition. Moreover, by using the Chernoff-Hoeffding inequality [24], we bound the actual reward and the empirical reward. The details are provided in Appendix A. □

Lemma 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 Lemma 1 to bound the regret.

We can decompose $\mathrm{Reg}(k)$ in (13) into three components:

$$\begin{aligned}\mathrm{Reg}(k) =& L(\pi^*, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^k, \boldsymbol{\lambda}^k, \mathbf{s}_1) \\ =& \underbrace{L(\pi^*, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1)}_{\mathrm{Term1}(\mathrm{Reg}^\pi(k))} \\ &+ \underbrace{L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^k, \mathbf{s}_1)}_{\mathrm{Term2}(\mathrm{Reg}^\lambda(k))} \\ &+ \underbrace{L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^k, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^k, \boldsymbol{\lambda}^k, \mathbf{s}_1)}_{\mathrm{Term3}(\mathrm{Reg}^\mu(k))}, \quad (26)\end{aligned}$$

where Term1 represents the regret associated with policy π^k relative to π^*, Term2 represents the regret due to the deviation of the Lagrange multiplier $\boldsymbol{\lambda}^k$ from its optimal value, and Term3 represents the corresponding regret due to the sequence

of Lagrange multipliers $\mu^k = (\mu^k(1), \mu^k(2), \ldots, \mu^k(\lfloor H/C \rfloor + 1))$ within episode k. Recall $\mu^k(\cdot)$ is updated after every C time-slots within episode k.

In this sequel, we have the following theorem.

Theorem 1. *With probability $1-\epsilon$, if the learning parameters $\gamma_\lambda = c_1/\sqrt{K}$ and $\gamma_\mu = c_2/\sqrt{K}$ for any $c_1, c_2 > 0$, then the cumulative regret of Algorithm 1 in K episodes is given by*

$$\text{Reg}(K) \leq O\big(V_{\max}|\mathcal{S}|\sqrt{|\mathcal{A}|}N\sqrt{K\log(K/\epsilon)} + N\sqrt{K}(1+C/H)\big), \quad (27)$$

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, K is the total number of episodes, and μ and $\boldsymbol{\lambda}$ are updated after C and H time slots, respectively.

Proof. We prove Theorem 1 by finding the regret bounds for Term1, Term2, and Term3. See Appendix B for details. □

As shown in Theorem 1, we achieve sub-linear regret in the number of episodes K. This regret bound is similar to the existing online learning algorithms without [14], [16], [20], [25], [26] or with fairness constraints [18], where [18] solves the problem with LP formulation which is difficult to implement in large-scale systems.

V. Simplification

A. Simplified Online-FMGF policy

Recall that (20) requires us to solve an inner maximization step that takes $O(|\mathcal{S}|)$ complexity. We can avoid this step by using the following lemma.

Lemma 2. *It holds that*

$$Q_{n,t}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s,a) - Q_{n,t}^{\hat{P}_n^k, r_n^k, \mu^k, \lambda_n^k}(s,a) \leq \rho_n^k(s,a) V_{\max}, \quad (28)$$

where the action-value function $Q_{n,t}(s,a)$ for any P_n, r_n, μ, and λ_n is defined in (11)*.*

Proof. See Appendix F. □

Lemma 2 provides an upper bound of the optimistic action-value functions associated with the empirical transition probability $\hat{P}_n^k(\cdot|s,a)$, which further simplifies Algorithm 1 and enables direct computation of the gain indices. 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.

The simplified algorithm is presented in Algorithm 2. The complexity of computing the gain indices in Algorithm 2 is $O(N|\mathcal{S}||\mathcal{A}|)$, which is smaller than that of Algorithm 2 and the earlier studies [15], [16], [18].

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 $1-\epsilon$, if the learning parameters $\gamma_\lambda = c_1/\sqrt{K}$ and $\gamma_\mu = c_2/\sqrt{K}$ for any $c_1, c_2 > 0$, then the cumulative regret in K episodes is given by*

$$\text{Reg}(K) \leq O\big(V_{\max}|\mathcal{S}|\sqrt{|\mathcal{A}|}N\sqrt{K\log(K/\epsilon)} + N\sqrt{K}(1+C/H)\big), \quad (29)$$

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, K is the total number of episodes, and μ and $\boldsymbol{\lambda}$ are updated after C and H time slots, respectively.

Proof. See Appendix G. □

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

VI. Simulation Results

We evaluate the performance of the following policies:

- 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 [18]
- Lyapunov+UCB-based policy in [17]
- Greedy Policy (Upper bound): This policy solves (1)-(2) without the fairness constraints (3), which yields an upper bound of the solution to problem (1)-(3).

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

A. Throughput Maximization with Markov SNR Model

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, which represent the random fluctuations of the SNR value at the receiver side. 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 $T = 10^5$ time-slots over 400 episodes.

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.

Figures 2 and 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

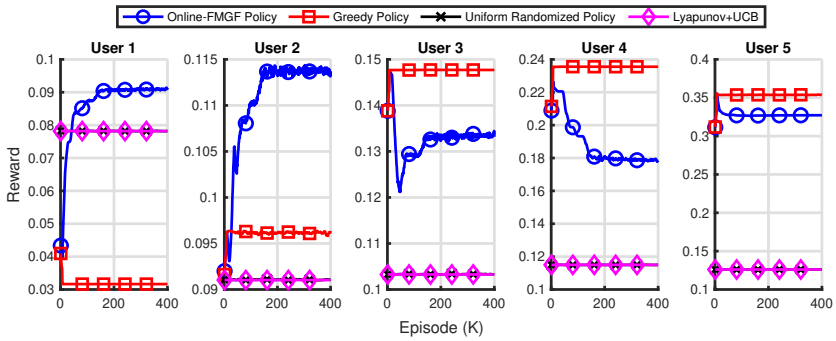

Fig. 2: Individual reward vs # episodes (K), with #agents $N = 5$ in Throughput maximization with Markov SNR Model.

