Provably Efficient Reinforcement Learning for Adversarial Restless Multi-Armed Bandits with Unknown Transitions and Bandit Feedback

Guojun Xiong [1] **Jian Li** [1]

Abstract

Restless multi-armed bandits (RMAB) play a central role in modeling sequential decision making problems under an instantaneous activation constraint that at most B arms can be activated at any decision epoch. Each restless arm is endowed with a state that evolves independently according to a Markov decision process regardless of being activated or not. In this paper, we consider the task of learning in episodic RMAB with unknown transition functions and adversarial rewards, which can change arbitrarily across episodes. Further, we consider a challenging but natural bandit feedback setting that only adversarial rewards of activated arms are revealed to the decision maker (DM). The goal of the DM is to maximize its total adversarial rewards during the learning process while the instantaneous activation constraint must be satisfied in each decision epoch. We develop a novel reinforcement learning algorithm with two key contributors: a novel biased adversarial reward estimator to deal with bandit feedback and unknown transitions, and a low-complexity index policy to satisfy the instantaneous activation constraint. We show $\tilde{\mathcal{O}}(H\sqrt{T})$ regret bound for our algorithm, where T is the number of episodes and H is the episode length. To our best knowledge, this is the first algorithm to ensure $\tilde{\mathcal{O}}(\sqrt{T})$ regret for adversarial RMAB in our considered challenging settings.

1. Introduction

Restless multi-armed bandits (RMAB) (Whittle, 1988) has been widely used to study sequential decision making problems with an *"instantaneous activation constraint"* that at most B out of N arms can be activated at any decision epoch, ranging from wireless scheduling (Sheng et al., 2014; Cohen et al., 2014; Xiong et al., 2022b), resource allocation (Glazebrook et al., 2011; Larrañaga et al., 2014; Xiong et al., 2023; 2022d), to healthcare (Killian et al., 2021). Each arm is described by a Markov decision process (MDP) (Puterman, 1994), and evolves stochastically according to two different transition kernels, depending on whether the arm is activated or not. Rewards are generated with each transition. The goal of the decision maker (DM) is to maximize the total expected reward over a finite-horizon (Xiong et al., 2022a) or an infinite-horizon (Wang et al., 2020; Avrachenkov & Borkar, 2022; Xiong et al., 2022c; Xiong & Li, 2023; Wang et al., 2024) under the instantaneous activation constraint.

The majority of the literature on RMAB consider a stochastic environment, where both the rewards and dynamics of the environments are assumed to be stationary over time. However, in real-world applications such as online advertising and revenue management, rewards[1] are not necessarily stationary but can change arbitrarily between episodes (Lee & Lee, 2023). To this end, we study the problem of learning *a finite-horizon adversarial RMAB (*`ARMAB`*) with unknown transitions* over T episodes. In each episode, all arms start from a fixed initial state, and the DM repeats the followings for a fixed number of H decision epochs: determine whether or not to activate each arm while the instantaneous activation constraint must be satisfied, receive adversarial rewards from each arm, which transits to the next state according to some unknown transition functions. Specifically, we consider a challenging but natural *bandit feedback* setting, where the adversarial rewards in each decision epoch are only revealed to the DM when the state-action pairs are visited. The goal of the DM is to minimize its regret, which is the difference between its total adversarial rewards and the total rewards received by an optimal fixed policy.

To achieve this goal, we develop an episodic reinforcement learning (RL) algorithm named `UCMD-ARMAB`. First, to handle unknown transitions of each arm, we construct confi-

[1]Stony Brook University. Correspondence to: Guojun Xiong <guojun.xiong@stonybrook.edu>, Jian Li <jian.li.3@stonybrook.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Although most existing literature on adversarial learning use the term "loss" instead of "reward", we choose to use the latter, or more specifically "adversarial reward" in this paper to be consistent with RMAB literature. One can translate between rewards and losses by taking negation.

Table 1: Comparison with existing works, where T is the number of episodes and H is the length of each episode.

Paper	Model	Setting	Feedback	Constraint	Algorithm	Regret
(Rosenberg & Mansour, 2019b)	MDP	Adversarial	Full	✗	OMD	$\tilde{\mathcal{O}}(H\sqrt{T})$
(Rosenberg & Mansour, 2019a)	MDP	Adversarial	Bandit	✗	OMD	$\tilde{\mathcal{O}}(H^{2/3}T^{3/4})$
(Jin et al., 2020)	MDP	Adversarial	Bandit	✗	OMD	$\tilde{\mathcal{O}}(H\sqrt{T})$
(Luo et al., 2021)	MDP	Adversarial	Bandit	✗	Policy Optimization	$\tilde{\mathcal{O}}(H^2\sqrt{T})$
(Qiu et al., 2020)	CMDP	Adversarial	Full	Average	Primal-dual	$\tilde{\mathcal{O}}(H\sqrt{T})$
(Germano et al., 2023)	CMDP	Adversarial	Full	Average	Primal-dual	$\tilde{\mathcal{O}}(HT^{3/4})$
(Wang et al., 2020)	RMAB	Stochastic	Full	Hard	Generative model	$\tilde{\mathcal{O}}(H^{2/3}T^{2/3})$
(Xiong et al., 2022a;c)	RMAB	Stochastic	Full	Hard	Index-based	$\tilde{\mathcal{O}}(\sqrt{HT})$
This Work	**RMAB**	**Adversarial**	**Bandit**	**Hard**	**Index-based OMD**	$\tilde{\mathcal{O}}(H\sqrt{T})$

dence sets to guarantee the true ones lie in these sets with high probability (Section 4.1). Second, to handle adversarial rewards, we apply Online Mirror Descent (OMD) to solve a relaxed problem, rather than directly on ARMAB, in terms of occupancy measures (Section 4.2). This is due to the fact that ARMAB is known to be computationally intractable even in the offline setting (Papadimitriou & Tsitsiklis, 1994). We note that OMD has also been used in adversarial MDP (Rosenberg & Mansour, 2019b; Jin et al., 2020) and CMDP (Qiu et al., 2020). However, they considered a stationary occupancy measure due to the existing of stationary policies, which is not the case for our finite-horizon RMAB with instantaneous activation constraint. This requires us to leverage a time-dependent occupancy measure.

Third, a key difference compared to stochastic RMAB (Wang et al., 2020; Xiong et al., 2022a;c) is that with bandit feedback and to apply the above OMD, we must construct adversarial reward estimators since the rewards of arms are not completely revealed to the DM. We address this challenge by developing a novel biased overestimated reward estimator (Section 4.3) based on the observations of counts for each state-action pairs. Finally, to handle the instantaneous activation constraint in ARMAB, we develop a low-complexity index policy (Section 4.4) based on the solutions from the OMD in Section 4.2. This is another key difference compared to adversarial MDP or CMDP, which requires us to explicitly characterize the regret due to the implementation of such an index policy.

We prove that UCMD-ARMAB achieves $\tilde{\mathcal{O}}(H\sqrt{T})$ regret, where T is the number of episodes and H is the episode length. Although our regret bound exhibits a gap (i.e., $\sqrt{H}$ times larger) to that of stochastic RMAB (Xiong et al., 2022a;c), to our best knowledge, our result is the first to achieve $\tilde{\mathcal{O}}(\sqrt{T})$ regret for adversarial RMAB with bandit feedback and unknown transition functions, a harder problem compared to stochastic RMABs.

Notations. We use calligraphy letter $\mathcal{A}$ to denote a finite set with cardinality $|\mathcal{A}|$, and $[N]$ to denote the set of integers $\{1, \cdots, N\}$.

2. Related Work

We discuss our related work from three categories: MDP, CMDP and RMAB. In particular, we mainly focus on the adversarial settings for the former two.

Adversarial MDP. Even-Dar et al. (2009) proposed the adversarial MDP model with arbitrarily changed loss functions and a fixed stochastic transition function. The first to consider unknown transition function with full information feedback is Neu et al. (2012), which proposed a Follow-the-Perturbed-Optimistic algorithm with an $\tilde{\mathcal{O}}(H|\mathcal{S}||\mathcal{A}|\sqrt{T})$ regret. Rosenberg & Mansour (2019b) improved the bound to $\tilde{\mathcal{O}}(H|\mathcal{S}|\sqrt{|\mathcal{A}|T})$ through a UCRL2-based online optimization algorithm. A more challenging bandit feedback setting was considered in Rosenberg & Mansour (2019a), which achieves an $\tilde{\mathcal{O}}(H^{3/2}|\mathcal{S}||\mathcal{A}|^{1/4}T^{3/4})$ regret. Jin et al. (2020) further achieves an improved $\tilde{\mathcal{O}}(H|\mathcal{S}|\sqrt{|\mathcal{A}|T})$ regret via a novel reward estimator and a modified radius of upper confidence ball. Under a similar setting, a policy optimization method is developed in Luo et al. (2021) with $\tilde{\mathcal{O}}(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|T})$ regret. Another line of recent works further consider the settings of linear function approximation (Neu & Olkhovskaya, 2021), the best-of-both-world (Jin et al., 2021), delay bandit feedback (Jin et al., 2022) and adversarial transition functions (Jin et al., 2023).

