Online Restless Multi-Armed Bandits with Long-Term Fairness Constraints

Shufan Wang, Guojun Xiong, Jian Li

Stony Brook University
{shufan.wang, guojun.xiong, jian.li.3}@stonybrook.edu

Abstract

Restless multi-armed bandits (RMAB) have been widely used to model sequential decision making problems with constraints. The decision maker (DM) aims to maximize the expected total reward over an infinite horizon under an "instantaneous activation constraint" that at most B arms can be activated at any decision epoch, where the state of each arm evolves stochastically according to a Markov decision process (MDP). However, this basic model fails to provide any fairness guarantee among arms. In this paper, we introduce RMAB-F, a new RMAB model with "long-term fairness constraints", where the objective now is to maximize the long-term reward while a minimum long-term activation fraction for each arm must be satisfied. For the online RMAB-F setting (i.e., the underlying MDPs associated with each arm are unknown to the DM), we develop a novel reinforcement learning (RL) algorithm named Fair-UCRL. We prove that Fair-UCRL ensures probabilistic sublinear bounds on both the reward regret and the fairness violation regret. Compared with off-the-shelf RL methods, our Fair-UCRL is much more computationally efficient since it contains a novel exploitation that leverages a low-complexity index policy for making decisions. Experimental results further demonstrate the effectiveness of our Fair-UCRL.

Introduction

The restless multi-armed bandits (RMAB) model (Whittle 1988) has been widely used to study sequential decision making problems with constraints, ranging from wireless scheduling (Sheng, Liu, and Saigal 2014; Cohen, Zhao, and Scaglione 2014), resource allocation in general (Glazebrook, Hodge, and Kirkbride 2011; Larrañaga, Ayesta, and Verloop 2014; Borkar, Ravikumar, and Saboo 2017), to healthcare (Bhattacharya 2018; Mate, Perrault, and Tambe 2021; Killian, Perrault, and Tambe 2021). In a basic RMAB setting, there is a collection of N "restless" arms, each of which is endowed with a state that evolves independently according to a Markov decision process (MDP) (Puterman 1994). If the arm is activated at a decision epoch, then it evolves stochastically according to one transition kernel, otherwise according to a different transition kernel. RMAB generalizes the Markovian multi-armed bandits (Lattimore and Szepesvári 2020) by allowing arms that are not activated to change state, which leads to "restless" arms, and hence extends its applicability. For simplicity, we refer to a restless arm as an arm in the rest of the paper. Rewards are generated with each transition depending on whether the arm is activated or not. The goal of the decision maker (DM) is to maximize the expected total reward over an infinite horizon under an "instantaneous activation constraint" that at most B arms can be activated at any decision epoch.

However, the basic RMAB model fails to provide any guarantee on how activation will be distributed among arms. This is also a salient design and ethical concern in practice, including mitigating data bias for healthcare (Mate, Perrault, and Tambe 2021; Li and Varakantham 2022a) and societal impacts (Yin et al. 2023; Biswas et al. 2023), providing quality of service guarantees to clients in network resource allocation (Li, Liu, and Ji 2019), just to name a few. In this paper, we introduce a new *RMAB model with fairness constraints*, dubbed as RMAB-F to address fairness concerns in the basic RMAB model. Specifically, we impose "long-term fairness constraints" into RMAB problems such that the DM must ensure a minimum long-term activation fraction for each arm (Li, Liu, and Ji 2019; Chen et al. 2020; D'Amour et al. 2020; Li and Varakantham 2022a), as motivated by aforementioned resource allocation and healthcare applications. The DM's goal now is to maximize the long-term reward while satisfying not only "*instantaneous* activation constraint" in *each decision epoch* but also "*long-term* fairness constraint" for *each arm*. Our objective is to develop *low-complexity* reinforcement learning (RL) algorithms with *order-of-optimal regret guarantees* to solve RMAB-F without knowing the underlying MDPs associated with each arm.

Though online RMAB has been gaining attentions, existing solutions cannot be directly applied to our online RMAB-F. First, existing RL algorithms including state-of-the-art colored-UCRL2 (Ortner et al. 2012) and Thompson sampling methods (Jung and Tewari 2019; Akbarzadeh and Mahajan 2022), suffer from an exponential computational complexity and regret bounds grow exponentially with the size of state space. This is because those need to repeatedly solve Bellman equations with an exponentially large state space for making decisions. Second, though much effort has been devoted to developing low-complexity RL algorithms with order-of-optimal regret for online RMAB, many challenges remain unsolved. For example, multi-timescale stochastic approximation algorithms (Fu et al. 2019; Avrachenkov and Borkar 2022) suffer from slow convergence and have no regret guarantee. Adding to these limitations is the fact that none of them were designed with fairness constraints in mind, e.g., (Wang, Huang, and Lui 2020; Xiong, Li, and Singh 2022; Xiong, Wang, and Li 2022; Xiong et al. 2022; Xiong and Li 2023) only focused on minimizing costs in RMAB, while the DM in our RMAB-F faces a **new dilemma** on how to manage the balance between maximizing the *long-term* reward and satisfying both *instantaneous* activation constraint and *long-term* fairness requirements. This adds a new layer of difficulty to designing low-complexity RL algorithms with order-of-optimal regret for RMAB that is already quite challenging.

To tackle this new dilemma, we develop Fair-UCRL, a novel RL algorithm for online RMAB-F. On one hand, we provide the first-ever regret analysis for online RMAB-F, and prove that Fair-UCRL ensures sublinear bounds (i.e., $\tilde{\mathcal{O}}(\sqrt{T})$) for both the reward regret (suboptimality of long-term rewards) and the fairness violation regret (suboptimality of long-term fairness violation) with high probability. On the other hand, Fair-UCRL is computationally efficient. This is due to the fact that Fair-UCRL contains a novel exploitation that leverages a low-complexity index policy for making decisions, which differs dramatically from aforementioned off-the-shelf RL algorithms that make decisions via solving complicated Bellman equations. Such an index policy in turn guarantees that the instantaneous activation constraint can be always satisfied in each decision epoch. To the best of our knowledge, Fair-UCRL is the first model-based RL algorithm that simultaneously provides (i) order-of-optimal regret guarantees on both the reward and fairness constraints; and (ii) a low computational complexity, for RMAB-F in the online setting. Finally, experimental results on real-world applications (resource allocation and healthcare) show that Fair-UCRL effectively guarantees fairness for each arm while ensures good regret performance.

