Reinforcement Learning Augmented Asymptotically Optimal Index Policy for Finite-Horizon Restless Bandits

Guojun Xiong[1], Jian Li[1] and Rahul Singh[2]
[1]SUNY-Binghamton University
[2]Indian Institute of Science
gxiong1@binghamton.edu, lij@binghamton.edu, rahulsingh@iisc.ac.in

Abstract

We study a finite-horizon restless multi-armed bandit problem with multiple actions, dubbed as $\text{R(MA)}^2\text{B}$. The state of each arm evolves according to a controlled Markov decision process (MDP), and the reward of pulling an arm depends on both the current state and action of the corresponding MDP. Since finding the optimal policy is typically intractable, we propose a computationally appealing index policy entitled *Occupancy-Measured-Reward Index Policy* for the finite-horizon $\text{R(MA)}^2\text{B}$. Our index policy is well-defined without the requirement of indexability condition and is provably asymptotically optimal. We then adopt a learning perspective where the system parameters are unknown, and propose $\text{R(MA)}^2\text{B}$-UCB, a generative model based reinforcement learning augmented algorithm that can fully exploit the structure of *Occupancy-Measured-Reward Index Policy*. Compared to existing algorithms, $\text{R(MA)}^2\text{B}$-UCB performs close to offline optimum, as well as achieves a sub-linear regret and a low computational complexity all at once. Experimental results show that $\text{R(MA)}^2\text{B}$-UCB outperforms existing algorithms in both regret and running time.

Introduction

We study the restless multi-armed, multi-action bandit problem, dubbed as $\text{R(MA)}^2\text{B}$ with a finite horizon. A restless multi-armed bandit (RMAB) problem (Whittle 1988) involves activating a fixed number of competing "arms" sequentially over time. Each arm is endowed with a state that evolves independently according to a controlled Markov decision process (MDP). In its original form (i.e., MAB), only the states of activated arms evolve, and rewards are generated upon the state evolution. The goal is to decide which arms need be activated at each decision epoch to maximize the total expected reward. A celebrated *Gittins index policy* (Gittins 1974) was proposed for MAB, where each arm is assigned with an index as a function of its current state and then activates the arm(s) with the largest indices. However, this policy only optimizes the *infinite-horizon* expected reward when *only one arm* can be activated at each decision epoch. Whittle (Whittle 1988) generalized the MAB to also allow the evolution of non-activated arms (dubbed as "a changing world setting"), giving rise to the RMAB problem.

The RMAB problem is a general model for a variety of sequential decision making problems ranging from job allocation (Niño-Mora 2007; Jacko 2010; Bertsimas and Niño-Mora 2000), wireless communication (Dai et al. 2011; Sheng, Liu, and Saigal 2014), sensor management (Mahajan and Teneketzis 2008; Ahmad et al. 2009) and healthcare (Deo et al. 2013; Lee, Lavieri, and Volk 2019; Mate, Perrault, and Tambe 2021; Killian, Perrault, and Tambe 2021). However, the RMAB is notoriously *intractable* (Papadimitriou and Tsitsiklis 1994) and the optimal policy for an RMAB is rarely an index policy. To that end, Whittle proposed a heuristic called the *Whittle index* for the *infinite-horizon* RMAB. However, the Whittle index is well-defined only when the so-called *indexability* condition is satisfied. Furthermore, even when an arm is indexable, finding its Whittle index can still be intractable, especially when the corresponding controlled Markov process is convoluted (Niño-Mora 2007). Finally, Whittle index policy is *only guaranteed* to be asymptotically optimal (Weber and Weiss 1990) under a difficult-to-verify condition that the fluid approximation has a globally asymptotically stable attractor.

Inspired by Whittle, many studies focus on finding the index policy for restless bandit problems, e.g., (Nino-Mora 2001; Verloop 2016; Hu and Frazier 2017; Zayas-Cabán, Jasin, and Wang 2019; Brown and Smith 2020). This line of works is primarily under the assumption that the system parameters are all already known. Since the true parameters are typically unavailable and possibly time-varying in many cases, it becomes important to examine RMAB from a learning perspective, e.g., (Dai et al. 2011; Liu, Liu, and Zhao 2011; Tekin and Liu 2011; Liu, Liu, and Zhao 2012; Tekin and Liu 2012; Ortner et al. 2012; Jung and Tewari 2019; Jung, Abeille, and Tewari 2019; Wang, Huang, and Lui 2020; Xiong, Singh, and Li 2021). However, analyzing a learning algorithm in RMAB is in general hard due to the learner's additional uncertainty, and there are still many open challenges for RMAB from both the policy design perspective and the learning perspective.

First, existing confidence bound based algorithms (Liu, Liu, and Zhao 2011; Tekin and Liu 2011, 2012; Liu, Liu, and Zhao 2012) may not perform close to the offline optimum. This is because *the baseline policy* is often heuristic without performance guarantee, e.g., only pulling one arm or a fixed set of arms. This is known to be weak in the RMAB setting, which makes the regret $\mathcal{O}(\log T)$ less meaningful. Second, the aforementioned learning algorithms with a theoretical guarantee of an $\tilde{\mathcal{O}}(\sqrt{T})$ regret are often *computationally expensive*. For example, colored-UCRL2 (Ortner et al. 2012) suffers from an exponential computation complexity, and the regret bound is exponential in the number of states and arms. This is because it needs to solve a set of Bellman equations with an exponentially large space set. Third, existing low-complexity policies such as (Wang, Huang, and Lui 2020; Xiong, Singh, and Li 2021) are often achieved with *no guarantee of an $\tilde{\mathcal{O}}(\sqrt{T})$ regret*, and also restricted to a specific Markov model, which is hard to be generalized. In a different line of works, the Thompson sampling based algorithms (Jung and Tewari 2019; Jung, Abeille, and Tewari 2019) that provide a theoretical guarantee in the Bayesian setting often suffers from a computationally expensive update method when the likelihood functions are complex. To the best of our knowledge, there are no provably optimal policies for RMAB problems (let alone the R(MA)^{2}B in consideration) with an efficient learning algorithm that performs *close to the offline optimum* and achieves *a sub-linear regret* and *a low computation complexity* all at once.