Adversarial Constrained MDP (CMDP). CMDP (Altman, 1999) plays an important role in control and planning, which aims to maximize a reward over all available policies subject to constraints that enforce the fairness or safety of the policies. Qiu et al. (2020) is one of the first to study CMDP with adversarial losses and unknown transition function. A primal-dual algorithm was proposed with $\tilde{\mathcal{O}}(H|\mathcal{S}|\sqrt{|\mathcal{A}|T})$ regret and constraint violation. Germano et al. (2023) further considered both adversarial losses and constraints, and proposed a best-of-both-world algorithm, which achieves $\tilde{\mathcal{O}}(HT^{3/4})$ regret and constraint violation.

RMAB. RMAB was first introduced in Whittle (1988), and has been widely studied, see Niño-Mora (2023) and references therein. In particular, RL algorithms have been proposed for RMAB with unknown transitions. Colored-

UCRL2 is the state-of-the-art method for online RMAB with $\tilde{\mathcal{O}}(\sqrt{HT})$ regret. To address the exponential computational complexity of colored-UCRL2, low-complexity RL algorithms have been developed. For example, Wang et al. (2020) proposed a generative model-based algorithm with $\tilde{\mathcal{O}}(H^{2/3}T^{2/3})$ regret, and Xiong et al. (2022a;c) designed index-aware RL algorithms for both finite-horizon and infinite-horizon average reward settings with $\tilde{\mathcal{O}}(\sqrt{HT})$ regret[2]. However, most of the existing literature on RMAB focus on the stochastic setting, where the reward functions are stochastically sampled from a fixed distribution, either known or unknown. To our best knowledge, this work is the first to study RMAB in adversarial settings with unknown transition functions and bandit feedback.

3. Model and Problem Formulation

We formally define the adversarial RMAB, and introduce the online settings considered in this paper.

3.1. ARMAB: Adversarial RMAB

Consider an episodic adversarial RMAB with N arms. Each arm $n \in [N]$ is associated with a unichain MDP denoted by a tuple $(\mathcal{S}, \mathcal{A}, P_n, \{r_n^t, \forall t \in [T]\}, H)$. $\mathcal{S}$ is a finite state space, and $\mathcal{A} := \{0, 1\}$ is the set of binary actions. Using the standard terminology from RMAB literature, an arm is *passive* when action $a = 0$ is applied to it, and *active* otherwise. $P_n : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the transition kernel with $P_n(s'|s, a)$ being the probability of transition to state s' from state s by taking action a. T is the number of episodes, each of which consists of H decision epochs. $r_n^t : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the adversarial reward function in episode t. For simplicity, let $r_n(s, 0) = 0, \forall s \in \mathcal{S}, n \in [N]$. We do not make any statistical assumption on the adversarial reward functions, *which can be chosen arbitrarily*[3].

At decision epoch $h \in [H]$ in episode $t \in [T]$, an arm can be either active or passive. A policy determines what action to apply to each arm at h under the instantaneous activation constraint that at most B arms can be activated. Denote such a feasible policy in episode t as π^t, and let $\{(S_n^{t,h}, A_n^{t,h}) \in \mathcal{S} \times \mathcal{A}, \forall t \in [T], h \in [H], n \in [N]\}$ be the random tuple generated according to transition functions $\{P_n, \forall n \in [N]\}$ and π^t. The corresponding expected reward in episode t is

$$R_t(\pi^t) := \mathbb{E}\left[\sum_{h=1}^{H}\sum_{n=1}^{N} r_n^t(S_n^{t,h}, A_n^{t,h}) \Big| \{P_n, \forall n\}, \pi^t\right]. \quad (1)$$

[2]For a fair comparison, the total time horizon for stochastic RMAB (Wang et al., 2020; Xiong et al., 2022a;c) is set to be HT.

[3]However, in stochastic RMAB, rewards follow a stochastic distribution, which is fixed between episodes. This leads to a different objective in adversarial RMAB as the DM must adapt to dynamic and potentially hostile reward structures while striving to find an optimal policy across all episodes.

The DM's goal is to find a policy π for all T episodes which maximizes the total expected adversarial reward under the instantaneous active constraint, i.e.,

$$\texttt{ARMAB}: \max_{\pi} \; R(T, \pi) := \sum_{t=1}^{T} R_t(\pi)$$
$$\text{s.t.} \sum_{n=1}^{N} A_n^{t,h} \le B, \forall h \in [H], t \in [T]. \quad (2)$$

It is known that even when the transition kernel of each arm and the adversarial reward functions in each episode are revealed to the DM at the very beginning, finding an optimal policy over T episodes, denoted as π^{opt} for ARMAB (2) is PSPACE-hard (Papadimitriou & Tsitsiklis, 1994). The fundamental challenge lies in the explosion of state space and the curse of dimensionality prevents computing optimal policies. Exacerbating this challenge is the fact that transition kernels are often unknown in practice, and adversarial reward functions are only revealed to the DM at the end of each episode in adversarial settings (see Section 3.2).

Remark 3.1. Most existing works on adversarial MDP (Rosenberg & Mansour, 2019b;a; Jin et al., 2020) and CMDP (Qiu et al., 2020; Germano et al., 2023) assume that the state space is loop-free. In other words, the state space $\mathcal{S}$ can be divided into L distinct layers, i.e., $\mathcal{S} := \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_L$ with a singleton initial layer $\mathcal{S}_1 = \{s_1\}$, a terminal layer $\mathcal{S}_L = \{s_L\}$, and $\mathcal{S}_\ell \cap \mathcal{S}_j = \emptyset, j \neq \ell$. Transitions only occur between consecutive layers, i.e., $P(s'|s, a) > 0$ if $s' \in \mathcal{S}_{\ell+1}, s \in \mathcal{S}_\ell, \forall \ell \in [L]$. On one hand, many practical problems do not have a loop-free MDP. On the other hand, this assumption requires any non-loop-free MDP to extend its state space L times to be transformed into a loop-free MDP with a fixed length L. This often enlarges the regret bound at least L times. In this paper, we consider a general MDP without such a restrictive loop-free assumption.

3.2. Online Setting and Learning Regret

We focus on the online adversarial settings where the underlying MDPs are unknown to the DM, and the adversarial reward is of bandit-feedback[4]. The interaction between the DM and the ARMAB environment is presented in Algorithm 1. Only the state and action spaces are known to the DM in advance, and the interaction proceeds in T episodes. At the beginning of episode t, the adversary determines the adversarial reward functions, each arm n starts from a fixed state $S_n^0, \forall n \in [N]$, and the DM determines a policy π^t and then executes this policy for each decision epoch $h \in [H]$ in this episode. Specifically, at decision epoch h, the DM chooses actions $\{A_n^{t,h}, \forall n \in \mathcal{N}\}$ for each arm

[4]We use the term "bandit-feedback" as in Rosenberg & Mansour (2019a); Jin et al. (2020) to denote that only the adversarial rewards of visited state-action pairs are revealed to the DM.