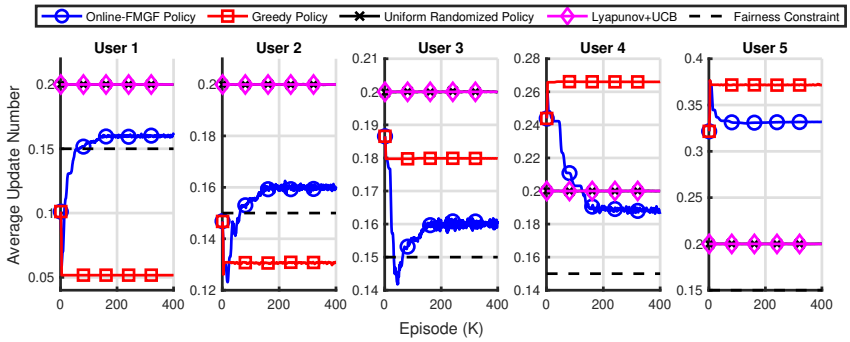

Fig. 3: Update Number vs # episodes (K), with # agents $N = 5$ in Throughput maximization with Markov SNR Model.

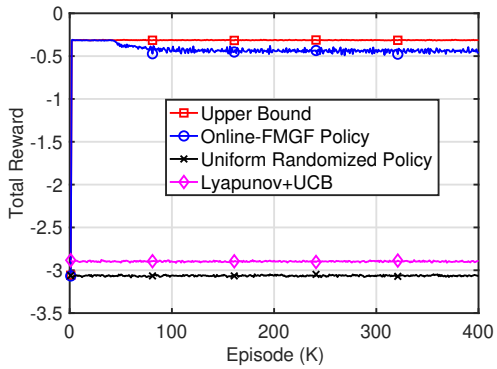

Fig. 4: Total reward vs # episodes (K), with # agents $N = 4$ in Throughput Maximization with Land Mobile Satellite System Model.

satisfy the fairness constraints. However, the individual reward performance of the Online-FMGF policy is better.

B. Throughput Maximization with Land Mobile Satellite System Model [27]

Consider a land mobile satellite system as presented in [27]. We consider 4 agents with four different elevation angles $(40°, 60°, 70°, 80°)$ of the antenna in an urban area [27, 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 [27]. 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 $T = 10^5$ time-slots over 400 episodes.

In Figure 4, 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 are worse than 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 [18], we consider a simple scenario with 2 agents and a two-state Markov chain which requires satisfying 41 constraint equations as described in [18]. 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.

VII. CONCLUSION

We investigated online learning algorithms for resource allocation in multi-agent systems. We formulated and solved an RMAB problem with fairness constraints and develop an Online-FMGF policy. Our proposed algorithm achieves smaller complexity and performs better than existing studies. The characterization of the regret for the large state-space using the function approximation constitutes an important future research direction.

REFERENCES

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

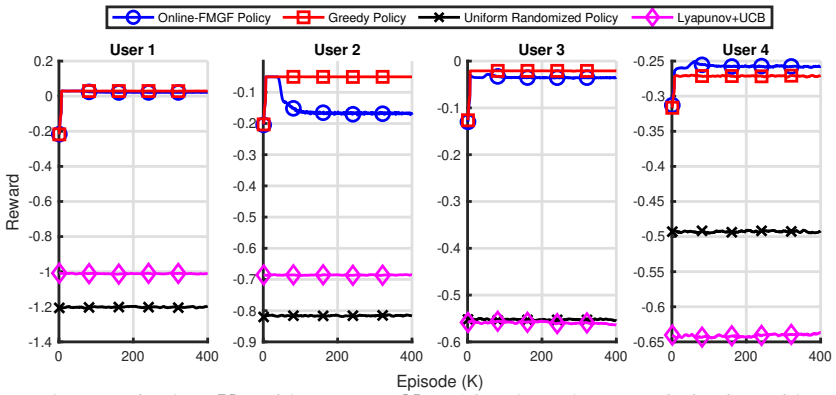

Fig. 5: Individual reward vs # episodes (K), with # agents $N = 4$ in Throughput Maximization with Land Mobile Satellite System Model.

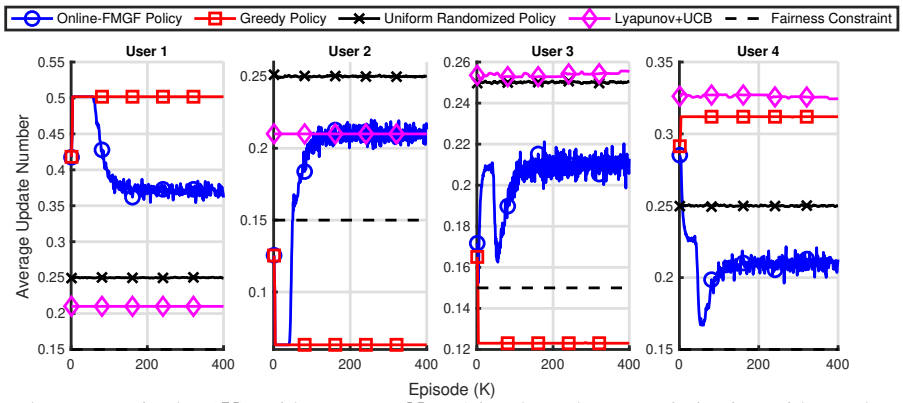

Fig. 6: Update Number vs # episodes (K), with # agents $N = 4$ in Throughput Maximization with Land Mobile Satellite System Model.