In this paper, we address above challenges for R(MA)^{2}B problems with finite horizon. In contrast to most of the existing literature aforementioned, each arm can take multiple actions. The rationality is that many applications are not limited to binary actions as in RMAB. For example, videos can be delivered through wireless channels at different levels of power (i.e., actions), which leads to different quality of experiences to users. However, the analysis of restless bandits with multiple actions largely remains elusive for general settings in the literature. We make progress toward R(MA)^{2}B problems by devising and analyzing two correlated algorithms. Our main contributions are as follows:

• **Asymptotically optimal index policy.** We propose an index policy for the general finite-horizon R(MA)^{2}B problem, entitled *Occupancy-Measured-Reward Index Policy*, in that the system parameters are known. We show that it is asymptotically optimal in the same limit considered by Whittle. Unlike Whittle index based policies, our index policy does not require the indexability condition to hold, and is well-defined for both indexable and nonindexable R(MA)^{2}B problems. This property is significantly appealing since the indexability condition is hard to verify or may not hold true in general, and the non-indexable settings have so far received little attention but arise in many practical problems.

• **Reinforcement learning augmented index policy.** We present one of the first generative model based reinforcement learning augmented algorithm toward an index policy in the context of finite-horizon R(MA)^{2}B problems, and we call it R(MA)^{2}B-UCB. Different from the state-of-the-art colored-UCRL2 that has a complexity exponential in the number of arms, our R(MA)^{2}B-UCB contains a novel optimistic planning step by obtaining an estimated model via sampling state-action pairs in an offline manner and solving a so-called extended linear programming problem in occupancy measures, with which the complexity is only linear in the number of arms. Furthermore, R(MA)^{2}B-UCB achieves an $\tilde{\mathcal{O}}(\sqrt{T})$ regret and performs close to the offline optimum since it contains a novel exploitation step by fully leveraging our *Occupancy-Measured-Reward Index Policy*, which significantly outperforms existing methods that often rely on a heuristic policy. Moreover, the multiplicative "pre-factor" that goes with the time-horizon dependent function in the regret is quite low due to the novel exploitation step, which is exponentially better than that of the colored-UCRL2. Our simulation results also show that R(MA)^{2}B-UCB outperforms existing algorithms in both regret and running time.

Notation. We denote the set of natural and real numbers by $\mathbb{N}$ and $\mathbb{R}$, respectively. We let T be the finite number of total decision epoch (time). We denote the cardinality of a finite set $\mathcal{A}$ by $A := |\mathcal{A}|$. We also use $[N]$ to represent the set of integers $\{1, \cdots, N\}$ for $N \in \mathbb{N}$.

System Model

Consider a finite-horizon R(MA)^{2}B problem with N arms. Each arm n is associated with a specific unichain Markov decision process (MDP) (Kallenberg 2003) $(\mathcal{S}_n, \mathcal{A}_n, P_n, r_n, \mathbf{s}_1, T)$, where $\mathcal{S}_n$ is the finite state space, $\mathcal{A}_n$ denotes the set of finite actions, $P_n : \mathcal{S}_n \times \mathcal{A}_n \times \mathcal{S}_n \mapsto \mathbb{R}$ is the transition kernel and $r_n : \mathcal{S}_n \times \mathcal{A}_n \mapsto \mathbb{R}$ is the reward function. For the ease of readability, we assume that all arms share the same state and action spaces, and denote as $\mathcal{S}$ and $\mathcal{A}$, respectively. Our results and analysis will still apply to different state and action spaces at the cost of complicated notations. In particular, we denote the action set $\mathcal{A} = \{0, 1, \cdots, A\}$ with $A < \infty$. Using the standard terminology from the RMAB literature, we call an arm *passive* when action $a = 0$ is applied to it, and *active* otherwise. An *activation cost* is incurred each time action a is applied to arm n. For the abuse of notation, we denote the activation cost to be a units by taking action a. The total activation cost associated with active arms at each time t is constrained by K units, which we call the *activation budget*. The initial state is chosen according to the initial distribution $\mathbf{s}_1$ and $T < \infty$ is the horizon.

At time $t \in [T]$, each arm n is at a specific state $s_n(t) \in \mathcal{S}$ and evolves to $s_n(t+1)$ independently as a controlled Markov process with the controlled transition probabilities $P_n(s_n(t), a_n(t), s_n(t+1))$ when action $a_n(t)$ is taken. The immediate reward earned from activating arm n at time t is denoted by $r_n(t) := r_n(s_n(t), a_n(t))$. Without loss of generality, we assume that $r_n \in [0, 1]$, $\forall n$ with mean $\bar{r}_n(s, a)$, and let $r_n(s, 0)$ be 0 $\forall s \in \mathcal{S}$, i.e., no reward is earned when the arm is passive. Denote the total reward earned at time t by $R(t)$, i.e., $R(t) := \sum_n r_n(t)$. Let $\mathcal{F}_t$ denote the operational history until t, i.e., the sigma-algebra (Shiryaev 2007) generated by the random variables $\{s_n(\ell) : n \in [N], \ell \in [t]\}, \{a_n(\ell) : n \in [N], \ell \in [t-1]\}$. Our goal is to derive a policy $\pi : \mathcal{F}_t \mapsto \mathcal{A}^N$ that makes decisions regarding which set of arms needs to be made active at each time $t \in [T]$ so as to maximize the expected value of the cumulative rewards subject to the activation budget, i.e.,

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

where the subscript indicates that the expectation is taken with respect to the measure induced by the policy π. We refer to the problem (1) as the "original problem", which suffers from the "curse of dimensionality" (Bellman 2010; Bertsekas 1995), and hence is computationally intractable. We overcome this difficulty by developing a computationally feasible and provably optimal index-based policy.

Asymptotically Optimal Index Policy

In this section, we focus on the scenario that the transition probabilities and reward functions are known. We will propose a powerful framework to design an asymptotically optimal index policy for finite-horizon R(MA)^{2}Bs. We begin by introducing the "relaxed problem", which can be posed as a linear programming (LP) in occupancy measures (Altman 1999). This forms a building block of our proposed lightweight index-based policy, which we show is asymptotically optimal with respect to the original problem.