Model and Problem Formulation

In this section, we provide a brief overview of the conventional RMAB, and then formally define our RMAB-F as well as the online settings considered in this paper.

Restless Multi-Armed Bandits

A RMAB problem consists of a DM and N arms (Whittle 1988). Each arm $n \in \mathcal{N} = \{1, ..., N\}$ is described by a unichain MDP M_n (Puterman 1994). Without loss of generality (W.l.o.g.), all MDPs $\{M_n, \forall n \in \mathcal{N}\}$ share the same finite state space $\mathcal{S}$ and action space $\mathcal{A} := \{0, 1\}$, but may have different transition kernels $P_n(s'|s, a)$ and reward functions $r_n(s, a)$, $\forall s, s' \in \mathcal{S}, a \in \mathcal{A}$. Denote the cardinalities of $\mathcal{S}$ and $\mathcal{A}$ as S and A, respectively. The initial state is chosen according to the initial state distribution $\boldsymbol{s}_0$ and T is the time horizon. At each time/decision epoch t, the DM observes the state of each arm n, denoted by $s_n(t)$, and activates a subset of B arms. Arm n is called *active* when being activated, i.e., $a_n(t) = 1$, and otherwise *passive*, i.e., $a_n(t) = 0$. Each arm n generates a stochastic reward $r_n(t) := r_n(s_n(t), a_n(t))$, depending on its state $s_n(t)$ and action $a_n(t)$. W.l.o.g., we assume that $r_n \in [0, 1]$ with mean $\bar{r}_n(s, a), \forall n, s, a$, and only active arms generate reward, i.e., $r_n(s, 0) = 0, \forall n, s$. Denote the sigma-algebra generated by random variables $\{(s_n(\tau), a_n(\tau)), \forall n, \tau < t\}$ as $\mathcal{F}_t$. The goal of the DM is to design a control policy $\pi : \mathcal{F}_t \mapsto \mathcal{A}^N$ to maximize the total expected reward, which can be expressed as $\liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=1}^{T}\sum_{n=1}^{N} r_n(t)$, under the "instantaneous activation constraint", i.e., $\sum_{n=1}^{N} a_n(t) \leq B, \forall t$.

RMAB with Long-Term Fairness Constraints

In addition to maximizing the long-term reward, ensuring long-term fairness among arms is also important for real-world applications (Yin et al. 2023). As motivated by applications in network resource allocation and healthcare (Li, Liu, and Ji 2019; Li and Varakantham 2022a), we impose a "long-term fairness constraint" on a minimum long-term activation fraction for each arm, i.e., $\liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=1}^{T} a_n(t) \geq \eta_n, \forall n \in \mathcal{N}$, where $\eta_n \in (0, 1)$ indicates the minimum fraction of time that arm n should be activated. To this end, the objective of RMAB-F is now to maximize the total expected reward while ensuring that both "instantaneous activation constraint" at each epoch and "long-term fairness constraint" for each arm are satisfied. Specifically, RMAB-F$(P_n, r_n, \forall n)$ is defined as:

$$\max_{\pi} \ \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{n=1}^{N} r_n(t)\right] \qquad (1)$$

$$\text{s.t.} \ \sum_{n=1}^{N} a_n(t) \leq B, \quad \forall t, \qquad (2)$$

$$\liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} a_n(t)\right] \geq \eta_n, \quad \forall n. \qquad (3)$$

Assumption 1. *We assume that the RMAB-F problem of (1)-(3) is feasible, i.e., there exists a policy π such that constraints (2) and (3) are satisfied.*

Note that in this paper, we only consider learning feasible RMAB-F by this assumption. When the underlying MDPs (i.e., P_n and r_n) associated with each arm $n \in \mathcal{N}$ are known to the DM, we can compute the offline optimal policy π^{opt} by treating the offline RMAB-F as an infinite-horizon average cost per stage problem using relative value iteration (Puterman 1994). However, it is well known that this approach suffers from the curse of dimensionality due to the explosion of state space (Papadimitriou and Tsitsiklis 1994).

Online Settings

We focus on online RMAB-F, where the DM repeatedly interacts with N arms $\{M_n = \{\mathcal{S}, \mathcal{A}, P_n, r_n\}, \forall n \in \mathcal{N}\}$ in an episodic manner. Specifically, the time horizon T is divided into K episodes and each episode consists of H consecutive frames, i.e., $T = KH$. The DM is not aware of the values of the transition kernel P_n and reward function $r_n, \forall n \in \mathcal{N}$. Instead, the DM estimates the transition kernels and reward

Algorithm 1: Fair-UCRL

```
Require: Initialize $C_n^0(s,a) = 0$, and $\hat{P}_n^0(s'|s,a) = 1/S, \forall n \in \mathcal{N}, s, s' \in \mathcal{S}, a \in \mathcal{A}$.
for $k = 1, 2, \cdots, K$ do
    // **Optimistic Planning** //
    Construct the set of plausible MDPs $\mathcal{M}^k$ as in (6);
    Relaxed the instantaneous activation constraint in RMAB-F ($\tilde{P}_n^k, \tilde{r}_n^k, \forall n$) to be "long-term activation constraint", and transform it into ELP($\mathcal{M}^k, z^k$) in (7);
    // **Policy Execution** //
    Establish the FairRMAB index policy $\pi^{k,*}$ on top of the solutions to the ELP and execute it.
end for
```

functions in an online manner by observing the trajectories over episodes. As a result, it is not possible for a learning algorithm to unconditionally guarantee constraint satisfaction in (2) and (3) over a finite number of episodes. To this end, we measure the performance of a learning algorithm with policy π using two types of *regret*.

First, the regret of a policy π with respect to the long-term reward against the offline optimal policy π^{opt} is defined as

$$\Delta_T^R := TV^{\pi^{opt}} - \mathbb{E}_\pi\left[\sum_{t=1}^{T}\sum_{n=1}^{N} r_n(t)\right], \tag{4}$$

where $V^{\pi^{opt}}$ is the long-term reward obtained under the offline optimal policy π^{opt}. Note that since finding π^{opt} for RMAB-F is intractable, we characterize the regret with respect to a feasible, asymptotically optimal index policy (see Theorem **??** in Wang, Xiong, and Li (2023)), similar to the regret definitions for online RMAB (Akbarzadeh and Mahajan 2022; Xiong, Wang, and Li 2022).