Algorithm 1 Online Interactions between the DM and the Adversarial RMAB Environment

Require: State space $\mathcal{S}$, action space $\mathcal{A}$, and unknown transition functions $\{\mathcal{P}_n, \forall n\}$;
for $t = 1$ to T **do**
  All arms start in state $S_n^0, \forall n$;
  Adversary decides the reward function $\{r_n^t, \forall n\}$, and the DM decides a policy π^t;
  for $h = 1$ to H **do**
    DM chooses actions $\{A_n^{t,h}, \forall n\}$ under the instantaneous activation constraint; observes adversarial reward $\{r_n^t(S_n^{t,h}, A_n^{t,h}), \forall n\}$;
    Arm $n, \forall n$ moves to the next state $S_n^{t,h+1} \sim P_n(\cdot|S_n^{t,h}, A_n^{t,h})$;
    DM observes states $\{S_n^{t,h+1}, \forall n\}$.
  end for
end for

according to $\pi^t(\cdot|S_n^{t,h})$ under the instantaneous activation constraint, i.e., $\sum_{n=1}^N A_n^{t,h} = B$, and each arm then moves to the next state $S_n^{t,h+1}$ sampled from $P_n(\cdot|S_n^{t,h}, A_n^{t,h})$. The DM records the trajectory of the current episode t and the adversarial rewards in each decision epoch are *only revealed to the DM when the state-action pairs are visited due to bandit-feedback.*

The DM's goal is to minimize its regret, as defined by

$$\Delta(T) := R(T, \pi^{opt}) - \sum_{t=1}^{T} R_t(\pi^t), \tag{3}$$

where $R(T, \pi^{opt})$ is the total expected adversarial rewards under the offline optimal policy π^{opt} by solving (2), and $R_t(\pi^t)$ is defined in (1). We simply refer to ARMAB as in the online setting in the rest of this paper.

Although the above definition is similar to that in stochastic settings, the fundamental difference is that the offline policy π^{opt} is only optimal when it is defined over all T episodes, and it is not guaranteed to be optimal in each episode, which is the case for stochastic setting. This is because the adversarial rewards can change arbitrarily between episodes rather than following some fixed (unknown) distribution as in the stochastic setting. This fundamental difference will necessitate new techniques in terms of regret characterization, which we will discuss in details in Section 5.

4. RL Algorithm for ARMAB

We show that it is possible to design a RL algorithm to solve the regret minimization problem (3) for the computationally intractable ARMAB. Specifically, we leverage the popular UCRL-based algorithm to the online adversarial RMAB setting, and develop an episodic RL algorithm named UCMD-ARMAB. There are four key components of our algorithm: (1) maintaining a confidence set of the transition functions; (2) using online mirror descent (OMD) to solve a relaxed version of ARMAB in terms of occupancy measure to deal with adversarial rewards; (3) constructing an adversarial reward estimator to deal with bandit feedback; and (4) designing a low-complexity index policy to ensure that the instantaneous activation constraint is satisfied in each decision epoch. We summarize our UCMD-ARMAB in Algorithm 2, which operates in an episodic manner with a total of T episodes and each episode including H decision epochs. For simplicity, let $\tau_t := H(t-1)+1$ be the starting time of the t-th episode.

Algorithm 2 UCMD-ARMAB

Require: Initialize $C_n^1(s,a) = 0$, $\hat{P}_n^1(s'|s,a) = 1/|\mathcal{S}|$
for $t = 1, 2, \cdots, T$ **do**
  Construct $\mathcal{P}_n^t(s,a)$ according to (5) at τ_t;
  Construct the adversarial reward estimator $\hat{r}_n^t(s,a), \forall s,a,n$ according to (14);
  Obtain a relaxed ARMAB (7) in terms of occupancy measure, and solve (12) with OMD;
  Construct an index policy π^t according to (15).
end for

4.1. Confidence Sets

As discussed in Section 3, ARMAB has two components: a stochastic transition function, and an adversarial reward function for each arm. Since transition functions are unknown to the DM, we maintain confidence sets via past sample trajectories, which contain true transition functions $P_n, \forall n$ with high probability. Specifically, UCMD-ARMAB maintains two counts for each arm n. Let $C_n^{t-1}(s,a), \forall n \in [N]$ be the number of visits to state-action pairs (s,a) until τ_t, and $C_n^{t-1}(s,a,s'), \forall n \in [N]$ be the number of transitions from s to s' under action a. At episode t, UCMD-ARMAB updates these two counts as: $\forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$

$$C_n^t(s,a) = C_n^{t-1}(s,a) + \sum_{h=1}^{H} \mathbb{1}(S_n^{t,h} = s, A_n^{t,h} = a),$$
$$C_n^t(s,a,s') = C_n^{t-1}(s,a,s') + \sum_{h=1}^{H} \mathbb{1}(S_n^{t,h+1} = s' | S_n^{t,h} = s, A_n^{t,h} = a).$$

UCMD-ARMAB estimates the true transition function by the corresponding empirical average as:

$$\hat{P}_n^t(s'|s,a) = \frac{C_n^{t-1}(s,a,s')}{\max\{C_n^{t-1}(s,a), 1\}}, \tag{4}$$

and then defines confidence sets at episode t as

$$\mathcal{P}_n^t(s,a) := \{\tilde{P}_n^t(s'|s,a), \forall s' : |\tilde{P}_n^t(s'|s,a) - \hat{P}_n^t(s'|s,a)| \le \delta_n^t(s,a)\}, \quad (5)$$

where the confidence width $\delta_n^t(s,a)$ is built according to the Hoeffding inequality (Maurer & Pontil, 2009) as: for $\epsilon \in (0,1)$

$$\delta_n^t(s,a) = \sqrt{\frac{1}{2C_n^{t-1}(s,a)} \log\Big(\frac{4|\mathcal{S}||\mathcal{A}|N(t-1)H}{\epsilon}\Big)}. \quad (6)$$

Lemma 4.1. *With probability at least $1-2\epsilon$, the true transition functions are within the confidence sets, i.e., $P_n \in \mathcal{P}_n^t$, $\forall n \in [N], t \in [T]$.*

4.2. Solving Relaxed ARMAB with OMD

Recall that solving ARMAB is computationally expensive even in the offline setting (Papadimitriou & Tsitsiklis, 1994). To tackle this challenge, we first relax the instantaneous activation constraint, i.e., the activation constraint is satisfied on average, and obtain the following relaxed problem of

$$\max_{\pi} \; R(T,\pi) := \sum_{t=1}^{T} R_t(\pi)$$
$$\text{s.t. } \mathbb{E}_\pi\Big[\sum_{n=1}^{N} A_n^{t,h}\Big] \le B, \; h \in [H], t \in [T]. \quad (7)$$

It turns out that this relaxed ARMAB can be equivalently transformed into a linear programming (LP) using occupancy measure (Altman, 1999). More specifically, the occupancy measure μ of a policy π for a finite-horizon MDP is defined as the expected number of visits to a state-action pair (s,a) at each decision epoch h. Formally,

$$\mu^\pi = \Big\{\mu_n(s,a;h) = \mathbb{P}(S_n^h = s, A_n^h = a) : \forall n \in [N], s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]\Big\}. \quad (8)$$

It can be easily checked that occupancy measures satisfy the following two properties. First,

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mu_n(s,a;h) = 1, \forall n \in [N], h \in [H], \quad (9)$$

with $0 \le \mu_n(s,a;h) \le 1$. Hence the occupancy measure μ_n, $\forall n$ is a probability measure. Second, the fluid balance exists in occupancy measure transitions as

$$\sum_{a\in\mathcal{A}} \mu_n(s,a;h) = \sum_{s'}\sum_{a'} \mu_n(s',a';h-1) P_n(s',a',s). \quad (10)$$

For ease of presentation, we relegate the details of the equivalent LP of (7) to the supplementary materials.

Remark 4.2. Occupancy measure has been widely used in adversarial MDP (Rosenberg & Mansour, 2019b; Jin et al., 2020) and CMDP (Qiu et al., 2020). Since there exists a stationary policy in these settings, the regret minimization problem in Rosenberg & Mansour (2019b); Jin et al. (2020); Qiu et al. (2020) can be equivalently reduced to an online linear optimization in terms of the stationary occupancy measure. Unlike these works, there is no such stationary policy for our considered finite-horizon RMAB with the instantaneous activation constraint, and hence we cannot reduce our regret (3) into a linear optimization using stationary occupancy measure. To address this additional challenge, we leverage the time-dependent occupancy measure (8), and the regret minimization calls for different proof techniques (see Section 5 for details).