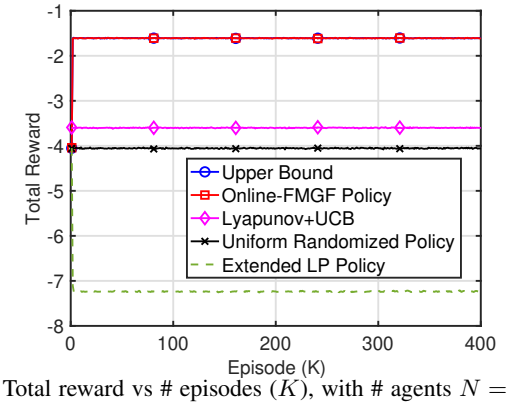

Fig. 7: Total reward vs # episodes (K), with # agents $N = 2$ in land mobile satellite system.

[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] 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.

[7] 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.

[8] 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.

[9] N. Gast, B. Gaujal, and C. Yan, "Linear program-based policies for restless bandits: Necessary and sufficient conditions for (exponentially fast) asymptotic optimality," *Mathematics of Operations Research*, 2023.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.
[19] D. P. Bertsekas *et al.*, "Dynamic programming and optimal control, 4th edition," *Belmont, MA: Athena Scientific*, vol. 1, 2011.
[20] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," *Advances in neural information processing systems*, vol. 21, 2008.
[21] A. Nedic and A. Ozdaglar, "Subgradient methods in network resource allocation: Rate analysis," in *IEEE CISS*, 2008, pp. 1189–1194.
[22] K. E. Avrachenkov and V. S. Borkar, "Whittle index based Q-learning for restless bandits with average reward," *Automatica*, vol. 139, p. 110186, 2022.
[23] 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.
[24] D. P. Dubhashi and A. Panconesi, *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] 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.

APPENDIX A
PROOF OF LEMMA 1

The L_1-deviation of the actual distribution and the empirical distribution over m distinct events from n samples is bounded by [23, Theorem 2.1]

$$\Pr(|p-\hat{p}|_1 \geq \alpha) \leq (2^m-2)e^{-\frac{n\alpha^2}{2}}. \tag{30}$$

We will utilize this result to compare the actual transition probabilities $P_n(\cdot|s,a)$ and the optimistic transition probabilities $P_n^k(\cdot|s,a)$ for all (s,a). In this sequel, we get

$$\Pr(|P_n^k(\cdot|s,a)-P_n(\cdot|s,a)|_1 \geq \alpha) \leq (2^{|\mathcal{S}|}-2)e^{-\frac{n\nu^2}{2}}. \tag{31}$$

Let $\nu = \sqrt{\frac{2}{n}\log\left(2^{|S|}|\mathcal{S}||\mathcal{A}|N\frac{k^2}{\epsilon}\right)}$. Substituting ν in (31) yields

$$\begin{aligned}\Pr(P_n^k(\cdot|c,\delta)-P_n(\cdot|c,\delta)|_1 \geq& \sqrt{\frac{2}{n}\log\left(2^{|S|}|\mathcal{S}||\mathcal{A}|N\frac{k^2}{\epsilon}\right)}\\ &\leq 2^{|\mathcal{S}|}e^{-\log\left(2^{|S|}|\mathcal{S}||\mathcal{A}|N\frac{k^2}{\epsilon}\right)}\\ &= \frac{\epsilon}{|\mathcal{S}||\mathcal{A}Nk^2}\end{aligned} \tag{32}$$

Denote $n = \max\{1, B_n^k(s,a)\}$ for all (s,a). Next, by taking summation over all states s, actions a and users $n \in \{1,2,\ldots,N\}$ implies

$$\Pr(\mathbf{P} \notin \mathcal{B}_p^k) \leq \frac{\epsilon}{k^2}. \tag{33}$$

Next, applying Chernoff-Hoeffding inequality [24], we get

$$\begin{aligned}&\Pr\left(|r_n(s,a)-\hat{r}_n^k(s,a)| > \rho_n^k(s,a)\right)\\ \leq& 2e^{-\left(\sqrt{\frac{2|\mathcal{S}|\log\left(|\mathcal{S}||\mathcal{A}|N\frac{k^2}{\epsilon}\right)}{\max(1,B_n^k(s,a))}}\right)^2\max(1,B_n^k(s,a))}\\ =& 2e^{-2|\mathcal{S}|\log\left(|\mathcal{S}||\mathcal{A}|N\frac{k^2}{\epsilon}\right)}\\ =& 2\left(\frac{\epsilon}{|\mathcal{S}||\mathcal{A}|Nk^2}\right)^{2|\mathcal{S}|}.\end{aligned} \tag{34}$$

By taking summation over all states s, actions a and users $n \in \{1,2,\ldots,N\}$ implies

$$\Pr(\mathbf{r} \notin \mathcal{B}_r^k) \leq \frac{2}{(|S||\mathcal{A}|N)^{2|S|-1}}\left(\frac{\epsilon}{k^2}\right)^{|S|}. \tag{35}$$

Therefore,

$$\Pr(\mathbf{P} \in \mathcal{B}_p^k \text{ and } \mathbf{r} \in \mathcal{B}_r^k) \geq 1-\frac{\epsilon}{k^2}-\frac{2}{(|S||\mathcal{A}|N)^{2|S|-1}}\left(\frac{\epsilon}{k^2}\right)^{|S|}. \tag{36}$$

This completes the proof.