Second, the regret of a policy π with respect to the long-term fairness against the minimum long-term activation fraction η_n for each arm n, or simply the fairness violation is

$$\Delta_T^{n,F} := T\eta_n - \mathbb{E}_\pi\left[\sum_{t=1}^{T} a_n(t)\right], \forall n \in \mathcal{N}. \tag{5}$$

Fair-UCRL and Regret Analysis

We first show that it is possible to develop an RL algorithm for the computationally intractable RMAB-F problem of (1)-(3). Specifically, we leverage the popular UCRL (Jaksch, Ortner, and Auer 2010) to online RMAB-F, and develop an episodic RL algorithm named Fair-UCRL. On one hand, Fair-UCRL strictly meets the "instantaneous activation constraint" (2) at each decision epoch since it leverages a low-complexity index policy for making decisions at each decision epoch, and hence Fair-UCRL is computationally efficient. On the other hand, we prove that Fair-UCRL provides probabilistic sublinear bounds for both reward regret and fairness violation regret. To our best knowledge, Fair-UCRL is the first model-based RL algorithm to provide such guarantees for online RMAB-F.

The Fair-UCRL Algorithm

Fair-UCRL proceeds in episodes as summarized in Algorithm 1. Let τ_k be the start time of episode k. Fair-UCRL maintains two counts for each arm n. Let $C_n^{k-1}(s,a)$ be the number of visits to state-action pairs (s,a) until τ_k, and $C_n^{k-1}(s,a,s')$ be the number of transitions from s to s' under action a until τ_k. Each episode consists of two phases:

Optimistic planning. At the beginning of each episode, Fair-UCRL constructs a confidence ball that contains a set of plausible MDPs (Xiong, Wang, and Li 2022) for each arm $\forall n \in \mathcal{N}$ with high probability. The "center" of the confidence ball has the transition kernel and reward function that are computed by the corresponding empirical averages as: $\hat{P}_n^k(s'|s,a) = \frac{C_n^{k-1}(s,a,s')}{\max\{C_n^{k-1}(s,a),1\}}$, $\hat{r}_n^k(s,a) = \frac{\sum_{l=1}^{k-1}\sum_{h=1}^{H} r_n^l(s,a)\mathbb{1}(s_n^l(h)=s,a_n^l(h)=a)}{\max\{C_n^{k-1}(s,a),1\}}$. The "radius" of the confidence ball is set to be $\delta_n^k(s,a)$ according to the Hoeffding inequality. Hence the set of plausible MDPs in episode k is:

$$\mathcal{M}^k = \big\{M_n^k = (\mathcal{S}, \mathcal{A}, \tilde{P}_n^k, \tilde{r}_n^k) : |\tilde{P}_n^k(s'|s,a) - \hat{P}_n^k(s'|s,a)| \\ \leq \delta_n^k(s,a), \tilde{r}_n^k(s,a) = \hat{r}_n^k(s,a) + \delta_n^k(s,a)\big\}, \tag{6}$$

Fair-UCRL then selects an optimistic MDP $M_n^k, \forall n$ and an optimistic policy with respect to RMAB-F ($\tilde{P}_n^k, \tilde{r}_n^k, \forall n$). Since solving RMAB-F ($\tilde{P}_n^k, \tilde{r}_n^k, \forall n$) is intractable, we first relax the instantaneous activation constraint so as to achieve a "long-term activation constraint", i.e., the activation. It turns out that this relaxed problem can be equivalently transformed into a linear programming (LP) via replacing all random variables in the relaxed RMAB-F ($\tilde{P}_n^k, \tilde{r}_n^k, \forall n$) with the occupancy measure corresponding to each arm n (Altman 1999). Due to lack of knowledge of transition kernels and rewards, we further rewrite it as an extended LP (ELP) by leveraging *state-action-state occupancy measure* $z_n^k(s,a,s')$ to express confidence intervals of transition probabilities: given a policy π and transition functions $\tilde{P}_n^k$, the occupancy measure $z_n^k(s,a,s')$ induced by π and $\tilde{P}_n^k$ is that $\forall n, s, s', a, k$: $z_n^k(s,a,s') := \lim_{H\to\infty}\frac{1}{H}\mathbb{E}_\pi[\sum_{h=1}^{H-1}\mathbb{1}(s_n(h)=s, a_n(h)=a, s_n(h+1)=s')]$. The goal is to solve the extended LP as

$$z^{k,*} = \arg\min_{z^k} \textbf{ELP}(\mathcal{M}^k, z^k), \tag{7}$$

with $z^{k,*} := \{z_n^{k,*}(s,a,s'), \forall n \in \mathcal{N}\}$. We present more details on ELP in Wang, Xiong, and Li (2023).

Policy execution. We construct an index policy, which is feasible for the online RMAB-F ($\tilde{P}_n^k, \tilde{r}_n^k, \forall n$) as inspired by Xiong, Wang, and Li (2022). Specifically, we derive our index policy on top of the optimal solution $z^{k,*} = \{z_n^{k,*}, \forall n \in \mathcal{N}\}$. Since $\mathcal{A} = \{0,1\}$, i.e, an arm can be either active or passive at time t, we define the index assigned to arm n in state $s_n(t) = s$ at time t to be as

$$\omega_n^{k,*}(s) := \frac{\sum_{s'} z_n^{k,*}(s,1,s')}{\sum_{a,s'} z_n^{k,*}(s,a,s')}, \quad \forall n \in \mathcal{N}. \tag{8}$$

We call this *the fair index* and rank all arms according to their indices in (8) in a non-increasing order, and activate the set of B highest indexed arms, denoted as $\mathcal{N}(t) \subset \mathcal{N}$ such that $\sum_{n\in\mathcal{N}(t)} a_n^*(t) \leq B$. All remaining arms are kept passive at time t. We denote the resultant index-based policy, which we call the `FairRMAB` index policy as $\pi^{k,*} := \{\pi_n^{k,*}, \forall n \in \mathcal{N}\}$, and execute this policy in this episode. More discussions on the property of the `FairRMAB` index policy are provided in Wang, Xiong, and Li (2023).