The Relaxed Problem

We consider the following relaxed problem by relaxing the "hard" constraint in (1) to the "relaxed" constraint, i.e., the activation cost at time $t \in [T]$ is limited by K on average

$$\max_{\pi} \mathbb{E}_\pi \left(\sum_{n=1}^{N} \sum_{t=1}^{T} r_n(t) \right) \text{ s.t. } \mathbb{E}_\pi \left\{ \sum_{n=1}^{N} a_n(t) \right\} \leq K, \forall t. \quad (2)$$

It is clear that the relaxed problem (2) achieves an upper bound of the optimal value of (1). Note that the optimal policy to (2) may be randomized (Altman 1999), i.e., an optimal deterministic policy may not exist for a finite-horizon MDP. Furthermore, we cannot use the conventional backward induction to find the optimal policy since the Bellman optimality equations no longer hold given the constraint in (2). It is well known (Altman 1999) that the relaxed problem (2) can be reduced to a LP in which the decision variables are the occupancy measures of the controlled process. More specifically, the occupancy measure μ of a policy π in a finite-horizon MDP is defined as the expected number of visits to a state-action pair (s, a) at each time t. Formally,

$$\mu = \left\{ \mu_n(s, a; t) = \mathbb{P}(s_n(t) = s, a_n(t) = a) : \forall n, t \right\}.$$

Using this definition, the relaxed problem (2) can be reformulated as the following LP (Altman 1999):

$$\max_{\mu} \sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{(s,a)} \mu_n(s, a; t) \bar{r}_n(s, a) \quad (3)$$

$$\text{s.t.} \sum_{n=1}^{N} \sum_{(s,a)} a \mu_n(s, a; t) \leq K, \quad (4)$$

$$\sum_{a} \mu_n(s, a; t) = \sum_{(s',a')} \mu_n(s', a'; t-1) P_n(s', a', s), \quad (5)$$

$$\sum_{a} \mu_n(s, a; 1) = \mathbf{s}_1(s), \quad (6)$$

where (4) is a restatement of the constraint in (2) for $\forall t \in [T]$, which indicates the activation budget; (5) represents the transition of the occupancy measure from time $t-1$ to time $t, \forall n \in [N]$ and $\forall t \in [T]$; and (6) indicates the initial condition for occupancy measure at time $1, \forall s \in \mathcal{S}$. From the constraints (5)-(6), it can be easily checked that the occupancy measure satisfies $\sum_{s,a} \mu_n(s, a, t) = 1, \forall t \in [T]$. Thus, the occupancy measure $\mu_n, \forall n \in [N]$ is a probability measure.

The optimal solutions of the LP define an optimal Markov policy of *the relaxed problem* through the occupancy measure (Altman 1999). Specifically, denote the solution to the above LP as $\mu^\star = \{\mu_n^\star(s, a; t) : n \in [N], t \in [T]\}$. Then a Markovian non-stationary randomized policy $\chi^\star = \{\chi_n^\star(t) : n \in [N], t \in [T]\}$ can be constructed as follows: if the state $s_n(t)$ is s at time t, then $\chi_n^\star(t)$ chooses an action a with a probability equaling to

$$\chi_n^\star(s, a; t) := \frac{\mu_n^\star(s, a; t)}{\sum_{a' \in \mathcal{A}} \mu_n^\star(s, a'; t)}. \quad (7)$$

If the denominator of (7) equals zero, i.e., state s for arm n is not reachable at time t, arm n can be simply made passive, i.e., $\chi_n^\star(s, 0; t) = 1$ and $\chi_n^\star(s, a; t) = 0, \forall a \in \mathcal{A} \setminus \{0\}$.

The Occupancy-Measured-Reward Index Policy

The solutions to the above LP and the constructed Markov policy $\chi^\star$ form the building block of our index policy for the original problem (1). Note that the optimal policy (7) is not always feasible for *the original problem* since in the latter at most K units of activation costs can be consumed at a time. To this end, our index policy assigns an index to each arm based on its current state and current time. We denote the index $\psi_n(s_n(t); t)$ associated with arm n at time t as

$$\psi_n(s_n(t); t) := \sum_{a \in \mathcal{A} \setminus \{0\}} \chi_n^\star(s_n(t), a; t) \bar{r}_n(s_n(t), a), \quad (8)$$

where $\chi_n^\star(s_n(t), a; t)$ is defined in (7). We call this the *occupancy-measured-reward index (OMR index)* since it is merely based on the optimal occupancy measures solved from the LP in (3)-(6) and the mean reward, representing the expected obtained reward for arm n at state $s_n(t)$ of time t. Let $\psi(t) := \{\psi_n(s_n(t); t) : n \in [N]\}$ be the OMR indices associated with the N arms at time t. Denote the action for arm n at state $s_n(t)$ of time t as $a_n^\star(s_n(t); t)$ and the set of active arms at time t as $\mathcal{B}(t)$. Our index policy then activates arms with OMR indices in a decreasing order. The activation process is terminated until the constraint $\sum_{n \in \mathcal{B}(t)} a_n^\star(s_n(t); t) \leq K$ is violated. The remaining arms $[N] \setminus \mathcal{B}(t)$ are passive at time t. Specifically, for each activated arm, its action is randomly selected according to the probability $\chi_n^\star(s_n(t), a; t)$ in (7). When multiple arms sharing the same OMR indices, we randomly activate one arm and allocate the remaining activation costs across all possible actions according to the probability $\chi_n^\star(s_n(t), a; t)$. If all indices are zero, then all remaining arms are made passive. We call this *an Occupancy-Measured-Reward Index Policy (OMR Index Policy)*, and denote it as $\pi^\star = \{\pi_n^\star, n \in [N]\}$, which is summarized in the supplementary material due to space constraints.