APPENDIX B
PROOF OF THEOREM 1

We will analyze the regret for three distinct components in (26). To that end, define

$$\text{Reg}^{\pi}(k) = L(\pi^*,\mathbf{P},\mathbf{r},\mu^*,\boldsymbol{\lambda}^*,\mathbf{s}_1) - L(\pi^k,\mathbf{P},\mathbf{r},\mu^*,\boldsymbol{\lambda}^*,\mathbf{s}_1), \tag{37}$$

$$\text{Reg}^{\lambda}(k) = L(\pi^k,\mathbf{P},\mathbf{r},\mu^*,\boldsymbol{\lambda}^*,\mathbf{s}_1) - L(\pi^k,\mathbf{P},\mathbf{r},\mu^*,\boldsymbol{\lambda}^k,\mathbf{s}_1), \tag{38}$$

$$\text{Reg}^{\mu}(k) = L(\pi^k,\mathbf{P},\mathbf{r},\mu^*,\boldsymbol{\lambda}^k,\mathbf{s}_1) - L(\pi^k,\mathbf{P},\mathbf{r},\mu^k,\boldsymbol{\lambda}^k,\mathbf{s}_1). \tag{39}$$

Therefore, $\text{Reg}(k)$ becomes

$$\begin{aligned}\text{Reg}(k) =& \text{Reg}^{\pi}(k) + \text{Reg}^{\lambda}(k) + \text{Reg}^{\mu}(k)\\ \leq& |\text{Reg}^{\pi}(k)| + |\text{Reg}^{\lambda}(k)| + |\text{Reg}^{\mu}(k)|.\end{aligned} \tag{40}$$

First, we analyze $\text{Reg}^{\pi}(k)$. In this sequel, we introduce the following lemma:

Lemma 3. *Given initial state $(s_1(1), s_2(1), \ldots, s_N(1))$, the following holds with probability $1-\epsilon$:*

$$\begin{aligned}&\text{Reg}^{\pi}(k) \leq\\ &\sum_{n=1}^{N} V_{n,1}^{P_n^k,r_n^k,\mu^*,\lambda_n^*}(s_n(1);\pi_n^k) - V_{n,1}^{P_n,r_n,\mu^*,\lambda_n^*}(s_n(1);\pi_n^k),\end{aligned} \tag{41}$$

where P_n^k and r_n^k 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, and $V_{n,1}^{P,r,\mu,\lambda}(\cdot;\pi)$ is the value function associated with policy π.

Proof. See Appendix C. □

When the actual transition probability and reward lies within the confidence ball, i.e., $\mathbf{P} \in \mathcal{B}_p^k$ and $\mathbf{r} \in \mathcal{B}_r^k$, from Lemma 3, we can write

$$\begin{aligned}
&\text{Reg}^{\pi}(k)\\
&\leq \sum_{n=1}^{N} V_{n,1}^{P_n^k, r_n^k, \mu^*, \lambda_n^*}(s_n(1); \pi_n^k) - V_{n,1}^{P_n, r_n, \mu^*, \lambda_n^*}(s_n(1); \pi_n^k)\\
&= \sum_{n=1}^{N} \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H} r_n^k(s_n(t), a_n(t)) - r_n(s_n(t), a_n(t))\\
&\quad \Bigg| s_n(t)\Bigg] + \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H}\sum_{s'}\Bigg(P_n^k(s'|s_n(t), a_n(t))\\
&\quad - P_n(s'|s_n(t), a_n(t))\Bigg) V_{n,t+1}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s')\Bigg], \qquad (42)
\end{aligned}$$

where (42) holds from [28, Lemma 10]. Continuing from (42), we can write

$$\begin{aligned}
\text{Reg}^{\pi}(k) \leq \sum_{n=1}^{N} \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H} \rho_n^k(s_n(t), a_n(t))\Bigg] + \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H}\sum_{s' \in \mathcal{S}}\\
\Bigg|P_n^k(s'|s_n(t), a_n(t)) - P_n(s'|s_n(t), a_n(t))\Bigg| V_{\max}\Bigg], \qquad (43)
\end{aligned}$$

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

$$\begin{aligned}
\text{Reg}^{\pi}(k) &\leq \sum_{n=1}^{N}(1+V_{\max})\mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H} \rho_n^k(s_n(t), a_n(t))\Bigg]\\
&= \sum_{n=1}^{N}(1+V_{\max})\mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H}\sqrt{\frac{2|\mathcal{S}|\log(2|\mathcal{S}||\mathcal{A}|N\frac{k^2}{\epsilon})}{\max(1, B_n^k(s_n(t), a_n(t)))}}\Bigg]\\
&= \sum_{n=1}^{N}(1+V_{\max})\sqrt{2|\mathcal{S}|\log(2|\mathcal{S}||\mathcal{A}|N\frac{k^2}{\epsilon})}\\
&\quad \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{t=1}^{H}\sqrt{\frac{1}{\max(1, B_n^k(s_n(t), a_n(t)))}}\Bigg]\\
&\leq \sum_{n=1}^{N}\frac{1}{\epsilon}(1+V_{\max})\sqrt{2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|NK)}\\
&\quad \mathbb{E}_{P_n, \pi_n^k}\Bigg[\sum_{s \in S, a \in A}\frac{\delta_n^k(s,a)}{\max(1, B_n^k(s_n(t), a_n(t)))}\Bigg], \qquad (44)
\end{aligned}$$

where $\delta_n^k(s,a)$ 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 [16, Lemma E.3] to (44) yields