Unfortunately, we cannot solve this LP since we have no knowledge about the true transition functions and adversarial rewards. Similar to the stochastic setting (Xiong et al., 2022a;c) and with the confidence sets defined in Section 4.1, we can further rewrite this LP as an extended LP by leveraging the *state-action-state occupancy measure* $z_n^t(s,a,s';h)$ defined as $z_n^t(s,a,s';h) = P_n(s'|s,a)\mu_n^t(s,a;h)$ to express the confidence intervals of the transition probabilities. Unlike Xiong et al. (2022a;c), the adversarial rewards can change arbitrarily between episodes, and hence we also need to guarantee that the updated occupancy measure in episode t does not deviate too much away from the previously chosen occupancy measure in episode $t-1$. Thus, we further incorporate $D(z^t||z^{t-1})$ into the objective function, which is the unnormalized Kullback-Leible (KL) divergence between two occupancy measures, which is defined as

$$D(z^t||z^{t-1}) := \sum_{h=1}^{H}\sum_{s,a,s'} z^t(s,a,s';h) \ln \frac{z^t(s,a,s';h)}{z^{t-1}(s,a,s';h)} - z^t(s,a,s';h) + z^{t-1}(s,a,s';h). \quad (11)$$

The DM needs to solve a non-linear problem over $z^t := \{z_n^t(s,a,s';h), \forall n \in [N]\}$ for a given parameter $\eta > 0$:

$$\max_{z^t} \sum_{h=1}^{H}\sum_{n=1}^{N}\sum_{(s,a,s')} \eta z_n^t(s,a,s';h)\hat{r}_n^{t-1}(s,a) - D(z^t||z^{t-1})$$
$$\text{s.t.} \sum_{n=1}^{N}\sum_{(s,a,s')} a z_n^t(s,a,s';h) \le B, \quad \forall h \in [H],$$
$$\sum_{a,s'} z_n^t(s,a,s';h) = \sum_{s',a'} z_n^t(s',a',s;h-1), \forall h \in [H],$$
$$\frac{z_n^t(s,a,s';h)}{\sum_y z_n^t(s,a,y;h)} - (\hat{P}_n^t(s'|s,a) + \delta_n^t(s,a)) \le 0,$$
$$-\frac{z_n^t(s,a,s';h)}{\sum_y z_n^t(s,a,y;h)} + (\hat{P}_n^t(s'|s,a) - \delta_n^t(s,a)) \le 0,$$
$$z_n^t(s,a,s';h) \ge 0, \forall s,s' \in \mathcal{S}, a \in \mathcal{A}, \forall n \in [N], \quad (12)$$

where $\hat{r}_n^{t-1}(s,a)$ is the estimated adversarial reward due to bandit feedback, and we formally define it in Section 4.3.

This problem has $O(|\mathcal{S}|^2|\mathcal{A}|HN)$ constraints and decision variables. Inspired by Rosenberg & Mansour (2019b), we solve (12) via OMD to choose the occupancy measure for each episode t. We first solve a unconstrained problem by setting $\tilde{z}^t(s,a,s';h) = z^{t-1}(s,a,s';h)e^{\eta\hat{r}^{t-1}(s,a)}$. Then we project the unconstrained maximizer $\tilde{z}^t$ into the feasible set $\mathcal{Z}^t$, which is defined by the constraints in (12). This can be reduced to a convex optimization problem, and can be efficiently solved using iterative methods (Boyd & Vandenberghe, 2004). For ease of readability, we relegate the details of solving (12) to the supplementary materials. Denote the optimal solution to (12) as $z^{t,\star}$.

4.3. Adversarial Reward Estimators

Since we consider a challenging bandit feedback setting, where the adversarial rewards in each decision epoch are revealed to the DM only when the state-action pairs are visited, we need to construct adversarial reward estimators based on observations. Specifically, we build upon the inverse importance-weighted estimators based on the observation of counts for each state-action pairs. Given the trajectory for episode t, a straightforward estimator is

$$\frac{r_n(s,a)}{\max\{c_n^t(s,a),1\}/H}\mathbb{1}(\exists h, S_n^{t,h}=s, A_n^{t,h}=a), \quad (13)$$

where $c_n^t(s,a) := \sum_{h=1}^{H}\mathbb{1}(S_n^{t,h}=s, A_n^{t,h}=a)$ is the number of visits to state-action pair (s,a) in episode t. For simplicity, we denote $\bar{r}_n^t(s,a) := \frac{r_n(s,a)}{\max\{c_n^t(s,a),1\}/H}$. Clearly, $\bar{r}_n^t(s,a)\mathbb{1}(\exists h, S_n^{t,h}=s, A_n^{t,h}=a)$ is an unbiased estimator of $r_n^t(s,a)$ from the above definition. A key difference between the unbiased estimator in (13) and those in previous works on adversarial MDPs (Jin et al., 2020; Rosenberg & Mansour, 2019b) lies in the construction of the denominator in (13). Specifically, Jin et al. (2020); Rosenberg & Mansour (2019b) considered MDPs with a underlying stationary policy, and hence the evolution of the dynamics in the considered MDPs will converge to this stationary policy. This enables the construction of the denominator in the estimator using the occupancy measure for each for each state-action pair under such a policy. In contrast, it is known that the dynamics in RMAB cannot converge to the stationary policy due to the fact that RMABs implement an index policy (see Section 4.4) to deal with the instantaneous activation constraint, and it is only provably in the asymptotic regime (Weber & Weiss, 1990; Verloop, 2016). This render the approaches in Jin et al. (2020); Rosenberg & Mansour (2019b) not applicable to ours, and necessitates different methods to construct the estimator as in (13).

Since we consider the bandit feedback, we further leverage the idea of implicit exploration as inspired by Neu (2015); Jin et al. (2020) to further encourage exploration. Specifically, we further increase $\bar{r}_n^t(s,a)$ with a bonus term $\delta_n^t(s,a)$ to obtain a biased estimator $(\bar{r}_n^t(s,a)+\delta_n^t(s,a))\mathbb{1}(\exists h, S_n^{t,h}=s, A_n^{t,h}=a)$. Since $r_n^t(s,a)\in[0,1]$ and to guarantee that this biased estimator is an overestimate, we further add the term $1-\mathbb{1}(\exists h, S_n^{t,h}=s, A_n^{t,h}=a)$. Thus, our final adversarial reward estimator is $\hat{r}_n^t(s,a)$

$$= \min\Big((\bar{r}_n^t(s,a)+\delta_n^t(s,a))\mathbb{1}(\exists h, S_n^{t,h}=s, A_n^{t,h}=a), 1\Big) + 1 - \mathbb{1}(\exists h, S_n^{t,h}=s, A_n^{t,h}=a). \quad (14)$$

Given the above constructions, it is clear that $\hat{r}_n^t(s,a)$ is a biased estimator upper-bounded by 1 and is overestimating $r_n^t(s,a), \forall s,a,n,t$ with high probability. Using overestimates for adversarial learning with bandit feedback can be viewed as an optimism principle to encourage exploration. This is beneficial for the regret characterization as the deterministic overestimation as in Jin et al. (2020) to guarantee a tighter regret bound.

Remark 4.3. The reward estimator in stochastic RMAB (Xiong et al., 2022a;c) is simply defined as the sample mean using all trajectories up to episode t, which cannot be applied to our adversarial RMAB. This is due to the fact that the rewards are assumed to be drawn from an unknown but fixed distribution across all episodes for stochastic RMAB, while rewards can change arbitrarily between episodes for adversarial RMAB. The bandit-feedback setting further differentiates our adversarial RMAB from classical stochastic ones. Finally, we note that Jin et al. (2020) considered the bandit feedback for adversarial MDP, and also constructed an adversarial loss/reward estimator. Since the regret minimization problem in Jin et al. (2020) can be reduced to an online linear optimization using stationary occupancy measure (see Remark 4.2), the estimator can be constructed directly using the stationary occupancy measure. Since there is no such stationary policy for our finite-horizon adversarial RMAB, this makes the estimator in Jin et al. (2020) not directly applicable to ours, and necessitates different construction techniques as discussed above.