Remark 1 *Our index policy is computationally appealing since it is only based on the "relaxed problem" by solving a LP. Furthermore, if all arms share the same MDP, the LP can be decomposed across arms as in (Whittle 1988), and hence the computational complexity does not scale with the number of arms. More importantly, our index policy is well-defined without the requirement of indexability condition (Whittle 1988). This is in contrast to most of the existing Whittle index-based policies that are only well defined in the case that the system is indexable, which is hard to verify and may not hold in general. Closest to our work is the parallel work on restless bandits (Zhang and Frazier 2021) with known transition probabilities and reward functions. In particular, (Zhang and Frazier 2021) explores index policies similar to ours, but under the assumption of homogeneous MDPs across arms in the binary action settings, and mainly focus on characterizing the asymptotic optimality gap. Our index policy in this section can be seen as the complement to it with the general heterogeneous MDPs across arms in the multiple action settings. Our design of reinforcement learning augmented index policy and the regret analysis in next section also distinguishes our work.*

Asymptotic Optimality

For the abuse of notation, we let the number of arms be ρN and the value of activation constraint be ρK in the limit with $\rho \to \infty$. In other words, it represents the scenarios where there are N different classes of arms and each class contains ρ arms. Our *OMR Index Policy* achieves asymptotic optimality when the number of arms ρN and the activation constraint ρK go to infinity while holding $\alpha = K/N$ constant[1]. Let $R(\pi, \rho K, \rho N)$ denote the expected reward of the original problem (1) obtained by an arbitrary policy π in this limit. Denote the optimal policy of the original problem (1) as π^{opt}.

Theorem 1 *The OMR Index Policy achieves the asymptotic optimality as follows*

$$\lim_{\rho\to\infty} \frac{1}{\rho}\Big(R(\pi^\star, \rho K, \rho N) - R(\pi^{opt}, \rho K, \rho N)\Big) = 0.$$

Remark 2 *Theorem 1 indicates that as the number of per-class arms (i.e., ρ) goes to infinity, the gap between the performance achieved by OMR Index Policy and the optimal policy π^{opt} is bounded, and thus per arm gap tends to be zero. The proof is available at (Xiong, Li, and Singh 2021).*

Reinforcement Learning for the Index Policy

The computation of the *OMR Index Policy* requires the knowledge about the transition probabilities and reward functions associated with the MDPs for each arm. However, these quantities are typically unavailable in practice. Hence, we now adopt a learning perspective where the parameters are unknown. We propose a generative model based reinforcement learning (RL) augmented algorithm entitled R(MA)^{2}B-UCB, which obtains samples initially and can ensure almost the same performance as *OMR Index Policy*. R(MA)^{2}B-UCB is also computationally efficient since it can fully exploit the structure of the *OMR Index Policy*.

Algorithm 1: R(MA)^{2}B-UCB Policy

Input: Learning horizon T, and learning counts $\Lambda(T) < T$.

```
1: for $n = 1, 2, ..., N$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$ do
2:     Sample pairs $(s, a)$ of arm $n$ for $\Lambda(T)$ times.
3: end for
4: Construct $\mathcal{P}_n(s, a)$ and $\mathcal{R}_n(s, a)$ according to (9);
5: Compute the optimal solution of the extended LP (10);
6: Establish the corresponding OMR Index Policy $\pi^\star$;
7: Execute $\pi^\star$ for the rest of the game.
```

The Learning Problem

We consider the setting aforementioned, where the learning agent repeatedly interacts with a finite-horizon R(MA)^{2}B in which each arm is associated with a controlled MDP $(\mathcal{S}, \mathcal{A}, P_n, r_n, \mathbf{s}_1, T)$. However, the learning agent does not know the transition probability P_n nor the reward function r_n, $\forall n \in [N]$, and it relies on the samples observed to make decisions. The performance of the learning agent is measured by the regret. More precisely, the regret is defined as the expected gap between the offline optimum, i.e., the best policy under which both the transition probabilities and reward functions are known, and the cumulative reward of the arm selecting algorithm. Specifically, denote the cumulative reward under an arbitrary policy π as $R(\pi, \mathbf{s}_1, T) := \sum_{t=1}^{T} r(t)$, which is a random variable. Then the expected average reward under policy π satisfies $\xi(\pi, \mathbf{s}_1) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}[R(\pi, \mathbf{s}_1, T)]$, and the optimal average reward is $\xi^{opt} := \sup_\pi \xi(\pi, \mathbf{s}_1)$, which is independent of the initial state for MDPs with finite diameter (Puterman 1994). Then the regret of policy π is defined as $\Delta(\pi, \mathbf{s}_1, T) := T\xi^{opt} - \mathbb{E}_\pi[R(\pi, \mathbf{s}_1, T)]$.

A Generative Model Based Learning Algorithm

We consider an adaptation of the upper confidence bound (UCB) (Auer, Cesa-Bianchi, and Fischer 2002) to the setting of R(MA)^{2}B problem with the generative model (Kearns, Mansour, and Ng 2002), which we call the R(MA)^{2}B-UCB policy, as presented in Algorithm 1.

More precisely, there are two phases in R(MA)^{2}B-UCB: a planning phase and a policy execution phase. The planning phase (lines 1-6 in Algorithm 1) leverages a novel construction of a set of plausible transition models (i.e., MDPs) based on the number of visits to state-action pairs (s, a) and transitions tuples (s, a, s') as accurate as possible. Specifically, we explore a generative approach with a single step simulator that can generate samples of the next state and reward given any state and action (Kearns, Mansour, and Ng 2002; HasanzadeZonuzy, Kalathil, and Shakkottai 2021). By solving an optimistic planning problem, which is expressed as an LP problem in occupancy measures, we can define the corre-

[1] We consider the asymptotic optimality in the same limit as by Whittle (Whittle 1988) and others (Weber and Weiss 1990; Verloop 2016; Hu and Frazier 2017; Zayas-Cabán, Jasin, and Wang 2019).

sponding *OMR Index Policy*. The planning problem, referred to as *an extended LP* in Algorithm 1 is detailed below. Our key contribution here is to choose the right value of $\Lambda(T)$ to balance the accuracy and complexity, which contributes to the properties of sub-linear regret and low-complexity of R(MA)^{2}B-UCB.

At the policy execution phase (line 7 in Algorithm 1), the derived *OMR Index Policy* is executed. Our key contribution here is to leverage our proposed *OMR Index Policy*, rather than using heuristic ones as in existing algorithms. This guarantees that R(MA)^{2}B-UCB performs close to the offline optimum since our proposed index policy is near-optimal. Moreover, this contributes to the low multiplicative "pre-factor" that goes with the time-horizon dependent function in the regret, which is exponentially better than that of the state-of-the-art colored-UCRL2 (Ortner et al. 2012).