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\delta_n^k(s,a)}{\sqrt{\max(1, B_n^k(s,a))}}\\
&\leq \Bigg(\sqrt{H+1}+1\Bigg)\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{B_n^k(s,a)}\\
&\leq \Bigg(\sqrt{H+1}+1\Bigg)\sqrt{|\mathcal{S}||\mathcal{A}|KH} \qquad (45)
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{k=1}^{K}\text{Reg}^{\pi}(k)\\
&\leq \sum_{n=1}^{N}\frac{1}{\epsilon}(1+V_{\max})\sqrt{2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|NK)}(\sqrt{H+1}+1)\\
&\qquad \sqrt{|\mathcal{S}||\mathcal{A}|KH}\\
&\leq O\Bigg(V_{\max}|\mathcal{S}|\sqrt{|\mathcal{A}|}NH\sqrt{K\log K}\Bigg). \qquad (46)
\end{aligned}$$

The regret outside the confidence ball vanishes with probability

$$\begin{aligned}
&\Pr(\mathbf{P} \in \mathcal{B}_p^k) \text{ and } \mathbf{r} \in \mathcal{B}_r^k, \forall\sqrt{K} \leq k \leq K)\\
&= 1 - \Pr(\mathbf{P} \in \mathcal{B}_p^k) - \Pr(\mathbf{r} \in \mathcal{B}_r^k)\\
&\geq 1 - \sum_{\sqrt{K}\leq k\leq K}\frac{\epsilon}{k^2} - \sum_{\sqrt{K}\leq k\leq K}\frac{m\epsilon^{2|\mathcal{S}|}}{k^{4|\mathcal{S}|}} = 1 - \int_{\sqrt{K}}^{K}\frac{\epsilon}{k^2} - \int_{\sqrt{K}}^{K}\frac{m\epsilon^{2|\mathcal{S}|}}{k^{4|\mathcal{S}|}}\\
&= 1 - \epsilon\Bigg(\frac{1}{\sqrt{K}} - \frac{1}{K}\Bigg) - m\epsilon^{2|\mathcal{S}|}\Bigg(\frac{1}{K^{2|\mathcal{S}|-(1/2)}} - \frac{1}{K^{4|\mathcal{S}|-1}}\Bigg), \qquad (47)
\end{aligned}$$

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 $1 - O(\epsilon)$.

Next, we analyze $\text{Reg}^{\lambda}(k)$. From (7), we can write

$$\begin{aligned}
&\text{Reg}^{\lambda}(k)\\
&= L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^k, \mathbf{s}_1)\\
&= \sum_{n=1}^{N}(\lambda_n^* - \lambda_n^k)\Bigg(\frac{1}{H}\sum_{t=1}^{H}\mathbb{E}[a_n^k(t)] - \beta_n\Bigg)\\
&= \sum_{n=1}^{N}(\lambda_n^* - \lambda_n^k)g_{\lambda,n}^k, \qquad (48)
\end{aligned}$$

where $g_{\lambda,n}^k = \frac{1}{H}\sum_{t=1}^{H}\mathbb{E}[a_n^k(t)] - \beta_n$ represents the gradient associated with the update rule of λ_n^k in (25) and

$$|\text{Reg}^{\lambda}(k)| = \Bigg|\sum_{n=1}^{N}(\lambda_n^* - \lambda_n^k)g_{\lambda,n}^k\Bigg| = \Bigg|\sum_{n=1}^{N}(\lambda_n^k - \lambda_n^*)g_{\lambda,n}^k\Bigg|. \qquad (49)$$

Lemma 4. *The following holds*

$$\sum_{n=1}^{N}\sum_{k=1}^{K}(\lambda_n^k-\lambda_n^*)g_{\lambda,n}^k\leq \sum_{n=1}^{N}\frac{1}{2\gamma_\lambda}|\lambda_n^*-\lambda_n^1|^2+\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\gamma_\lambda}{2}(g_{\lambda,n}^k)^2. \tag{50}$$

Proof. See Appendix D. □

Substituting (49) into (50), we get

$$\sum_{k=1}^{K}|\text{Reg}^\lambda(k)| \leq \frac{1}{2\gamma_\lambda}\sum_{n=1}^{N}|\lambda_n^*-\lambda_n^1|^2+\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\gamma_\lambda}{2}(g_{\lambda,n}^k)^2. \tag{51}$$

Our algorithm sets $\lambda_n^1=0$ and λ_n^* is bounded. Hence, by substituting $\gamma_\lambda=\frac{c_1}{\sqrt{K}}$ for any $c_1>0$, we have

$$\frac{1}{2\gamma_\lambda}\sum_{n=1}^{N}|\lambda_n^*-\lambda_n^1|^2\leq O\bigg(\sqrt{K}\bigg). \tag{52}$$

Moreover, the gradient $g_{\lambda,n}^k$ is upper bounded by $1-\beta_n$. Thus, we get

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\gamma_\lambda}{2}(g_{\lambda,n}^k)^2\leq\frac{N\sqrt{K}}{2c_1}. \tag{53}$$

Utilizing (52) and (53) in (51) yields

$$\sum_{k=1}^{K}|\text{Reg}^\lambda(k)|\leq O\bigg(N\sqrt{K}\bigg). \tag{54}$$

Finally, we analyze $\text{Reg}^\mu(k)$. From (7), we can write

$$\begin{aligned}&\text{Reg}^\mu(k)\\ =&L(\pi^k,\mathbf{P},\mathbf{r},\mu^*,\boldsymbol{\lambda}^k,\mathbf{s}_1)-L(\pi^k,\mathbf{P},\mathbf{r},\mu^k,\boldsymbol{\lambda}^k,\mathbf{s}_1)\\ \leq&\frac{C}{H}\sum_{j=1}^{\lfloor H/C\rfloor}\sum_{n=1}^{N}(\mu^k(j)-\mu^*)\bigg(\frac{1}{C}\sum_{t=(jC+1)}^{(j+1)C}\mathbb{E}[a_{\mu,n}^k(t)]-M\bigg)\\ =&\frac{C}{H}\sum_{j=1}^{\lfloor H/C\rfloor}\sum_{n=1}^{N}(\mu^k(j)-\mu^*)g_{\mu,n}^k,\end{aligned} \tag{55}$$