4.4. Index Policy for ARMAB

Unfortunately, the optimal solution $z^{t,\star}$ to (12) is not always feasible for our ARMAB due to the fact that the instantaneous activation constraint in ARMAB must be satisfied in each decision epoch rather than on the average sense as in (12). Inspired by Xiong et al. (2022a;c), we further construct an index policy on top of $z^{t,\star}$ that is feasible for ARMAB. Specifically, since $\mathcal{A} := \{0,1\}$, i.e., an arm can be either active or passive at each decision epoch h, we define the index assigned to arm n in state $S_n^{t,h}=s$ at decision epoch

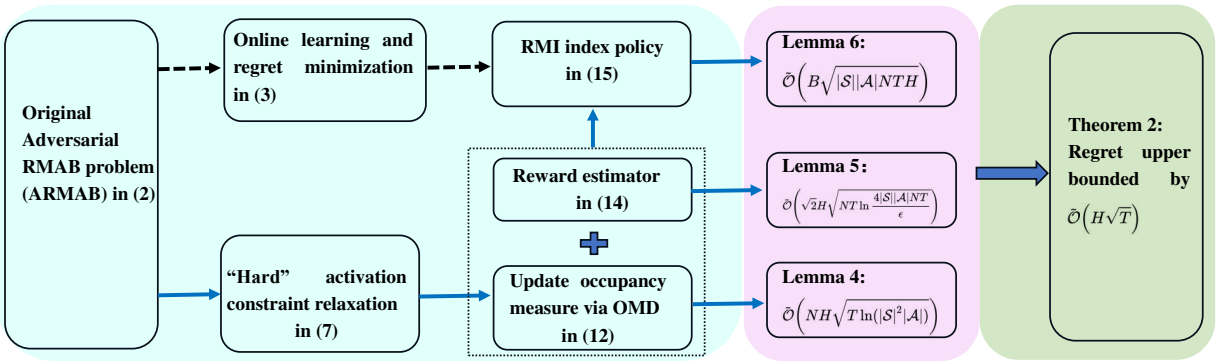


Figure 1: The workflow of UCMD−ARMAB and its regret analysis. The dashed arrows present the aimed procedures for solving the original problem in (2), and the solid arrows show the true procedures of UCMD−ARMAB. By relaxing the "hard" activation constraint as shown in (7), UCMD−ARMAB updates occupancy measure via OMD as in (12) (see Section 4.2), combined with the adversarial reward estimator in (14) (see Section 4.3). Then, it establishes the RMI index policy in (15) (see Section 4.4). These correspond to the **three sources of learning regret**, i.e., regret due to (i) OMD online optimization (Lemma 5.4), (ii) bandit-feedback adversarial reward (Lemma 5.5), and (iii) the RMI index policy (Lemma 5.6).

h of episode t to be as

$$\mathcal{I}_n^t(s;h) := \frac{\sum_{s'} z_n^{t,\star}(s,1,s';h)}{\sum_{b,s'} z_n^{t,\star}(s,b,s';h)}, \quad \forall n \in [N]. \tag{15}$$

We call this the reward-maximizing index (RMI) since $\mathcal{I}_n^t(s;h)$ indicates the probability of activating arm n in state s at decision epoch h of episode t towards maximizing the total expected adversarial rewards. To this end, we rank all arms according to their indices in a non-increasing order, and activate the set of B highest indexed arms at each decision epoch h. All remaining arms are kept passive at decision epoch h. We denote the resultant RMI policy as $\pi^t := \{\pi_n^t, \forall n\}$, and execute it in episode t.

Theorem 4.4. *The RMI policy is asymptotically optimal when both the number of arms and instantaneous activation constraint are enlarged by ρ times with $\rho \to \infty$.*

5. Analysis

In this section, we bound the regret of our UCMD−ARMAB.

5.1. Main Results

Theorem 5.1. *With probability at least $1-3\epsilon$, the regret of UCMD−ARMAB with $\eta = \sqrt{\frac{\ln(|\mathcal{S}|^2|\mathcal{A}|)}{T}}$ satisfies*

$$\Delta(T) = \tilde{\mathcal{O}}\Bigg(NH\sqrt{T\ln(|\mathcal{S}|^2|\mathcal{A}|)} + H\sqrt{2NT\ln\frac{4|\mathcal{S}||\mathcal{A}|NT}{\epsilon}} + B\sqrt{|\mathcal{S}||\mathcal{A}|NTH}\Bigg). \tag{16}$$

The regret in (16) contains three terms. The first term is the regret due to the OMD online optimization for occupancy measure updates (Section 4.2). The second term represents the regret due to bandit-feedback of adversarial rewards (Section 4.3). The third term comes from the implementation of our RMI policy for ARMAB (to satisfy the instantaneous activation constraint, Section 4.4). Clearly, the regret of UCMD−ARMAB is in the order of $\tilde{\mathcal{O}}(H\sqrt{T})$. This is the same as that for adversarial MDP (Rosenberg & Mansour, 2019b; Jin et al., 2020) and CMDP (Qiu et al., 2020). However, none of them consider an instantaneous activation constraint as in our ARMAB, which requires us to design a low-complexity index policy, and explicitly characterize its impact on the regret. Although our regret bound exhibits a gap (i.e., $\sqrt{H}$ times larger) to that of stochastic RMAB (Xiong et al., 2022a;c), to the best of our knowledge, our result is the first to achieve $\tilde{\mathcal{O}}(\sqrt{T})$ regret. Recall that comparing to the stochastic RMAB, we are considering a harder problem with a challenging setting, i.e., the rewards can change arbitrarily between episodes rather than following a fixed distribution, and with bandit feedback where only the adversarial reward of visited state-action pairs are revealed to the DM. This challenging setting thus requires us to design a novel adversarial reward estimator coupled with the OMD online optimization procedure.

5.2. Proof Sketch

As discussed earlier, ARMAB (2) is computationally intractable, and hence we cannot directly solve it and transform the regret minimization in (3) into an online linear optimization problem. This makes existing regret analysis for adversarial MDP (Rosenberg & Mansour, 2019b;

Jin et al., 2020) and CMDP (Qiu et al., 2020) not directly applicable to ours, and necessitates different proof techniques. To address this challenge and inspired by stochastic RMAB (Xiong et al., 2022a;c), we instead work on the relaxed problem (7), which achieves a provably upper bound on the adversarial rewards of ARMAB (2). In other words, the occupancy measure-based solutions to (7) provide an upper bound of the optimal adversarial reward $R(T, \pi^{opt})$ achieved by the offline optimal policy π^{opt} of ARMAB (2). We state this result formally in the following lemma.

Lemma 5.2. *There exists a set of occupancy measures $\mu_\pi^* := \{\mu_n^*(s,a;h), \forall n \in [N], s \in \mathcal{S}, a \in \mathcal{A}, h \in [h]\}$ under policy π^* that optimally solve the equivalent LP of the relaxed problem (7). In addition, $\sum_{t=1}^{T}\langle \mu_\pi^*, r^t\rangle$ is no less than $R(T, \pi^{opt})$, where $r^t := \{r_n^t(s,a), \forall n, s, a\}$.*

The proof of Theorem 5.1 then starts with a regret decomposition. Unlike stochastic RMAB (Xiong et al., 2022a;c), we cannot simply decompose the regret over episodes. This is because in our ARMAB, the adversarial rewards can change arbitrarily between episodes, and the offline optimal policy π^{opt} is only optimal when it is defined over all T episodes and it is not guaranteed to be optimal in each episode (see Section 3). As a result, the regret analysis for stochastic RMAB (Xiong et al., 2022a;c) is not applicable to ours. To address this challenge, we instead decompose the regret in terms of its coming sources. Specifically, we first need to characterize the regret due to solving the relaxed problem with OMD in terms of occupancy measure (Section 4.2). We then bound the regret due to the biased overestimated reward estimator (Section 4.3). Finally, we bound the regret of implementing our RMI policy (Section 4.4). We visualize these three steps in Figure 1 and provide a proof sketch herein. Combining them together gives rise to our main theoretical results in Theorem 5.1.

Regret decomposition. We formally state our regret decomposition in the following lemma.

Lemma 5.3. *Denote π^* as the optimal policy to equivalent LP of the relaxed problem (7) and $\{\tilde{\pi}^t, \forall t\}$ are polices executed on the MDPs at each episode with transition kernels selected from the confidence set defined in (5). Let $R(T, \pi^*, \{\hat{r}_n^t, \forall n, t\})$ be the total adversarial rewards achieved by policy π^* with overestimated rewards $\{\hat{r}_n^t, \forall n, t\}$. Denote $R(T, \{\tilde{\pi}^t, \forall t\}, \{\hat{r}_n^t, \forall n, t\})$ and $R(T, \{\tilde{\pi}^t, \forall t\}, \{r_n^t, \forall n, t\})$ as the total adversarial rewards achieved by policy $\{\tilde{\pi}^t, \forall t\}$ with overestimated rewards $\{\hat{r}_n^t, \forall n, t\}$ and true reward $\{r_n^t, \forall n, t\}$, respectively. The regret in (3) can be upper bounded as*

$$\begin{aligned}\Delta(T) \leq & \underbrace{R(T, \pi^{opt}) - R(T, \pi^*, \{\hat{r}_n^t, \forall n, t\}), \forall n, t\})}_{Term_0 \leq 0} \\ & + \underbrace{R(T, \pi^*, \{\hat{r}_n^t, \forall n, t\}) - R(T, \{\tilde{\pi}^t, \forall t\}, \{\hat{r}_n^t, \forall n, t\})}_{Term_1} \\ & + \underbrace{R(T, \{\tilde{\pi}^t, \forall t\}, \{\hat{r}_n^t, \forall n, t\}) - R(T, \{\tilde{\pi}^t, \forall t\}, \{r_n^t, \forall n, t\})}_{Term_2} \\ & + \underbrace{R(T, \{\tilde{\pi}^t, \forall t\}, \{r_n^t, \forall n, t\}) - \sum_{t=1}^{T} R_t(\pi^t)}_{Term_3}. \end{aligned} \tag{17}$$