Remark 1. *Although `Fair-UCRL` draws inspiration from the infinite-horizon UCRL (Jaksch, Ortner, and Auer 2010; Xiong, Wang, and Li 2022), there exist a major difference. `Fair-UCRL` modifies the principle of optimism in the face of uncertainty for making decisions which is utilized by UCRL based algorithms, to not only maximize the long-term rewards but also to satisfy the long-term fairness constraint in our `RMAB-F`. This difference is further exacerbated since the objective of conventional regret analysis, e.g., colored-UCRL2 (Ortner et al. 2012; Xiong, Wang, and Li 2022) for `RMAB` is to bound the reward regret, while due to the long-term fairness constraint, we also need to bound the fairness violation regret for each arm for `Fair-UCRL`, which will be discussed in details in Theorem 1. We note that the designs of our `Fair-UCRL` and the `FairRMAB` index policy are largely inspired by the LP based approach in Xiong, Wang, and Li (2022) for online `RMAB`. However, Xiong, Wang, and Li (2022) only considered the instantaneous activation constraint, and hence is not able to address the new dilemma faced by our online `RMAB-F`, which also needs to ensure the long-term fairness constraints. Finally, our `G-Fair-UCRL` with no fairness violation further distinguishes our work.*

Regret Analysis of `Fair-UCRL`

We now present our main theoretical results on bounding the regrets defined in (4) and (5), realizable by `Fair-UCRL`.

Theorem 1. *When the size of the confidence intervals $\delta_n^k(s,a)$ is built for $\epsilon \in (0,1)$ as*

$$\delta_n^k(s,a) = \sqrt{\frac{1}{2C_n^k(s,a)} \log\left(\frac{SAN(k-1)H}{\epsilon}\right)},$$

with probability at least $1-(\frac{\epsilon}{SANT})^{\frac{1}{2}}$, `Fair-UCRL` achieves the reward regret as:

$$\Delta_T^R = \tilde{\mathcal{O}}\left(B\epsilon\log T + (\sqrt{2}+2)\sqrt{SANT}\sqrt{\log\frac{SANT}{\epsilon}}\right),$$

and with probability at least $(1-(\frac{\epsilon}{SANT})^{\frac{1}{2}})^2$, `Fair-UCRL` achieves the fairness violation regret for each arm $\forall n \in \mathcal{N}$ as:

$$\Delta_T^{n,F} = \tilde{\mathcal{O}}\left(\eta_n\epsilon\log T + C_0 T_{Mix}^n\sqrt{SANT}\log\frac{SANT}{\epsilon}\right),$$

where B is the activation budget, ϵ is the constant defined to build confidence interval, T_{mix}^n is the mixing time of the true MDP associated with arm n, $C_0 = 4(\sqrt{2}+1)\big(\hat{n} + \frac{C\rho^{\hat{n}}}{1-\rho}\big)$ and $\hat{n} = \lceil\log_\rho C^{-1}\rceil$ with ρ and C being constants (see Corollary **??** *in Wang, Xiong, and Li (2023)).*

As discussed in Remark 1, the design of `Fair-UCRL` differs from UCRL type algorithms in several aspects. These differences further necessitate different proof techniques for regret analysis. First, we leverage the relative value function of Bellman equation for long-term average MDPs, which enables us to transfer the regret to the difference of relative value functions. Thus, only the first moment behavior of the transition kernels are needed to track the regret, while state-of-the-art (Wang, Huang, and Lui 2020) leveraged the higher order moment behavior of transition kernels for a specific MDP, which is hard for general MDPs. Closest to ours is Xiong, Wang, and Li (2022), which however bounded the reward regret under the assumption that the diameter D of the underlying MDP associated with each arm is known. Unfortunately, this knowledge is often unavailable in practice and there is no easy way to characterize the dependence of D on the number of arms N (Akbarzadeh and Mahajan 2022). Finally, in conventional regret analysis of RL algorithms for `RMAB`, e.g., Akbarzadeh and Mahajan (2022); Xiong, Li, and Singh (2022); Xiong, Wang, and Li (2022); Wang, Huang, and Lui (2020), only the reward regret is bounded. However, for our `RMAB-F` with long-term fairness among each arm, we also need to characterize the fairness violation regret, for which, we leverage the mixing time of the underlying MDP associated with each arm. This is one of our main theoretical contributions that differentiates our work.

We note that another line of works on constrained MDPs (CMDPs) either considered a similar extended LP approach (Kalagarla, Jain, and Nuzzo 2021; Efroni, Mannor, and Pirotta 2020) in a finite-horizon setting, which differ from our infinite-horizon setting, or are only with a long-term cost constraint (Singh, Gupta, and Shroff 2020; Chen, Jain, and Luo 2022), while our `RMAB` problem not only has a long-term fairness constraint, but also an instantaneous activation constraint that must be satisfied at each decision epoch. This makes their approach not directly applicable to ours.

Proof Sketch of Theorem 1

We present some lemmas that are essential to prove Theorem 1. Our proof consists of three steps: regret decomposition and regret characterization when the true MDPs are in the confidence ball or not. A key challenge lies in bounding the fairness violation regret, for which the decision variable is the action a in our `Fair-UCRL`, while most recent works, e.g., Efroni, Mannor, and Pirotta (2020); Xiong, Wang, and Li (2022); Akbarzadeh and Mahajan (2022) focused on the reward function of the proposed policy. This challenge differentiates the proof, especially on bounding the fairness violation regret when the true MDP belongs to the confidence ball. To start with, we first introduce a lemma for the decomposition of reward and fairness violation regrets:

Lemma 1. *The reward and fairness violation regrets of `Fair-UCRL` can be decomposed into the summation of k episodic regrets with a constant term with probability at least $1-(\frac{\epsilon}{SANT})^{\frac{1}{2}}$., i.e.*

$$\Delta_T^R\{\pi^{*,k}, \forall k\} \leq \sum_{k=1}^{K} \Delta_k^R\{\pi^{*,k}\} + \sqrt{\frac{1}{4}T\log\frac{SANT}{\epsilon}},$$

$$\Delta_T^{n,F}\{\pi^{*,k},\forall k\} \leq \sum_{k=1}^{K} \Delta_k^{n,F}\{\pi^{*,k}\} + \sqrt{\frac{1}{4} T \log \frac{SANT}{\epsilon}},$$

where Δ_k^R and $\Delta_k^{n,F}$ are the reward/ fairness violation regret in episode k under policy $\pi^{*,k}$.

Proof Sketch: With probability of at least $1-(\frac{\epsilon}{SANT})^{\frac{1}{2}}$, the difference between reward until time T and the episodic reward for all K episodes can be bounded with a constant term $\sqrt{\frac{1}{4}T\log\frac{SANT}{\epsilon}}$ via Chernoff-Hoeffding's inequality. This is in parallel with several previous works, e.g. Akbarzadeh and Mahajan (2022); Xiong, Wang, and Li (2022); Efroni, Mannor, and Pirotta (2020).