where $g_{\mu,n}^k=\frac{1}{C}\sum_{t=jC+1}^{(j+1)C}\mathbb{E}[a_{\mu,n}^k(t)]-M$ represents the gradient associated the with update rule of $\mu^k(j)$ in (23) and

$$\begin{aligned}|\text{Reg}^\mu(k)|\leq&\bigg|\frac{C}{H}\sum_{j=1}^{\lfloor H/C\rfloor}\sum_{n=1}^{N}(\mu^k(j)-\mu^*)g_{\mu,n}^k\bigg|\\ =&\bigg|\frac{C}{H}\sum_{j=1}^{\lfloor H/C\rfloor}\sum_{n=1}^{N}(\mu^*-\mu^k(j)g_{\mu,n}^k\bigg|.\end{aligned} \tag{56}$$

Similar to Lemma 4, we get the following lemma.

Lemma 5. *The following holds*

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{j=1}^{\lfloor H/C\rfloor}(\mu^*-\mu^k(j))g_{\mu,n}^k\leq \sum_{n=1}^{N}\frac{1}{2\gamma_\mu}|\mu^1(1)-\mu^*|^2+\sum_{k=1}^{K}\sum_{n=1}^{N}\lfloor\frac{H}{C}\rfloor\frac{\gamma_\mu}{2}(g_{\mu,n}^k)^2. \tag{57}$$

Proof. See Appendix E. □

Substituting (56) into (57), we get

$$\sum_{k=1}^{K}|\text{Reg}^\mu(k)| \leq\frac{C}{H}\sum_{n=1}^{N}\frac{1}{2\gamma_\mu}|\mu^k(1)-\mu^*|^2+\sum_{k=1}^{K}\sum_{n=1}^{N}\frac{\gamma_\mu}{2}(g_{\mu,n}^k)^2. \tag{58}$$

Our algorithm sets $\mu^1(1)=0$ and μ^* is bounded. Hence, by substituting $\gamma_\mu=\frac{c_2}{\sqrt{K}}$ for any $c_2>0$, we have

$$\sum_{n=1}^{N}\frac{1}{2\gamma_\mu}|\mu^1(1)-\mu^*|^2\leq O\bigg(\frac{\sqrt{K}C}{H}\bigg). \tag{59}$$

Moreover, the gradient $g_{\mu,n}^k$ is upper bounded by $1-M$. Thus, we get

$$\sum_{n=1}^{N}\frac{\gamma_\mu}{2}(g_{\mu,n}^k)^2\leq\frac{N\sqrt{K}}{2c_2}. \tag{60}$$

Utilizing (59) and (60) in (58) yields

$$\sum_{k=1}^{K}|\text{Reg}^\mu(k)|\leq O\bigg(\frac{N\sqrt{K}C}{H}\bigg). \tag{61}$$

From (46), (54), and (61), we get the cumulative regret of all K episodes, which is given by

$$\begin{aligned}\text{Reg}(K)\leq O\big(&V_{\max}|\mathcal{S}|\sqrt{|\mathcal{A}|}N\sqrt{K\log K}\\ &+N\sqrt{K}(1+C/H)\big).\end{aligned} \tag{62}$$

This completes the proof of Theorem 1.

Appendix C
Proof of Lemma 3

For any policy π, we have from (21)

$$\begin{aligned}L(\pi,\mathbf{P},\mathbf{r},\mu^k,\boldsymbol{\lambda}^k,\mathbf{s}_1)=&\sum_{n=1}^{N}V_{n,1}^{P_n,r_n,\mu^k,\lambda_n^k}(s_n(1);\pi_n)\\ &-1/T\sum_{t=1}^{T}\sum_{n=1}^{N}\lambda_n^k\beta_n+1/T\sum_{t=1}^{T}\sum_{n=1}^{N}\mu^kM,\end{aligned} \tag{63}$$

where $V_{n,1}^{P,r,\mu,\lambda}(\cdot)$ is the optimal value function associated with the parameters P,r,μ,λ.

Next, we show the following

$$\sum_{n=1}^{N} V_{n,1}^{P_n,r_n,\mu^k,\lambda_n^k}(s_n(1);\pi_n) \le \sum_{n=1}^{N} V_{n,1}^{P_n,r_n,\mu^k,\lambda_n^k}(s_n(1)) \le \sum_{n=1}^{N} V_{n,1}^{P_n^k,r_n^k,\mu^k,\lambda_n^k}(s_n(1)), \tag{64}$$

where the last inequality can be proven from (20) and backward induction.

For any $t = H, H-1, \ldots, 1$, we will show by backward induction the following holds

$$V_{n,t}^{P_n,r_n,\mu^k,\lambda_n^k}(s) \le V_{n,t}^{P_n^k,r_n^k,\mu^k,\lambda_n^k}(s). \tag{65}$$

It is trivial for $t = H+1$, as $V_{n,H+1}(s) = 0$ for all s. Let the following holds for $t = h+1$

$$V_{n,h+1}^{P_n,r_n,\mu^k,\lambda_n^k}(s) \le V_{n,h+1}^{P_n^k,r_n^k,\mu^k,\lambda_n^k}(s). \tag{66}$$

We have to show that if (66) holds, then the following is true:

$$V_{n,h}^{P_n,r_n,\mu^k,\lambda_n^k}(s) \le V_{n,h}^{P_n^k,r_n^k,\mu^k,\lambda_n^k}(s). \tag{67}$$