Specifically, $Term_0 \leq 0$ holds due to Lemma 5.2 and the overestimation of adversarial rewards. $Term_1$ is the performance gap between the optimal policy π^* of the relaxed problem (7) and the OMD updated policy $\{\tilde{\pi}^t, \forall t\}$ under the plausible MDP selected from the confidence set in (5). Since we leverage the OMD to update the occupancy measures, to bound $Term_1$, the key is to connect items in $Term_1$ with the occupancy measure, which leverages the result in Lemma 5.2. $Term_2$ is the regret due to the overestimation of the adversarial reward estimator, i.e., $\hat{r}_n^t \geq r_n^t, \forall n, t$. $Term_3$ is the performance gap between the policy $\{\tilde{\pi}^t, \forall t\}$ in the optimistic plausible MDP and the learned RMI index policy $\{\pi^t, \forall t\}$ for the true MDP. We bound it based on the count of visits of each state-action pair.

Bounding $Term_1$. We first bound $Term_1$, i.e., the regret due to OMD online optimization.

Lemma 5.4. *With probability $1 - 2\epsilon$, we have $Term_1 \leq \frac{NH\ln(|\mathcal{S}|^2|\mathcal{A}|)}{\eta} + \eta NH(T+1)$.*

Bounding $Term_1$ is equivalent to bound the inner product $\sum_{t=1}^{T}\sum_{n=1}^{N}\langle \mu_n^* - z_n^t, \hat{r}_n^t\rangle$, with $z_n^t, \forall n \in [N]$ being controlled by the OMD updates, a proper η (as given in Theorem 5.1) is required to guarantee $\tilde{\mathcal{O}}(H\sqrt{T})$ regret.

Bounding $Term_2$. We then bound $Term_2$, i.e., the regret due to bandit-feedback adversarial reward.

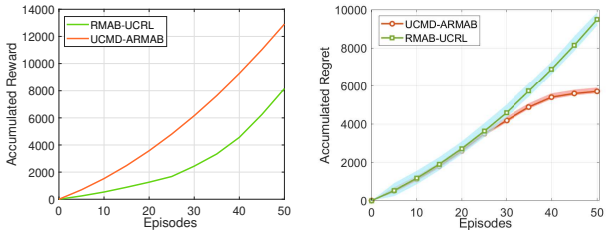
Lemma 5.5. *With probability $1 - 3\epsilon$, we have $Term_2 \leq H\sqrt{2NT\ln\frac{4|\mathcal{S}||\mathcal{A}|NT}{\epsilon}} + HN\sqrt{NT\ln\frac{1}{\epsilon}}$.*

Bounding $Term_2$ requires us to bound the inner product $\sum_{t=1}^{T}\sum_{n=1}^{N}\langle z_n^t, \hat{r}_n^t - r_n^t\rangle$, which is dominated by the gap of $\hat{r}_n^t - r_n^t$. An overestimation can guarantee that $\langle z_n^t, \hat{r}_n^t - r_n^t\rangle \geq 0, \forall n \in [N], t \in [T]$.

Bounding $Term_3$. Finally, we bound $Term_3$, i.e., the regret due to RMI index policy.

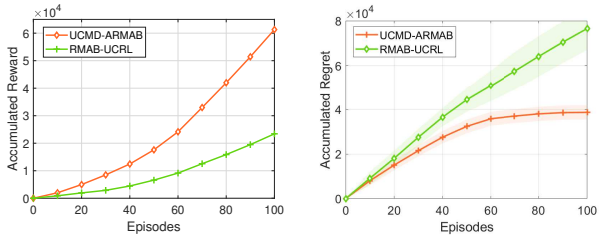
Lemma 5.6. *With probability $1 - 2\epsilon$, we have $Term_3 \leq \left(\sqrt{2\ln\frac{4|\mathcal{S}||\mathcal{A}|NTH}{\epsilon}} + 2B\right)\sqrt{|\mathcal{S}||\mathcal{A}|NTH}$.*

Since the adversarial rewards (i.e., $\hat{r}_n^t$) do not impact $Term_3$, we decompose $Term_3$ into each episode. The key is to characterize the number of visits of each state-action pair, which is related to the instantaneous activation constraint B and has a $\tilde{\mathcal{O}}(B\sqrt{TH})$ regret.



(a) Accumulated Reward. (b) Accumulated Regret.

Figure 2: The learning performance comparison for case study-CPAP.



(a) Accumulated reward. (b) Accumulated regret.

Figure 3: The learning performance comparison for case study-A Deadline Scheduling Problem.

6. Numerical Study

In this section, we demonstrate the utility of UCMD−ARMAB by evaluating it under two real-world applications of RMAB in the presence of adversarial rewards, i.e., the continuous positive airway pressure therapy (CPAP) (Kang et al., 2013; Herlihy et al., 2023; Li & Varakantham, 2022; Wang et al., 2024) and a deadline scheduling problem (Xiong et al., 2022a). The detailed setup of these two problems are provided in Appendix C.

We compare our UCMD−ARMAB with a benchmark named RMAB-UCRL (Xiong et al., 2022c), which was developed for stochastic RMAB, in terms of accumulated rewards and accumulated regret. As observed from Figure 2 and Figure 3, our UCMD−ARMAB significantly outperforms RMAB-UCRL in adversarial settings. In particular, our UCMD−ARMAB achieves a significant improvement over the accumulated reward as shown in in Figure 2a and Figure 3a, and exhibits a provably sublinear regret as shown in Figure 2b and Figure 3b, while the regret of RMAB-UCRL increases exponentially under adversarial settings. These verify the effectiveness of the proposed UCMD−ARMAB on handling adversarial RMABs.

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.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

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

Bercu, B., Delyon, B., Rio, E., et al. *Concentration inequalities for sums and martingales*. Springer, 2015.

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

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

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Germano, J., Stradi, F. E., Genalti, G., Castiglioni, M., Marchesi, A., and Gatti, N. A best-of-both-worlds algorithm for constrained mdps with long-term constraints. *arXiv preprint arXiv:2304.14326*, 2023.

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

Herlihy, C., Prins, A., Srinivasan, A., and Dickerson, J. P. 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*, pp. 732–740, 2023.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pp. 409–426, 1994.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2020.

Jin, T., Huang, L., and Luo, H. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 34:20491–20502, 2021.

Jin, T., Lancewicki, T., Luo, H., Mansour, Y., and Rosenberg, A. Near-optimal regret for adversarial mdp with delayed bandit feedback. *Advances in Neural Information Processing Systems*, 35:33469–33481, 2022.

Jin, T., Liu, J., Rouyer, C., Chan, W., We, C.-Y., and Luo, H. No-regret online reinforcement learning with adversarial losses and transitions. *arXiv preprint arXiv:2305.17380*, 2023.

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

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

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

Lee, D. and Lee, D. Online resource allocation in episodic markov decision processes. *arXiv preprint arXiv:2305.10744*, 2023.

Li, D. and Varakantham, P. Towards soft fairness in restless multi-armed bandits. *arXiv preprint arXiv:2207.13343*, 2022.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021.

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

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Neu, G. and Olkhovskaya, J. Online learning in mdps with linear function approximation and bandit feedback. *Advances in Neural Information Processing Systems*, 34: 10407–10417, 2021.

Neu, G., Gyorgy, A., and Szepesvári, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pp. 805–813. PMLR, 2012.

Niño-Mora, J. Markovian restless bandits and index policies: A review. *Mathematics*, 11(7):1639, 2023.

Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of optimal queueing network control. In *Proceedings of IEEE 9th annual conference on structure in complexity Theory*, pp. 318–322. IEEE, 1994.

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

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287, 2020.

Rosenberg, A. and Mansour, Y. Online stochastic shortest path with bandit feedback and unknown transition function. volume 32, 2019a.

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

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

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

Wang, S., Huang, L., and Lui, J. Restless-ucb, an efficient and low-complexity algorithm for online restless bandits. *Advances in Neural Information Processing Systems*, 33: 11878–11889, 2020.

Wang, S., Xiong, G., and Li, J. Online restless multi-armed bandits with long-term fairness constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15616–15624, 2024.