Proof Sketch of Fairness Violation Regret. The proof of fairness violation regret is one of our main theoretical contributions in this paper. To our best knowledge, this is the first result for online RMAB-F, i.e. with both instantaneous activation constraint and long-term fairness constraint. We now present two key lemmas which are essential to bound the fairness violation regret when combining with Lemma 1. First, we show that the fairness violation regret can be bounded when the transition and reward function of true MDP (denoted by M) does not belong to the confidence ball, i.e. $M \notin \mathcal{M}_k$.

Lemma 2. *The fairness violation regret for failing confidence ball for all K episodes is bounded by*

$$\sum_{k=1}^{K} \Delta_k^{n,F}\{\pi^{*,k},\forall k\}\mathbb{1}(M \notin \mathcal{M}_k) \leq \frac{1}{2}\eta_n \epsilon \log T.$$

Proof Sketch: With the probability of failing event $P(M \notin \mathcal{M}_k) \leq \frac{\epsilon}{kH}$, one can bound the fairness violation term since $\eta_n - a_n(t) \leq \eta_n$. The final bound is obtained by summing over all episodes.

Now, we present the dominated term in bounding the fairness violation regret when the true MDP belongs to the confidence ball.

Lemma 3. *The fairness violation regret when the true MDP belongs to the confidence ball in each episode k is bounded by*

$$\begin{aligned}&\sum_{k=1}^{K} \Delta_k^{n,F}\{\pi^{*,k},\forall k\}\mathbb{1}(M \notin \mathcal{M}_k)\\ \leq& C_0 T_{Mix}^{n}\Bigg((\sqrt{2}+1)\sqrt{SANT}\sqrt{\log\frac{SANT}{\epsilon}}\\ &\qquad + \frac{1}{2}\sqrt{T}\log\frac{SANT}{\epsilon}\Bigg) + \sqrt{T}\frac{C}{1-\rho}.\end{aligned}$$

Proof Sketch: We first define a new variable $\overline{F}_n(\pi^k,p) = \frac{1}{T}\lim_{T\to\infty}(\sum_{t=1}^{T} a_n(t)|\pi^k,p)$ as the long term average fairness variable under policy π^k for arm n with MDP that has the true transition probability matrix p. We show that the fairness violation regret when the true MDP belongs to confidence ball can be upper bounded by $\mathbb{E}\big[H\eta_n - \overline{F}_n(\pi_n^k,p)\big]$ with a constant term.

Next we introduce another variable close to $\overline{F}_n(\pi^k,p)$, that is $\overline{F}_n(\pi^k,\theta) = \frac{1}{T}\lim_{T\to\infty}(\sum_{t=1}^{T} a_n(t)|\pi^k,\theta)$ as the fairness variable under policy π in episode k for arm n with MDP whose transition matrix θ belongs to the confidence ball. By comparing the total variance norm of $\overline{F}_n(\pi^k,p)$ and $\overline{F}_n(\pi^k,\theta)$, we can upper bound $\mathbb{E}\big[H\eta_n - \overline{F}_n(\pi_n^k,p)\big]$ as $\beta_n^k(\pi_k) := 2\big(\hat{n}+\frac{C\rho^{\hat{n}}}{1-\rho}\big)\max_s\sum_a \pi_k(s,a)\delta_n^k(s,a)$, where π_k is the policy in episode k. In order to bound $\beta_n^k(\pi_k)$ with the expected number of counts of (s,a) pair in episode k $\mathbb{E}[c_n^k(s,a)]$, we leverage the mixing time T_{mix}^n.

The regret is further split into two terms, one of which $\sum_{k=1}^{K}\sum_{(s,a)}\sum_n \frac{c_n^k(s,a)}{C_n^{k-1}(s,a)}$ can be bounded as $(\sqrt{2}+1)\sqrt{SANT}$ through the induction of sequence summation, while the other term $\sum_{k=1}^{K}\sum_{(s,a)}\frac{\mathbb{E}[c_n^k(s,a)]-c_n^k(s,a)}{\sqrt{2C_n^{k-1}(s,a)}}$ can be upper bounded by $\sqrt{T}\sqrt{\frac{1}{4}\log\frac{SANT}{\epsilon}}$ via Azuma-Hoeffding's inequality, as it can be considered as a martingale difference sequence.

Proof Sketch of Reward Regret. Similar to the fairness violation regret, we first bound the reward regret when the MDP does not belong to the confidence ball.

Lemma 4. *The reward regret for failing the confidence ball for all K episodes is bounded by*

$$\sum_{k=1}^{K} \Delta_k^{R}\{\pi^{*,k},\forall k\}\mathbb{1}(M \notin \mathcal{M}_k) \leq B\epsilon\log T.$$

Proof Sketch: Similar to Lemma 2, the probability of failing confidence ball is bounded by $P(M \notin \mathcal{M}_k) \leq \frac{\epsilon}{kH}$. Summing over all episodes yields the bound.

We then present the dominated term in the reward regret.

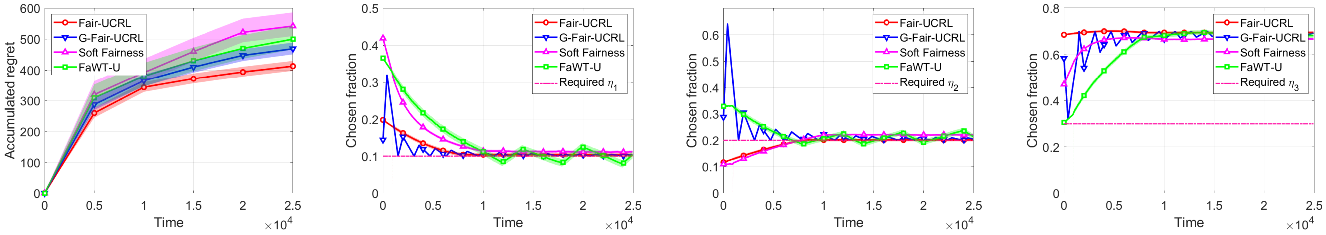
Lemma 5. *The reward regret when the true MDP belongs to the confidence ball in each episode k is bounded by*

$$\begin{aligned}&\sum_{k=1}^{K} \Delta_k^{R}\{\pi^{*,k},\forall k\}\mathbb{1}(M \in \mathcal{M}_k)\\ &\qquad\qquad \leq (\sqrt{2}+2)\sqrt{SANT}\sqrt{\log\frac{SANT}{\epsilon}}.\end{aligned}$$