Optimistic planning. We sample each state-action pair of arm n for $\Lambda(T)$ (the value of $\Lambda(T)$ will be specified later) number of times uniformly across all state-action pairs. We denote the number of times that a transition tuple (s,a,s') was observed within $\Lambda(T)$ as $T_n(s,a,s')$, satisfying $T_n(s,a,s') = \sum_{h=1}^{\Lambda(T)} \mathbf{1}(s_n(h+1) = s'|s_n(h) = s, a_n(h) = a), \forall (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, where $s_n(h)$ represents the state for arm n at time h and $a_n(h)$ is the corresponding action. Then R(MA)^{2}B-UCB estimates the true transition probability $\forall(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and the true reward $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$ by the corresponding empirical averages as $\hat{P}_n(s'|s,a) = T_n(s,a,s')/\Lambda(T)$, $\hat{r}_n(s,a) = \frac{1}{\Lambda(T)}\sum_{h=1}^{\Lambda(T)} r_n(s,a;h)\mathbf{1}(s_n(h)=s, a_n(h)=a)$.

R(MA)^{2}B-UCB further defines confidence intervals for the transition probabilities (resp. the rewards), such that the true transition probabilities (resp. true rewards) lie in them with high probability. Formally, for $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$, we define

$$\begin{aligned}\mathcal{P}_n(s,a) :=& \{\tilde{P}_n(s'|s,a), \forall s'\in\mathcal{S}:\\ & \quad |\tilde{P}_n(s'|s,a) - \hat{P}_n(s'|s,a)| \leq \delta_n(s,a)\},\\ \mathcal{R}_n(s,a) :=& \{\tilde{r}_n(s,a) : \tilde{r}_n(s,a) = \hat{r}_n(s,a) + \delta_n(s,a)\}, \quad (9)\end{aligned}$$

where the size of the confidence intervals $\delta_n(s,a)$ is built using the empirical Hoeffding inequality (Maurer and Pontil 2009). For any $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, and $\eta\in(0,1)$, it is defined as $\delta_n(s,a) = \sqrt{\frac{1}{2\Lambda(T)}\log\left(SAN\Lambda(T)/\eta\right)}$.

The set of plausible MDPs associated with the confidence intervals is $\mathcal{M} = \{M_n = (\mathcal{S},\mathcal{A},\tilde{P}_n,\tilde{r}_n) : \tilde{P}_n(\cdot|s,a)\in\mathcal{P}_n(s,a), \tilde{r}_n(s,a)\in\mathcal{R}_n(s,a), \forall n\}$. Then R(MA)^{2}B-UCB computes a policy by performing optimistic planning. Given the set of plausible MDPs, it selects an optimistic transition (resp. reward) function and an optimistic policy by solving a "modified LP", which is similar to the LP defined in (3)-(6), but with the transition and reward functions replaced by $\tilde{P}(\cdot|\cdot,\cdot)$ and $\tilde{r}(\cdot,\cdot)$ in the confidence balls (9) since the corresponding true values are not available.

The extended LP problem. The modified LP can be further expressed as an extended LP by leveraging the state-action-state occupancy measure $z_n(s,a,s',t)$ defined as $z_n(s,a,s',t) = P_n(s'|s,a)\mu_n(s,a;t)$ to express the confidence intervals of the transition probabilities. The extended LP over z is as follows:

$$\begin{aligned}\max & \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{(s,a,s')} z(s,a,s';t)\tilde{r}_n(s,a)\\ \text{s.t.} & \sum_{n=1}^{N}\sum_{(s,a,s')} z_n(s,a,s';t)a \leq K, \ \forall t,\\ & \sum_{a,s'} z_n(s,a,s';t) = \sum_{s',a'} z_n(s',a',s,t-1), \ \forall t,\\ & \sum_{a,s'} z_n(s,a,s';1) = \mathbf{s}_1(s), \ \forall s,\\ & \frac{z_n(s,a,s';t)}{\sum_y z_n(s,a,y;t)} - (\hat{P}_n(s'|s,a) + \delta_n(s,a)) \leq 0,\\ & -\frac{z_n(s,a,s';t)}{\sum_y z_n(s,a,y;t)} + (\hat{P}_n(s'|s,a) - \delta_n(s,a)) \leq 0, \quad (10)\end{aligned}$$

where the last two constraints indicate that the transition probabilities lie in the desired confidence interval for $\forall(s,a,s',t)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[T]$. Such an approach was also used in (Jin et al. 2019; Rosenberg and Mansour 2019) in the context of adversarial MDPs and (Efroni, Mannor, and Pirotta 2020; Kalagarla, Jain, and Nuzzo 2021; HasanzadeZonuzy, Kalathil, and Shakkottai 2021) in constrained MDPs. Once we compute z, the policy is recovered from the computed occupancy measures as

$$\chi_n(s,a;t) = \frac{\sum_{s'} z_n(s,a,s';t)}{\sum_{b,s'} z_n(s,b,s';t)}. \quad (11)$$

Finally, we compute the *OMR index* as in (8) using (11), from which we construct the *OMR Index Policy*, and execute this policy to the end.

Remark 3 *Although R(MA)^{2}B-UCB has a similar form as an "explore-then-commit" policy, e.g., (Ortner et al. 2012), one key novelty of R(MA)^{2}B-UCB lies in leveraging the approach of optimism-in-the-face-of-uncertainty (Jaksch, Ortner, and Auer 2010) to balance exploration and exploitation in a non-episodic offline manner. As a result, there is no need for R(MA)^{2}B-UCB to search for a better MDP instance as in (Ortner et al. 2012; Wang, Huang, and Lui 2020), which is computationally expensive (i.e., exponential in the number of arms). The second key novelty is that R(MA)^{2}B-UCB only relies on samples initially obtained by a generative model to construct a upper-confidence ball, from which a policy can be derived by solving an extend LP for only once with a complexity of $\mathcal{O}(NSAT)$ (which is $\mathcal{O}(SAT)$ if all arms are identical). However, existing algorithms, e.g., colored UCRL2 is computationally expensive as it relies on a complex recursive Bellman equation to derive the policy. The last key novelty is that R(MA)^{2}B-UCB further leverages the structure of our proposed near-optimal index policy in the policy execution phase rather than using a heuristic one as in existing algorithms e.g., (Liu, Liu, and Zhao 2011; Tekin and Liu 2011, 2012; Liu, Liu, and Zhao 2012). These key novelties jointly guarantee that R(MA)^{2}B-UCB achieves almost the same performance as the offline optimum, a sub-linear regret and a low computation complexity all at once.*

Regret Bound

We present our main theoretical results in this section.