Weber, R. R. and Weiss, G. On An Index Policy for Restless Bandits. *Journal of applied probability*, pp. 637–648, 1990.

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

Xiong, G. and Li, J. Finite-time analysis of whittle index based q-learning for restless multi-armed bandits with neural network function approximation. *Advances in Neural Information Processing Systems*, 36:29048–29073, 2023.

Xiong, G., Li, J., and Singh, R. Reinforcement learning augmented asymptotically optimal index policy for finite-horizon restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8726–8734, 2022a.

Xiong, G., Qin, X., Li, B., Singh, R., and Li, J. Index-aware reinforcement learning for adaptive video streaming at the wireless edge. In *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 81–90, 2022b.

Xiong, G., Wang, S., and Li, J. Learning infinite-horizon average-reward restless multi-action bandits via index awareness. *Advances in Neural Information Processing Systems*, 35:17911–17925, 2022c.

Xiong, G., Wang, S., Li, J., and Singh, R. Whittle index based q-learning for wireless edge caching with linear function approximation. *arXiv preprint arXiv:2202.13187*, 2022d.

Xiong, G., Wang, S., Yan, G., and Li, J. Reinforcement learning for dynamic dimensioning of cloud caches: A restless bandit approach. *IEEE/ACM Transactions on Networking*, 2023.

A. Details and Proofs in Section 4

We provide all omitted details and proofs of the key lemmas of the main paper in this appendix.

A.1. Efficient Solver of Updating Occupancy Measure (12)

In this subsection, we provide details on how to efficiently solve the updating occupancy measure problem in (12). Similar to Rosenberg & Mansour (2019b), the solution of (12) can be yielded by decomposing the original problem into two subproblems. Specifically, the first subproblem is to solve the objective in (12) as an unconstrainted problem as

$$\tilde{z}^t = \arg\max_{z^t} \sum_{h=1}^{H} \sum_{n=1}^{N} \sum_{(s,a,s')} \eta z_n^t(s,a,s',h)\hat{r}_n^{t-1}(s,a) - D(z^t||z^{t-1}), \tag{18}$$

which can be easily solved and has a closed-form solution as

$$\tilde{z}_n^t(s,a,s';h) = z_n^{t-1}(s,a,s';h)e^{\eta \hat{r}_n^{t-1}(s,a)}, \forall s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], n \in [N]. \tag{19}$$

The second subproblem is then to project the solution in (19) to the feasible set of z^t in episode t by minimizing the KL-divergence of z^t and $\tilde{z}^t$. To simplify the notations, we denote the feasible set of z^t that satisfies the constraints in (12) as $\mathcal{Z}^t$, and thus the subproblem is expresses as follows

$$z^t = \arg\min_{z \in \mathcal{Z}^t} D(z||\tilde{z}^t). \tag{20}$$

Defining some Lagrangian parameters as $\beta : \mathcal{S} \mapsto \mathbb{R}$ and $\mu = (\mu^+, \mu^-)$ with $\mu^+, \mu^- : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}_{\geq 0}$, and $B_{n,t}^{\mu,\beta}(s,a,s',h)$ as

$$\begin{aligned} B_{n,t}^{\mu,\beta}(s,a,s',h) &= \mu^-(s,a,s',h) - \mu^+(s,a,s',h) + (\mu^+(s,a,s',h) + \mu^-(s,a,s',h))\delta_n^t(s,a) \\ &\quad + \beta(s',h) - \beta(s,h) + \eta\hat{r}_n^t(s,a) + \sum_{s''} \hat{P}_n^t(s''|s,a)(\mu^+(s,a,s'',h) - \mu^-(s,a,s'',h)). \end{aligned}$$

Hence, the following theorem from Rosenberg & Mansour (2019b) provides the solution of (20) and the detailed proof can be found in Jin et al. (2020).

Theorem A.1 (Theorem 4.2 (Rosenberg & Mansour, 2019b)). *Let $t > 1$ and define the function*

$$Z_n^{t,h}(\mu,\beta) = \sum_{s,a,s'} z_n^t(s,a,s',h)e^{B_{n,t}^{\mu,\beta}(s,a,s',h)},$$

the solution to (20) is

$$z_n^{t+1}(s,a,s',h) = \frac{z_n^t(s,a,s',h)e^{B_{n,t}^{\mu^t,\beta^t}(s,a,s',h)}}{Z_n^{t,h}(\mu^t,\beta^t)},$$

where μ^t, β^t are obtained by solving a convex optimization with non-negativity constraints as

$$\mu^t, \beta^t = \arg\min_{\mu,\beta>0} \sum_{h=1}^{H} \ln Z_n^{t,h}(\mu,\beta).$$

A.2. Index Policy Design and proof of Theorem 4.4

Following the method of the celebrated Whittle index (Whittle, 1988) for conventional RMAB problems, we first relax the "hard" activation constraint in (2) as an averaged constraint, which gives rise to the following relaxed problem.

$$\max_{\pi} \sum_{t=1}^{T} R_t \quad \text{s.t. } \mathbb{E}_\pi \left[\sum_{n=1}^{N} A_n^{t,h} \right] \leq B, h \in [H], t \in [T]. \tag{21}$$

Provided the definition of occupancy measure in (8), the relaxed problem (21) can be reformulated as a linear programming (LP) (Altman, 1999) expressed in the following lemma.

Lemma A.2. *The relaxed problem (21) is equivalent to the following LP*

$$\max_{\mu} \sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{(s,a)} \mu_n(s,a;h) r_n^t(s,a) \tag{22}$$

$$\text{s.t.} \sum_{n=1}^{N}\sum_{(s,a)} a\mu_n(s,a;h) \le B, \quad \forall h \in [H], \tag{23}$$

$$\sum_{a\in\mathcal{A}} \mu_n(s,a;h) = \sum_{(s',a')} \mu_n(s',a';h-1) P_n(s',a',s), \tag{24}$$

$$\mu_n(s,a;h) \ge 0, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], n \in [N], \tag{25}$$

where (23) is a restatement of the constraint in (21); (24) indicates the transition of the occupancy measure from time slot $h-1$ to time slot h; and (25) guarantees that the occupancy measures are non-negative.

The RMI is designed upon the optimal occupancy measure $\mu^\star = \{\mu_n^\star(s,a;h) : n \in [N], h \in [1,\dots,H]\}$ to the above LP. The first step is to construct a Markovian randomized policy $\xi_n^\star(s,a;h) := \frac{\mu_n^\star(s,a;h)}{\sum_{a\in\mathcal{A}} \mu_n^\star(s,a;h)}$, and then set $\xi_n^\star(s,1;h)$ as the RMI. Thereafter, we prioritize the users according to a decreasing order of their RMIs, and then activate the B arms with the highest indices. Note that our proposed RMI policy is well-defined even when the problem is not indexable (Whittle, 1988; Weber & Weiss, 1990).

Proof of Theorem 4.4. We now show that our RMI policy is asymptotically optimal when both the number of arms N and the activation constraint B go to infinity while holding B/N constant as that in Whittle (Whittle, 1988) and others (Weber & Weiss, 1990). For abuse of notation, let the number of arms be ρN and the resource constraint be ρW in the asymptotic regime with $\rho \to \infty$. In other words, we consider N different classes of arms with each class containing ρ arms. Let $R^\pi(\rho B, \rho N)$ denote the expected total reward of the original problem (2) under an arbitrary policy π for such a system. To show the asymptotical optimality of RMI, according to Xiong et al. (2022a), it is sufficient to show that

$$\lim_{\eta\to\infty} \frac{1}{\rho}\Big(R^{\pi^{opt}}(\eta W, \eta N) - R^{\pi^\star}(\eta W, \eta N)\Big) = 0, \tag{26}$$

with π^* being the RMI index policy and π^{opt} is the optimal one. The equation in (26) indicates that as the number of per-class arms goes to infinity, the average gap between the performance achieved by our RMI index policy $\pi^\star$ and the optimal policy π^{opt} of the original problem in (2) tends to be zero.