Proof Sketch: We split the reward regret into two terms, $\sum_{(s,a)}\sum_n c_n^k(s,a)(\mu^*/B - \tilde{r}_n(s,a))$ and $\sum_{(s,a)}\sum_n c_n^k(s,a)2\sqrt{\frac{1}{2C_n^{k-1}(s,a)}\log\frac{SANkH}{\epsilon}}$. The first term is upper bounded by 0 due to the fact that for any episode k, the optimistic average reward $\tilde{r}_n(s,a)$ of the optimistically chosen policy $\tilde{\pi}_k$ within the confidence ball is equal or larger than the true optimal average reward μ^*, provided that the true MDP belongs to confidence ball. Similar to Lemma 3, the second term can be bounded with $(\sqrt{2}+1)\sqrt{SANT}$.

Experiments

In this section, we first evaluate the performance of Fair-UCRL in simulated environments, and then demonstrate the utility of Fair-UCRL by evaluating it under three real-world applications of RMAB.



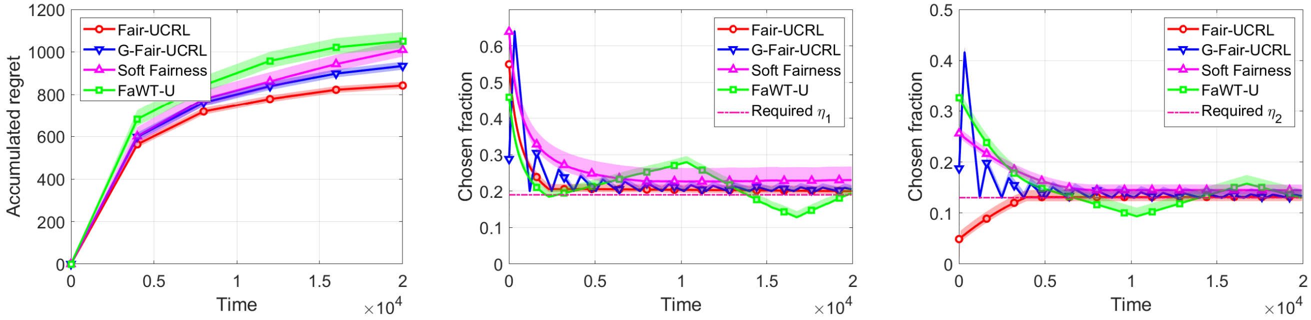
(a) Reward regret.

(b) Arm 1.

(c) Arm 2.

(d) Arm 3.

Figure 1: Evaluation in simulated environments: (a) Reward regret; and (b)-(d) Fairness constraint violation.



(a) Reward regret.

(b) Fairness constraint violation for arm 1.

(c) Fairness constraint violation for arm 2.

Figure 2: Continuous positive airway pressure therapy.

Evaluation in Simulated Environments

Settings. We consider 3 classes of arms, each including 100 duplicates with state space $\mathcal{S} \in \{0, 1, 2, 3, 4, 5\}$. Class-$n$ arm arrives with rate $\lambda_n = 3n$ for $n = 1, 2, 3$, and departs with a fixed rate of $\mu = 5$. We consider a controlled Markov chain in which states evolve as a specific birth-and-death process, i.e., state s only transits to $s + 1$ or $s - 1$ with probability $P(s, s + 1) = \lambda/(\lambda + \mu)$ or $P(s, s - 1) = \mu/(\lambda + \mu)$, respectively. Class-$n$ arm generates a random reward $r_n(s) \sim Ber(sp_n)$, with p_n uniformly sampled from $[0.01, 0.1]$. The activation budget is set to 100. The minimum activation fraction η is set to be 0.1, 0.2 and 0.3 for the three classes of arms, respectively. We set $K = H = 160$. We use Monte Carlo simulations with $1,000$ independent trials.

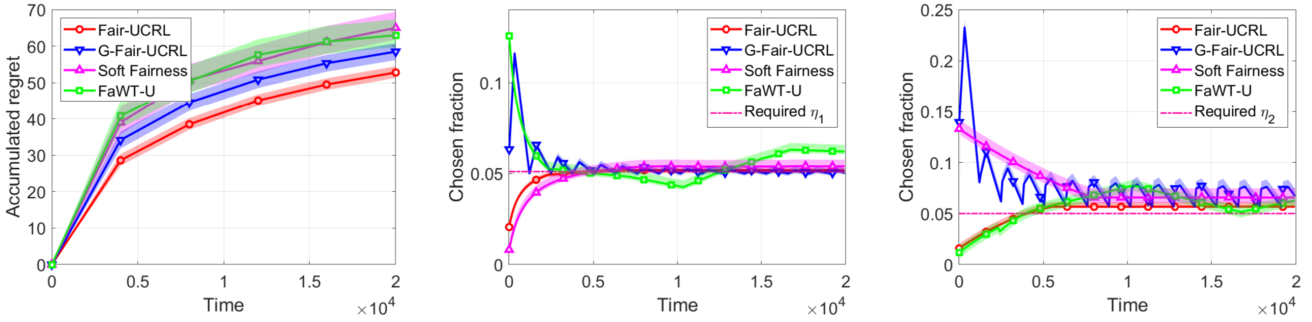
Baselines. (1) *FaWT-U* (Li and Varakantham 2022a) activates arms based on their Whittle indices. If the fairness constraint is not met for an arm after a certain time, FaWT-U always activates that arm regardless of its Whittle index. (2) *Soft Fairness* (Li and Varakantham 2022b) incorporates softmax based value iteration method into the RMAB setting. Since both algorithms are designed with discounted rewards, we choose the discounted factor to be 0.999 for fair comparisons with our Fair-UCRL, which is designed for infinite-horizon average-reward settings. (3) G-Fair-UCRL: We modify Fair-UCRL by greedily enforcing the fairness constraint satisfaction in each episode. Specifically, at the beginning of each episode, G-Fair-UCRL randomly pulls an arm to force each arm n to be pulled $H\eta_n$ times. This greedy exploration will take $\lceil \frac{\sum_{n=1}^{N} H\eta_n}{B} \rceil$ decision epochs in total in each episode. G-Fair-UCRL operates in the same manner as Fair-UCRL in the rest of this episode. See Wang, Xiong, and Li (2023) for details on G-Fair-UCRL.