Theorem 2 *The regret of the R(MA)^{2}B-UCB policy with $\Lambda(T) = \mathcal{O}(T^{1/2})$ satisfies:*

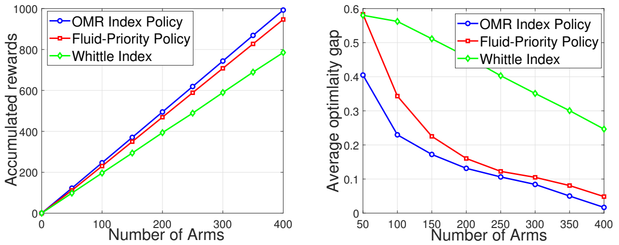
$$\Delta(\pi^\star, \mathbf{s}_1, T) = \mathcal{O}\Big((SAK + 2K(1+\eta))\sqrt{T}\Big). \qquad (12)$$

Since there are two phases in R(MA)^{2}B-UCB, we decompose the regret as $\Delta(\pi^\star, \mathbf{s}_1, T) = \Delta(T_1) + \Delta(\pi^\star, \mathbf{s}_1, T_2)$, where $\Delta(T_1)$ is the regret for the planning phase and $\Delta(\pi^\star, \mathbf{s}_1, T_2)$ is the regret for the policy execution phase with $T_2 = T - T_1$. The first term $\mathcal{O}(SAK\sqrt{T})$ in (12) is the worst regret from $\Lambda(T)$ explorations of each state-action pair under the generative model with $\mathcal{O}(SA\sqrt{T})$ time steps for sampling and at most K arms being activated each time. The second term $\mathcal{O}(2K(1+\eta)\sqrt{T})$ comes from the policy execution phase. Specifically, the $\mathcal{O}(2K\eta\sqrt{T})$ regret occurs when $\Lambda(T)$ explorations for each state-action pair construct a set of plausible MDPs that do not contain the true MDP $\mathcal{M}$ in line 4 of Algorithm 1, which is a rare event with probability $2\eta/\Lambda(T)$. The key then is to characterize the regret when the event that the true MDP $\{(\mathcal{S}, \mathcal{A}, P_n, r_n), \forall n\}$ lies in the set of plausible MDP $\mathcal{M}$ occurs. Based on the optimism of plausible MDPs, the optimal average reward $\tilde{\xi}$ for the optimistic MDP $\{(\mathcal{S}, \mathcal{A}, \tilde{P}_n, \tilde{r}_n), \forall n\}$ is no less than ξ^{opt}. Thus the expected regret is bounded by $T_2\tilde{\xi} - T_2\xi^{opt}$, which is directly related with the occupancy measure we defined. The proof is available at (Xiong, Li, and Singh 2021).

Remark 4 *Though R(MA)^{2}B-UCB is an offline non-episodic algorithm, it still achieves an $\tilde{\mathcal{O}}(\sqrt{T})$ regret no worse than the episodic colored-UCRL2. Note that for colored-UCRL2, the regret bound is instance-dependent due to the online episodic manner such that the regret bound tends to be logarithmic in the horizon as well. However, R(MA)^{2}B-UCB adopts explore-then-commit mechanism which uses generative model based sampling and constructs the plausible MDPs sets only once. This removes the instance-dependent regret with order of $\log T$. Though the state-of-the-art Restless-UCB (Wang, Huang, and Lui 2020) has a similar mechanism as ours in obtaining samples in an offline manner, it lowers its implementation complexity by sacrificing the regret performance to $\mathcal{O}(T^{2/3})$ since it heavily depends on the performance of an offline oracle approximator for policy execution. Instead, we leverage our proposed provably optimal and computationally appealing index policy for the policy execution phase. This also contributes to the low multiplicative "pre-factor" in the regret.*

Experiments

In this section, we present our experimental results to validate our model and theoretical results, including the asymptotic optimality of the *OMR Index Policy*, and the sublinear regret of the R(MA)^{2}B-UCB policy. Due to space constraints, we relegate some experimental results including case studies to (Xiong, Li, and Singh 2021).



(a) Accumulated reward.

(b) Average optimality gap.

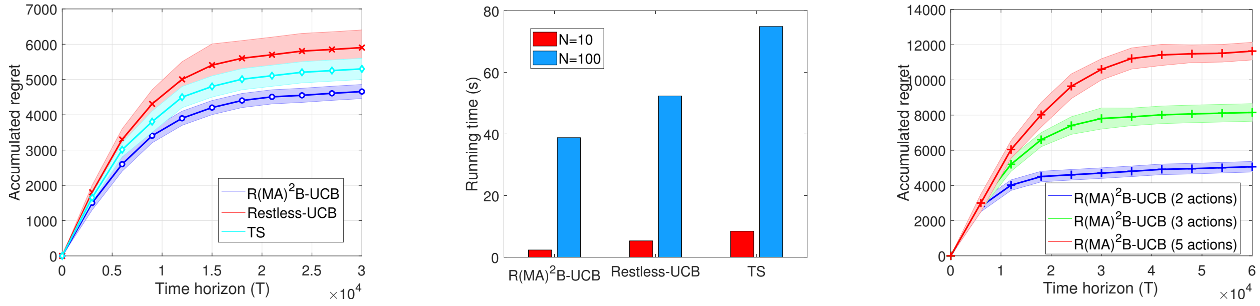
Figure 1: Evaluation of the *OMR Index Policy*.

Evaluation of the *OMR Index Policy*

Since most existing index policies are designed only for binary action settings, i.e., arms being active or passive, and hard to be generalized to the multi-action setting studied in this paper, we first consider that the states evolve as a specific birth-and-death process where state s can only transit to $s-1$ or $s+1$. We compare with two state of the arts, i.e., Whittle index policy (Whittle 1988), and Fluid-priority policy (Zhang and Frazier 2021), a priority based policy as defined in (Verloop 2016). Consider a setting with 10 classes of arms, and a state space $\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The arrival rates are set as $\{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$ with a departure rate 20. When a class-i arm is activated, it receives a random reward $r_i(s)$ which is a Bernoulli random variable with a state dependent rate $s \cdot p_i$, i.e., $r_i(s) \sim Ber(sp_i)$ with p_i uniformly distributed in $[0.01, 0.1]$. Otherwise, it receives a zero reward. The horizon is $T = 100$ and the activation ratio is $\alpha = K/N = 0.3$. For ease of exposition, the number of arms vary from 50 to 400.