We can write

$$\begin{aligned}
&V_{n,h}^{P_n,r_n,\mu^k,\lambda_n^k}(s) - V_{n,h}^{P_n^k,r_n^k,\mu^k,\lambda_n^k}(s)\\
=&\max_a \Bigg(r_n(s,a) - r_n^k(s,a) + \sum_{s'} P_n(s'|s,a) V_{n,h+1}^{P_n,r_n}(s')\\
&- \max_{P_n^k \in \mathcal{B}_p^k} \sum_{s'} P_n^k(s'|s,a) V_{n,h+1}^{P_n^k,r_n^k}(s') \Bigg)\\
\le&\max_a \Bigg(\sum_{s'} P_n(s'|s,a) V_{n,h+1}^{P_n,r_n}(s')\\
&- \max_{P_n^k \in \mathcal{B}_p^k} \sum_{s'} P_n^k(s'|s,a) V_{n,h+1}^{P_n,r_n}(s') \Bigg),\\
\le&0,
\end{aligned} \tag{68}$$

where we utilize (66) and the fact that $\rho_n^k(s,a) \ge 0$. Hence, (65) is proven.

From (63) and (64), we can establish that

$$L(\pi, \mathbf{P}, \mathbf{r}, \mu^k, \boldsymbol{\lambda}^k, \mathbf{s}_1) \le L(\pi^k, \mathbf{P}^k, \mathbf{r}^k, \mu^k, \boldsymbol{\lambda}^k, \mathbf{s}_1). \tag{69}$$

When the confidence bound holds, we can write

$$\begin{aligned}
&\mathrm{Reg}^{\pi}(k)\\
=&L(\pi^*, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1)\\
\le&L(\pi^*, \mathbf{P}, \mathbf{r}, \mu^k, \boldsymbol{\lambda}^k, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) &(70)\\
\le&L(\pi^k, \mathbf{P}^k, \mathbf{r}^k, \mu^k, \boldsymbol{\lambda}^k, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) &(71)\\
\le&L(\pi^k, \mathbf{P}^k, \mathbf{r}^k, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) - L(\pi^k, \mathbf{P}, \mathbf{r}, \mu^*, \boldsymbol{\lambda}^*, \mathbf{s}_1) &(72)\\
=&\sum_{n=1}^{N} V_{n,1}^{P_n^k,r_n^k,\mu^*,\lambda_n^*}(s_n(1);\pi_n^k) - V_{n,1}^{P_n,r_n,\mu^*,\lambda_n^*}(s_n(1);\pi_n^k),
\end{aligned} \tag{73}$$

where (70) holds because $\boldsymbol{\lambda}^*, \mu^*$ minimizes $L(\pi^*, \mathbf{P}, \mathbf{r}, \mu, \boldsymbol{\lambda}, \mathbf{s}_1)$, (71) holds due to (69) and the fact that $\pi^k, \mathbf{P}^k, \mathbf{r}^k$ maximizes $L(\pi, \mathbf{P}, \mathbf{r}, \mu^k, \boldsymbol{\lambda}^k, \mathbf{s}_1)$, and (72) holds because $\boldsymbol{\lambda}^k, \mu^k$ minimizes $L(\pi^k, \mathbf{P}^k, \mathbf{r}^k, \mu, \boldsymbol{\lambda}, \mathbf{s}_1)$.

This completes the proof.

Appendix D
Proof of Lemma 4

From (25), we can write

$$\begin{aligned}
\sum_{n=1}^{N} |\lambda_n^* - \lambda_n^{k+1}|^2 \le& \sum_{n=1}^{N} \left| \lambda_n^* - \lambda_n^k + \gamma_\lambda g_{\lambda,n}^k \right|^2\\
=& \sum_{n=1}^{N} |\lambda_n^* - \lambda_n^k|^2 + 2\gamma_\lambda g_{\lambda,n}^k (\lambda_n^* - \lambda_n^k)\\
&+ \gamma_\lambda^2 (g_{\lambda,n}^k)^2.
\end{aligned} \tag{74}$$

Taking telescopic sum and dividing both sides of (74) with $2\gamma_\lambda$, we get

$$\begin{aligned}
\sum_{n=1}^{N} \frac{1}{2\gamma_\lambda} |\lambda_n^* - \lambda_n^{k+1}|^2 \le& \sum_{n=1}^{N} \frac{1}{2\gamma_\lambda} |\lambda_n^* - \lambda_n^1|^2 +\\
&\sum_{k=1}^{K} (\lambda_n^* - \lambda_n^k) g_{\lambda,n}^k + \sum_{k=1}^{K} \frac{\gamma_\lambda}{2} (g_{\lambda,n}^k)^2,
\end{aligned} \tag{75}$$

which yields

$$\begin{aligned}
\sum_{n=1}^{N} \sum_{k=1}^{K} (\lambda_n^k - \lambda_n^*) g_{\lambda,n}^k \le& \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\gamma_\lambda}{2} (g_{\lambda,n}^k)^2\\
&+ \frac{1}{2\gamma_\lambda} |\lambda_n^* - \lambda_n^1|^2,
\end{aligned} \tag{76}$$

from which Lemma 4 follows.