Reward Regret. The accumulated reward regrets are presented in Figure 1a, where we use Monte Carlo simulations with $1,000$ independent trials. Fair-UCRL achieves the lowest accumulated reward regret. More importantly, this is consistent with our theoretical analysis (see Theorem 1), while neither FaWT-U nor Soft Fairness provides a finite-time analysis, i.e., nor provable regret bound guarantees.

Fairness Constraint Violation. The activation fraction for each arm over time under different policies are presented in Figures 1b, 1c and 1d, respectively. After a certain amount of time, the minimum activation fraction for each arm under Fair-UCRL is always satisfied, and a randomized initialization may cause short term fairness violation, for example, after $6,500$ time steps for arm 2, even though the constraint needs to be satisfied on average. Similar observations hold for Soft Fairness, while for FaWT-U, fairness constraint violation repeatedly occurs over time for arm 1 and arm 2.

Continuous Positive Airway Pressure Therapy

We study the continuous positive airway pressure therapy (CPAP) as in Herlihy et al. (2023); Li and Varakantham (2022b), which is a highly effective treatment when it is used consistently during the sleeping for adults with obstructive sleep apnea. Similar non-adherence to CPAP in patients hinders the effectiveness, we adapt the Markov model of CPAP adherence behavior (Kang et al. 2013) to a two-state system

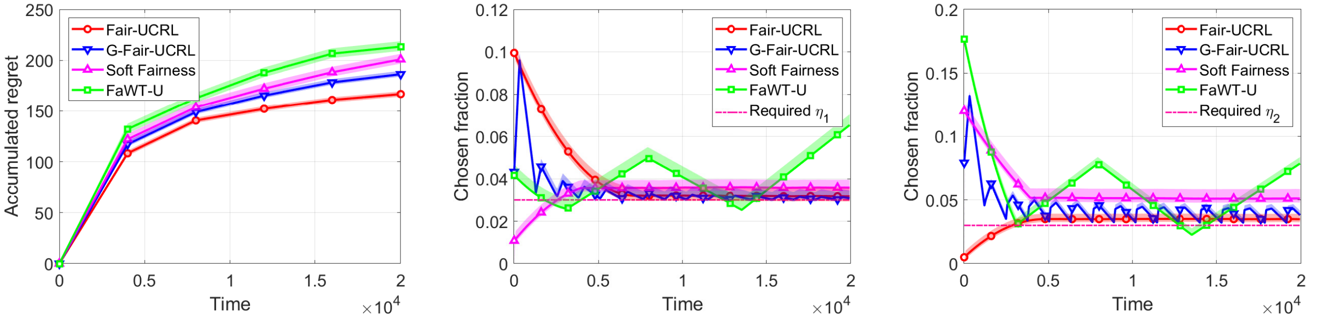


(a) Reward regret.

(b) Fairness constraint violation for arm 1.

(c) Fairness constraint violation for arm 2.

Figure 3: PASCAL recognizing textual entailment task.



(a) Reward regret.

(b) Fairness constraint violation for arm 1.

(c) Fairness constraint violation for arm 2.

Figure 4: Land mobile satellite system.

with the clinical adherence criteria. Specifically, there are 3 states, representing low, intermediate and acceptable adherence levels. Patients are clustered into two groups, "Adherence" and "Non-Adherence". The first group has a higher probability of staying in a good adherence level. There are 20 arms/patients with 10 in each group. The transition matrix of arms in each group contains a randomized, small noise from the original data. The intervention, which is the action applied to each arm, results in a 5% to 50% increase in adherence level. The budget is $B = 5$ and the fairness constraint is set to be a random number between [0.1, 0.7]. The objective is to maximize the total adherence level. The accumulated reward regret and the activation fraction for two randomly selected arms are presented in Figures 2a, 2b and 2c, respectively. Again, we observe that `Fair-UCRL` achieves a much smaller reward regret and the fairness constraint is always satisfied after a certain amount of time.

PASCAL Recognizing Textual Entailment

We study the PASCAL recognizing textual entailment task as in Snow et al. (2008). Workers are assigned with tasks that determine if *hypothesis* can be inferred from *text*. There are 10 workers. Due to lack of background information, a worker may not be able to correctly annotate a task. We assign a "successful annotation probability" to each worker, which is based on the average success rate over 800 tasks in the dataset. Each worker is a MDP with state 1 (correctly annotated) and 0 (otherwise). The transition probability from state 0 to 1 with $a = 1$ is the same as that of staying at state 1 with $a = 1$, which is set as the successful annotation probability. Reward is 1 if a selected worker successfully annotates the task, and 0 otherwise. At each time, 3 tasks are generated (i.e., $B = 3$) and distributed to workers. Fairness constraints for all workers are set to be $\eta = 0.05$. Again, both proposed algorithms outperform two baselines and maintain higher selection fraction as shown in Figures 3a, 3b and 3c for two randomly selected arms, respectively.

Land Mobile Satellite System

We study the land mobile satellite system problem as in Prieto-Cerdeira et al. (2010), in which the land mobile satellite broadcasts a signal carrying multimedia services to handheld devices. There are 4 arms with different elevation angles $(40°, 60°, 70°, 80°)$ of the antenna in urban area. Only two states (*Good* and *bad*) are considered and we leverage the same transition matrix as in Prieto-Cerdeira et al. (2010). Similar, we use the average direct signal mean as the reward function. The budget is $B = 2$. We apply the fairness constraint $\eta = 0.03$ to all angles. Again, `Fair-UCRL` outperforms the considered baselines in reward regret (Figure 4a), while satisfies long term average fairness constraint (Figures 4b and 4c for two randomly selected arms).

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) grants 2148309 and 2315614, and was supported in part by funds from OUSD R&E, NIST, and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program. This work was also supported in part by the U.S. Army Research Office (ARO) grant W911NF-23-1-0072, and the U.S. Department of Energy (DOE) grant DE-EE0009341. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Akbarzadeh, N.; and Mahajan, A. 2022. On learning Whittle index policy for restless bandits with scalable regret. *arXiv preprint arXiv:2202.03463*.

Altman, E. 1999. *Constrained Markov decision processes*, volume 7. CRC Press.

Avrachenkov, K. E.; and Borkar, V. S. 2022. Whittle index based Q-learning for restless bandits with average reward. *Automatica*, 139: 110186.