The accumulated rewards achieved by these policies are presented in Figure 1a. We observe that *OMR Index Policy* performs slightly better than the Fluid-priority policy. We conjecture that this is due to the fact that *OMR Index Policy* prioritizes the arms directly based on their contributions to the cumulative reward, while Fluid-priority policy does not differentiate arms in the same priority category. More importantly, both *OMR Index Policy* and Fluid-priority policy significantly outperform the Whittle index policy.

We further validate the asymptotic optimality of *OMR Index Policy* (see Theorem 1). In particular, we compare the rewards obtained by *OMR Index Policy* and the two baselines, with that obtained from the theoretical upper bound achieved by solving the LP in (3)-(6). The difference is called the optimality gap. The average optimality gap, i.e., the ratio between the optimality gap and the number of arms of different policies is illustrated in Figure 1b. Again, we observe that *OMR Index Policy* slightly outperforms the Fluid-priority in terms of the vanishing speed of the average optimality gap since *OMR Index Policy* achieves a higher accumulated reward as shown in Figure 1a. Moreover, both *OMR Index Policy* and Fluid-priority significantly outperform the Whittle index policy. This is due to the fact that the optimality gap of the Fluid-priority index policy (i.e. a constant $\mathcal{O}(1)$) does not scale with the number of arms, while that of Whittle index policy does (Zhang and Frazier 2021).



(a) Accumulated regret: binary action setting.
(b) Average running time.
(c) Accumulated regret: multi-action setting.

Figure 2: Evaluation of the R(MA)^{2}B-UCB Policy.

Evaluation of the R(MA)^{2}B-UCB Policy

We compare with two state-of-the-art algorithms including Restless-UCB (Wang, Huang, and Lui 2020) and a Thompson sampling (TS) based policy (Jung and Tewari 2019) for restless bandits. Note that Restless-UCB is also an offline learning policy similar to ours while the TS-based policy is online with a sub-linear regret in the Bayesian setting but suffers from a high computation complexity. Also need to mention that there exists another popular algorithm for RMAB problems named colored-UCRL2 (Ortner et al. 2012). However, it is well known that the computation complexity of colored-UCRL2 grows exponentially with the number of arms. Furthermore, it has been shown that Restless-UCB outperforms colored-UCRL2 in (Wang, Huang, and Lui 2020), hence we omit the comparison here.

We leverage the same settings as described above for index policy evaluation and consider the number of arms to be $N = 100$. The results hold for larger number of arms. For the TS-based policy, we set the prior distribution to be uniform over a finite support $\{0, 0.1, 0.2, \ldots, 0.9, 1.0\}$. The regrets of these algorithms are shown in Figure 2a, in which we use the Monte Carlo simulation with $1,000$ independent trials. R(MA)^{2}B-UCB achieves the lowest accumulated regret. The reason explaining this phenomenon is that Restless-UCB sacrifices the regret performance for a lower computation complexity and thus performs worse compared to the online TS-based policy. On the other hand, R(MA)^{2}B-UCB achieves the best performance, partly due to leveraging our near-optimal index policy (see Remark 3). When the number of samples are sufficiently large (i.e, T is large), R(MA)^{2}B-UCB achieves near optimal performance.

We further compare the average running time. In this experiment, the horizon is $T = 60,000$. The results are presented in Figure 2b, which are averaged over 100 Monte Carlo runs of a single-threaded program on Intel Core i5-6400 desktop with 16 GB RAM. It is clear that R(MA)^{2}B-UCB is more efficient in terms of running time. For example, R(MA)^{2}B-UCB reduces the running time by up to 52% (resp. 70%) compared to Restless-UCB (resp. TS-based policy) when there are 10 arms, and reduces the corresponding running time by up to 26% (resp. 48%) when there are 100 arms. The improvement over colored-UCRL2 is even more significant with a larger number of arms since the time complexity of colored-UCRL2 grows exponentially with the number of arms. Hence we omit the comparison here. The significant improvement comes from the intrinsic design of our policy which only needs to solve an LP once, while the Restless-UCB needs a computation-intensive numerical approximation of the Oracle (e.g., Whittle index policy) and the TS-based policy is an online episodic algorithm which solves a Bellman equation for every episode.

We further evaluate R(MA)^{2}B-UCB under multi-action settings by considering a more general Markov process in which any two arbitrary states may communicate with each other and the transition probability matrices are randomly generated. The other settings remain the same as in the index policy evaluation. For the ease of exposition, we consider the number of actions to be $2, 3$ and 5. Figure 2c shows the accumulated regret vs. time for R(MA)^{2}B-UCB under different numbers of actions. Since the Restless-UCB and TS-based policies are hard to be extended to the multi-action setting, we do not consider them in this comparison. From Figure 2c, we observe that R(MA)^{2}B-UCB achieves $\sqrt{T}$ regret under multi-action settings, which validates our theoretical contributions in the paper (see Theorem 2). Furthermore, when the number of actions increases, it takes a larger number of time steps for the accumulated regret to converge. In other words, the planning phase in R(MA)^{2}B-UCB (see Algorithm 1) will take a longer time to learn the system parameters.

Conclusion

In this paper, we studied the restless multi-armed, multi-action bandit problem (R(MA)^{2}B) with a finite horizon. Since the problem is typically intractable, we first proposed an asymptotically optimal index policy entitled *OMR Index Policy*, which is computationally feasible. Since the system parameters are often unavailable in practice, we then adopted a learning perspective toward the index policy. We proposed a generative model based reinforcement learning augmented algorithm named R(MA)^{2}B-UCB, which can fully exploit the structure of the proposed *OMR Index Policy*. We proved that R(MA)^{2}B-UCB achieves a sub-linear regret with a low computation complexity. Our experimental results further validated our theoretical results.