Appendix E
Proof of Lemma 5

From (23), we can write

$$\begin{aligned}
&\sum_{n=1}^{N} |\mu^k(j+1) - \lambda^*|^2\\
\le& \sum_{n=1}^{N} \left| \mu^k(j) + \gamma_\mu g_{\mu,n}^k - \mu^* \right|^2\\
=& \sum_{n=1}^{N} |\mu^k(j) - \mu^*|^2 + 2\gamma_\mu g_{\mu,n}^k (\mu^k(j) - \mu^*) + \gamma_\mu^2 (g_{\mu,n}^k)^2.
\end{aligned} \tag{77}$$

Taking telescopic sum and dividing both sides of (74) with $2\gamma_\mu$, we get

$$\begin{aligned}
\sum_{n=1}^{N} \frac{1}{2\gamma_\mu} |\mu^{K+1}(\lfloor \frac{H}{C} \rfloor + 1) - \mu^*|^2 \le& \sum_{n=1}^{N} \frac{1}{2\gamma_\mu} |\mu^1(1) - \mu^*|^2 +\\
&\sum_{k=1}^{K} \sum_{j=1}^{\lfloor H/C \rfloor} (\mu^k(j) - \mu^*) g_{\mu,n}^k + \frac{H}{C} \sum_{k=1}^{K} \frac{\gamma_\mu}{2} (g_{\mu,n}^k)^2,
\end{aligned} \tag{78}$$

which holds because we set $\mu^{k+1}(1) = \mu^k(\lfloor \frac{H}{C} \rfloor + 1)$. From (78), we get

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{j=1}^{\lfloor H/C \rfloor}(\mu^* - \mu^k(j))g_{\mu,n}^k \leq \sum_{k=1}^{K}\sum_{n=1}^{N}\frac{H}{C}\frac{\gamma_\mu}{2}(g_{\mu,n}^k)^2 + \sum_{n=1}^{N}\frac{1}{2\gamma_\mu}|\mu^k(1) - \mu^*|^2, \tag{79}$$

from which Lemma 5 follows.

Appendix F
Proof of Lemma 2

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

$$Q_{n,t}^k(s,a) = Q_{n,t}^{\hat{P}_n^k}(s,a) + \rho_n^k(s,a)V_{\max}. \tag{80}$$

We can write

$$\begin{aligned}
&Q_{n,t}^k(s,a) - Q_{n,t}^{P_n^k}(s,a)\\
=&Q_{n,t}^{\hat{P}_n^k}(s,a) + V_{\max}\rho_n^k(s,a) - Q_{n,t}^{P_n^k}(s,a)\\
=&\rho_n^k(s,a)(V_{\max})+\\
&\sum_{s'}\left(\hat{P}_n^k(s'|s,a)V_{n,t+1}^k(s') - P_n^k(s'|s,a)V_{n,t+1}^{P_n^k}(s')\right)\\
=&\rho_n^k(s,a)(V_{\max})+\\
&\sum_{s'}\Bigg(\hat{P}_n^k(s'|s,a)V_{n,t+1}^k(s') - P_n^k(s'|s,a)V_{n,t+1}^k(s')+\\
&P_n^k(s'|s,a)V_{n,t+1}^k(s') - P_n^k(s'|s,a)V_{n,t+1}^{P_n^k}(s')\Bigg),
\end{aligned} \tag{81}$$

In (81), we can apply the following relationship from [29]

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

which yields

$$\begin{aligned}
&Q_{n,t}^k(s,a) - Q_{n,t}^{P_n^k}(s,a) \geq\\
&\sum_{s'}\left(P_n^k(s'|s,a)\left\{V_{n,t+1}^k(s') - V_{n,t+1}^{P_n^k}(s')\right\}\right).
\end{aligned} \tag{82}$$

Next, we have to show that $V_{n,t}^k(s) \geq V_{n,t}^{P_n^k}(s)$ for any $t = H, H-1, \ldots, 1$. We will prove this result by backward induction. It is trivially true for $t = H+1$. Next, we assume that it is true for $t = h+1$, and will show that it is true for $t = h$. As

$$V_{n,h+1}^k(s) \geq V_{n,h+1}^{P_n^k}(s), \tag{83}$$

utilizing (83) in (82) yields

$$Q_{n,h}^k(s,a) \geq Q_{n,h}^{P_n^k}(s,a). \tag{84}$$

Since, $V_{n,h}^k(s) = \max_a Q_{n,h}^k(s,a)$, we get $V_{n,h}^k(s) \geq V_{n,h}^{P_n^k}(s)$. Hence, $V_{n,t}^k(s) \geq V_{n,t}^{P_n^k}(s)$ for any t. Utilizing the fact that $V_{n,t}^k(s) \geq V_{n,t}^{P_n^k}(s)$ for any t in (82), we get Lemma 2.

Appendix G
Proof of Theorem 2

From Lemma 3, we get

$$\text{Reg}(k) \leq \sum_{n=1}^{N} V_{n,1}^{P_n^k, r_n^k, \mu^k, \lambda_n^k}(s_n(1)) - V_{n,1}^{P_n, r_n, \mu^k, \lambda_n^k}(s_n(1)). \tag{85}$$

We show that $V_{n,t}^k(s) \geq V_{n,t}^{P_n^k}(s)$ in the proof Lemma 2 in Appendix F, where $V_{n,t}^k(s) = \max_a Q_{n,t}^k(s,a)$, and $Q_{n,t}^k(s,a)$ is defined in (80). Using this in (85) implies

$$\begin{aligned}
\text{Reg}(k) \leq &\sum_{n=1}^{N} V_{n,1}^{\hat{P}_n^k, r_n^k, \mu^k, \lambda_n^k}(s_n(1)) - V_{n,1}^{P_n, r_n, \mu^k, \lambda_n^k}(s_n(1)) +\\
&\mathbb{E}_{\mathcal{F}_k, \pi^k}\rho_n^k(s,a)V_{\max}.
\end{aligned} \tag{86}$$

Next, applying the similar argument as presented in Appendix B with the additional constant term in (86), (29) follows. In particular, adapting the above with respect to the filtration $\mathcal{F}_k$ till episode k, we can apply the standard concentration lemma. □