Bhattacharya, B. 2018. Restless bandits visiting villages: A preliminary study on distributing public health services. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 1–8.

Biswas, A.; Killian, J. A.; Diaz, P. R.; Ghosh, S.; and Tambe, M. 2023. Fairness for Workers Who Pull the Arms: An Index Based Policy for Allocation of Restless Bandit Tasks. *arXiv preprint arXiv:2303.00799*.

Borkar, V. S.; Ravikumar, K.; and Saboo, K. 2017. An index policy for dynamic pricing in cloud computing under price commitments. *Applicationes Mathematicae*, 44: 215–245.

Chen, L.; Jain, R.; and Luo, H. 2022. Learning Infinite-Horizon Average-Reward Markov Decision Processes with Constraints. *arXiv preprint arXiv:2202.00150*.

Chen, Y.; Cuellar, A.; Luo, H.; Modi, J.; Nemlekar, H.; and Nikolaidis, S. 2020. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, 181–190. PMLR.

Cohen, K.; Zhao, Q.; and Scaglione, A. 2014. Restless multi-armed bandits under time-varying activation constraints for dynamic spectrum access. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, 1575–1578. IEEE.

D'Amour, A.; Srinivasan, H.; Atwood, J.; Baljekar, P.; Sculley, D.; and Halpern, Y. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 525–534.

Efroni, Y.; Mannor, S.; and Pirotta, M. 2020. Exploration-Exploitation in Constrained MDPs. *arXiv preprint arXiv:2003.02189*.

Fu, J.; Nazarathy, Y.; Moka, S.; and Taylor, P. G. 2019. Towards q-learning the whittle index for restless bandits. In *2019 Australian & New Zealand Control Conference (ANZCC)*, 249–254. IEEE.

Glazebrook, K. D.; Hodge, D. J.; and Kirkbride, C. 2011. General notions of indexability for queueing control and asset management. *The Annals of Applied Probability*, 21(3): 876–907.

Herlihy, C.; Prins, A.; Srinivasan, A.; and Dickerson, J. P. 2023. Planning to fairly allocate: Probabilistic fairness in the restless bandit setting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 732–740.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-Optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(4).

Jung, Y. H.; and Tewari, A. 2019. Regret Bounds for Thompson Sampling in Episodic Restless Bandit Problems. *Proc. of NeurIPS*.

Kalagarla, K. C.; Jain, R.; and Nuzzo, P. 2021. A Sample-Efficient Algorithm for Episodic Finite-Horizon MDP with Constraints. In *Proc. of AAAI*.

Kang, Y.; Prabhu, V. V.; Sawyer, A. M.; and Griffin, P. M. 2013. Markov models for treatment adherence in obstructive sleep apnea. In *IIE Annual Conference. Proceedings*, 1592. Institute of Industrial and Systems Engineers (IISE).

Killian, J. A.; Perrault, A.; and Tambe, M. 2021. Beyond" To Act or Not to Act": Fast Lagrangian Approaches to General Multi-Action Restless Bandits. In *Proc.of AAMAS*.

Larrañaga, M.; Ayesta, U.; and Verloop, I. M. 2014. Index Policies for A Multi-Class Queue with Convex Holding Cost and Abandonments. In *Proc. of ACM Sigmetrics*.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Li, D.; and Varakantham, P. 2022a. Efficient Resource Allocation with Fairness Constraints in Restless Multi-Armed Bandits. In *Proc. of UAI*.

Li, D.; and Varakantham, P. 2022b. Towards Soft Fairness in Restless Multi-Armed Bandits. *arXiv preprint arXiv:2207.13343*.

Li, F.; Liu, J.; and Ji, B. 2019. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3): 1799–1813.

Mate, A.; Perrault, A.; and Tambe, M. 2021. Risk-Aware Interventions in Public Health: Planning with Restless Multi-Armed Bandits. In *Proc.of AAMAS*.

Ortner, R.; Ryabko, D.; Auer, P.; and Munos, R. 2012. Regret Bounds for Restless Markov Bandits. In *Proc. of Algorithmic Learning Theory*.

Papadimitriou, C. H.; and Tsitsiklis, J. N. 1994. The Complexity of Optimal Queueing Network Control. In *Proc. of IEEE Conference on Structure in Complexity Theory*.

Prieto-Cerdeira, R.; Perez-Fontan, F.; Burzigotti, P.; Bolea-Alamañac, A.; and Sanchez-Lago, I. 2010. Versatile two-state land mobile satellite channel model with first application to DVB-SH analysis. *International Journal of Satellite Communications and Networking*, 28(5-6): 291–315.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Sheng, S.-P.; Liu, M.; and Saigal, R. 2014. Data-Driven Channel Modeling Using Spectrum Measurement. *IEEE Transactions on Mobile Computing*, 14(9): 1794–1805.

Singh, R.; Gupta, A.; and Shroff, N. B. 2020. Learning in Markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 254–263.

Wang, S.; Huang, L.; and Lui, J. 2020. Restless-UCB, an Efficient and Low-complexity Algorithm for Online Restless Bandits. In *Proc. of NeurIPS*.

Wang, S.; Xiong, G.; and Li, J. 2023. Online Restless Multi-Armed Bandits with Long-Term Fairness Constraints. *arXiv preprint arXiv:2312.10303*.

Whittle, P. 1988. Restless Bandits: Activity Allocation in A Changing World. *Journal of Applied Probability*, 287–298.

Xiong, G.; and Li, J. 2023. Finite-Time Analysis of Whittle Index based Q-Learning for Restless Multi-Armed Bandits with Neural Network Function Approximation. In *Proc. of NeurIPS*.

Xiong, G.; Li, J.; and Singh, R. 2022. Reinforcement Learning Augmented Asymptotically Optimal Index Policy for Finite-Horizon Restless Bandits. In *Proc. of AAAI*.

Xiong, G.; Wang, S.; and Li, J. 2022. Learning Infinite-Horizon Average-Reward Restless Multi-Action Bandits via Index Awareness. In *Proc. of NeurIPS*.

Xiong, G.; Wang, S.; Yan, G.; and Li, J. 2022. Reinforcement Learning for Dynamic Dimensioning of Cloud Caches: A Restless Bandit Approach. In *Proc. of IEEE INFOCOM*.

Yin, T.; Raab, R.; Liu, M.; and Liu, Y. 2023. Long-Term Fairness with Unknown Dynamics. *arXiv preprint arXiv:2304.09362*.