Acknowledgements

This work was supported in part by NSF grant CRII-CNS-2104880 and the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technologies Office Award Number 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

Ahmad, S. H. A.; Liu, M.; Javidi, T.; Zhao, Q.; and Krishnamachari, B. 2009. Optimality of Myopic Sensing in Multichannel Opportunistic Access. *IEEE Transactions on Information Theory*, 55(9): 4040–4050.

Altman, E. 1999. *Constrained Markov Decision Processes*, volume 7. CRC Press.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-Time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2): 235–256.

Bellman, R. 2010. *Dynamic Programming*. USA: Princeton University Press. ISBN 0691146683.

Bertsekas, D. P. 1995. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific Belmont, MA.

Bertsimas, D.; and Niño-Mora, J. 2000. Restless Bandits, Linear Programming Relaxations, and A Primal-Dual Index Heuristic. *Operations Research*, 48(1): 80–90.

Brown, D. B.; and Smith, J. E. 2020. Index Policies and Performance Bounds for Dynamic Selection Problems. *Management Science*, 66(7): 3029–3050.

Dai, W.; Gai, Y.; Krishnamachari, B.; and Zhao, Q. 2011. The Non-Bayesian Restless Multi-Armed Bandit: A Case of Near-Logarithmic Regret. In *Proc. of IEEE ICASSP*.

Deo, S.; Iravani, S.; Jiang, T.; Smilowitz, K.; and Samuelson, S. 2013. Improving Health Outcomes Through Better Capacity Allocation in A Community-based Chronic Care Model. *Operations Research*, 61(6): 1277–1294.

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

Gittins, J. 1974. A Dynamic Allocation Index for the Sequential Design of Experiments. *Progress in Statistics*, 241–266.

HasanzadeZonuzy, A.; Kalathil, D.; and Shakkottai, S. 2021. Learning with Safety Constraints: Sample Complexity of Reinforcement Learning for Constrained MDPs. In *Proc. of AAAI*.

Hu, W.; and Frazier, P. 2017. An Asymptotically Optimal Index Policy for Finite-Horizon Restless Bandits. *arXiv preprint arXiv:1707.00205*.

Jacko, P. 2010. Restless Bandits Approach to the Job Scheduling Problem and Its Extensions. *Modern Trends in Controlled Stochastic Processes: Theory and Applications*, 248–267.

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

Jin, C.; Jin, T.; Luo, H.; Sra, S.; and Yu, T. 2019. Learning Adversarial MDPs with Bandit Feedback and Unknown Transition. *arXiv preprint arXiv:1912.01192*.

Jung, Y. H.; Abeille, M.; and Tewari, A. 2019. Thompson Sampling in Non-Episodic Restless Bandits. *arXiv preprint arXiv:1910.05654*.

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*.

Kallenberg, L. 2003. Finite State and Action MDPs. In *Handbook of Markov Decision Processes*, 21–87. Springer.

Kearns, M.; Mansour, Y.; and Ng, A. Y. 2002. A Sparse Sampling Algorithm for Near-Optimal Planning in Large Markov Decision Processes. *Machine Learning*, 49(2): 193–208.

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*.

Lee, E.; Lavieri, M. S.; and Volk, M. 2019. Optimal Screening for Hepatocellular Carcinoma: A Restless Bandit Model. *Manufacturing & Service Operations Management*, 21(1): 198–212.

Liu, H.; Liu, K.; and Zhao, Q. 2011. Logarithmic Weak Regret of Non-Bayesian Restless Multi-Armed Bandit. In *Proc. of IEEE ICASSP*.

Liu, H.; Liu, K.; and Zhao, Q. 2012. Learning in A Changing World: Restless Multi-Armed Bandit with Unknown Dynamics. *IEEE Transactions on Information Theory*, 59(3): 1902–1916.

Mahajan, A.; and Teneketzis, D. 2008. Multi-Armed Bandit Problems. In *Foundations and Applications of Sensor Management*, 121–151. Springer.

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

Maurer, A.; and Pontil, M. 2009. Empirical Bernstein Bounds and Sample Variance Penalization. *arXiv preprint arXiv:0907.3740*.

Nino-Mora, J. 2001. Restless Bandits, Partial Conservation Laws and Indexability. *Advances in Applied Probability*, 76–98.

Niño-Mora, J. 2007. Dynamic Priority Allocation via Restless Bandit Marginal Productivity Indices. *Top*, 15(2): 161–198.

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*.

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

Rosenberg, A.; and Mansour, Y. 2019. Online Convex Optimization in Adversarial Markov Decision Processes. In *Proc. of ICML*.

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.

Shiryaev, A. N. 2007. *Optimal Stopping Rules*, volume 8. Springer Science & Business Media.

Tekin, C.; and Liu, M. 2011. Adaptive Learning of Uncontrolled Restless Bandits with Logarithmic Regret. In *Proc. of Allerton*.

Tekin, C.; and Liu, M. 2012. Online Learning of Rested and Restless Bandits. *IEEE Transactions on Information Theory*, 58(8): 5588–5611.

Verloop, I. M. 2016. Asymptotically Optimal Priority Policies for Indexable and Nonindexable Restless Bandits. *The Annals of Applied Probability*, 26(4): 1947–1995.

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

Weber, R. R.; and Weiss, G. 1990. On An Index Policy for Restless Bandits. *Journal of Applied Probability*, 637–648.

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

Xiong, G.; Li, J.; and Singh, R. 2021. Reinforcement Learning for Finite-Horizon Restless Multi-Armed Multi-Action Bandits. *arXiv preprint arXiv:2109.09855*.

Xiong, G.; Singh, R.; and Li, J. 2021. Learning Augmented Index Policy for Optimal Service Placement at the Network Edge. *arXiv preprint arXiv:2101.03641*.

Zayas-Cabán, G.; Jasin, S.; and Wang, G. 2019. An Asymptotically Optimal Heuristic for General Nonstationary Finite-Horizon Restless Multi-Armed, Multi-Action Bandits. *Advances in Applied Probability*, 51(3): 745–772.

Zhang, X.; and Frazier, P. I. 2021. Restless Bandits with Many Arms: Beating the Central Limit Theorem. *arXiv preprint arXiv:2107.11911*.