The proof comes from both side of (26). First, we have that for any policy $\pi^\star$, the left-hand side of (26) is non-negative since any policy cannot achieve a higher reward compared with that of the optimal policy π^{opt}, i.e.,

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

Hence we need to show the other direction of (26) that for induced RMI index policy $\pi^\star$, the following holds

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

Let $B_n(s;h)$ be the number of class n arms in state s at time h and $D_n(s,a;h)$ be the number of class n arms in state s at time h that are being served with action $a \in \mathcal{A}\setminus\{0\}$. Similar to Xiong et al. (2022a), by using induction, we show that $\rho \to \infty$, the following event occurs almost surely,

$$B_n(s;h)/\rho \to P_n(s;h),$$
$$D_n(s,a;h)/\rho \to P_n(s;h)\xi_n^\star(s,a;h)$$

respectively. This leads to the fact that

$$\lim_{\rho\to\infty} \frac{1}{\rho} R^{\pi^\star}(\rho B, \rho N) = \sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{(s,a)} \mu_n^\star(s,a;h) r_n(s,a),$$

which converges the optimal solution to the LP in Lemma A.2, which is an upper bound of $\lim_{\rho\to\infty} \frac{1}{\rho} R^{\pi^{opt}}(\rho B, \rho N)$ due to Lemma 5.2. This completes the proof.

B. Proofs in Section 5

In this section, we provide proofs for Lemmas in Section 5.

B.1. Proof of Lemma 5.2

There exists a set of occupancy measures $\mu_\pi^* := \{\mu_n^*(s,a;h), \forall n \in [N], s \in \mathcal{S}, a \in \mathcal{A}, h \in [h]\}$ under policy π that optimally solve (2) with a relaxed constraint such that $\mathbb{E}_\pi\left[\sum_{n=1}^N A_n^{t,h}\right] \leq B, h \in [H], t \in [T]$. In addition, $\sum_{t=1}^T \langle \mu_\pi^*, r^t \rangle$ is no less than $R(\mathbf{s}_1, T, \pi^{opt})$, where μ_π^* is the stacked vector of all occupancy measures $\mu_n^*(s,a;h)$ and r^t is the stacked vector of all rewards $r_n^t(s,a)$ for all n, s, a. According to Lemma A.2, if the "hard" activation constraint in (2) is relaxed to an averaged one as $\mathbb{E}_\pi\left[\sum_{n=1}^N A_n^{t,h}\right] \leq B, h \in [H], t \in [T]$, there is an equivalent LP expressed as in (22)-(25). To prove Lemme 5.2, it is sufficient to show that the relaxed problem achieves no less average reward than the original problem in (2). The proof is straightforward since the constraints in the relaxed problem expand the feasible region of the original problem in (2). Denote the feasible region of the original problem as

$$\Gamma := \left\{ A_n^{t,h}, \forall h, t \middle| \sum_{n=1}^N A_n^{t,h} \leq B, \forall h \right\},$$

and the feasible region of the relaxed problem as

$$\Gamma' := \left\{ A_n^{t,h}, \forall h, t \middle| \mathbb{E}_\pi \left[\sum_{n=1}^N A_n^{t,h} \leq B, \forall h \right] \right\}.$$

It is clear that the relaxed problem expands the feasible region of the original problem in (2), i.e., $\Gamma \subseteq \Gamma'$. Therefore, the relaxed problem achieves an objective value no less than that of the original problem in (2) because the original optimal solution is also inside the relaxed feasibility set (Altman, 1999), i.e., Γ'. Denote the optimal occupancy measures of LP in (22)-(25) as $\mu_\pi^* := \{\mu_n^*(s,a;h), \forall n \in [N], s \in \mathcal{S}, a \in \mathcal{A}, h \in [h]\}$ under a stationary policy π induced by $\{\mu_n^*(s,a;h), \forall n \in [N], s \in \mathcal{S}, a \in \mathcal{A}, h \in [h]\}$, and hence the maximum reward achieved for the LP in (22)-(25) is equal to $\sum_{t=1}^T \langle \mu_\pi, r^t \rangle$. Therefore, it follows that $\sum_{t=1}^T \langle \mu_\pi, r^t \rangle \geq R(\mathbf{s}_1, T, \pi^{opt})$, which completes the proof.

B.2. Proof of Lemma 5.3

According to the definition of regret in (3), we have the following inequality

$$\begin{aligned}
\Delta(T) &= R(\mathbf{s}_1, T, \pi^{opt}) - R(\mathbf{s}_1, T, \{\pi^t, \forall t\}) \\
&= \underbrace{R(\mathbf{s}_1, T, \pi^{opt}) - R(\mathbf{s}_1, T, \pi^{opt}, \{\hat{r}_n^t, \forall n, t\})}_{Term_0 \leq 0} \\
&\quad + \underbrace{R(\mathbf{s}_1, T, \pi^{opt}, \{\hat{r}_n^t, \forall n, t\}) - R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{\hat{r}_n^t, \forall n \in [N], t \in [T]\})}_{Term_1} \\
&\quad + \underbrace{R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{\hat{r}_n^t, \forall n \in [N], t \in [T]\}) - R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{r_n^t, \forall n \in [N], t \in [T]\})}_{Term_2} \\
&\quad + \underbrace{R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{r_n^t, \forall n \in [N], t \in [T]\}) - R(\mathbf{s}_1, T, \{\pi^t, \forall t\})}_{Term_3} \\
&\leq \underbrace{R(\mathbf{s}_1, T, \pi^{\mu^*}, \{\hat{r}_n^t, \forall n, t\}) - R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{\hat{r}_n^t, \forall n \in [N], t \in [T]\})}_{Term_1} \\
&\quad + \underbrace{R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{\hat{r}_n^t, \forall n \in [N], t \in [T]\}) - R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{r_n^t, \forall n \in [N], t \in [T]\})}_{Term_2} \\
&\quad + \underbrace{R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{r_n^t, \forall n \in [N], t \in [T]\}) - R(\mathbf{s}_1, T, \{\pi^t, \forall t\})}_{Term_3}.
\end{aligned}$$

The first inequality holds due to two reasons. First, $Term_0$ is non-positive as $\{\hat{r}_n^t, \forall n, t\}$ is an overestimation of the true $r_n^t, \forall n, t$. Second, $R(\mathbf{s}_1, T, \pi^{\mu^*}, \{\hat{r}_n^t, \forall n, t\})$ is no less than $R(\mathbf{s}_1, T, \pi^{opt}, \{\hat{r}_n^t, \forall n, t\})$ according to Lemma 5.2.

B.3. Proof of Lemma 5.4

Denote the true MDP as $M := \{M_n, \forall n\}$. Hence, we have the following event occurs with high probability when the true transition kernel is inside the confidence ball defined in (5), i.e.,

$$\mathcal{E}_p^t := \{\forall (s,a), n, |P_n(s'|s,a) - \hat{P}_n^t(s'|s,a)| < \delta_n^t(s,a)\}. \tag{27}$$

The cumulative probability that the failure events occur is bounded as follows.

Lemma B.1. *With probability at least $1-2\epsilon$, we have the true transition P is within the confidence set $\mathcal{P}^t := \{\mathcal{P}_n^t, \forall n \in [N]\}$, i.e., event $\mathcal{E}_p^t$ occurs, when $\delta_n^t(s,a) = \sqrt{\frac{1}{2C_n^{t-1}(s,a)} \log\left(\frac{4|\mathcal{S}||\mathcal{A}|N(t-1)H}{\epsilon}\right)}$.*

Proof. By Chernoff-Hoeffding inequality (Hoeffding, 1994), we have

$$\mathbb{P}\big(|P_n(s'|s,a) - \hat{P}_n^t(s'|s,a)| > \delta_n^t(s,a)\big) \le \frac{2\epsilon}{|\mathcal{S}||\mathcal{A}|N(t-1)H}.$$

Using union bound on all states, actions and users, we have

$$\begin{aligned}\mathbb{P}(\mathbb{1}_{\{\mathcal{E}_p^t\}}) &\ge 1 - \sum_{n=1}^{N}\sum_{(s,a)} \mathbb{P}\big(|P_n(s'|s,a) - \hat{P}_n^t(s'|s,a)| > \delta_n^t(s,a)\big) \\ &\ge 1 - \frac{2\epsilon}{(t-1)H}.\end{aligned}$$

□

Next, we have the inequality related with OMD update in Algorithm 1, similar to Jin et al. (2020) as follows.

Lemma B.2. *The OMD update with $z_n^1(s,a,s') = \frac{1}{|\mathcal{S}|^2|\mathcal{A}|}$ for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, and $z_n^{t+1} = \arg\max_{z_n \in \Delta^t} \eta\langle z_n, \hat{r}_t\rangle - D(z_n \,\|\, z_n^t)$ ensures*

$$\sum_{t=1}^{T}\sum_{n=1}^{N}\langle z_n - z_n^t, \hat{r}_n^t\rangle \le \frac{NH\ln(|\mathcal{S}|^2|\mathcal{A}|)}{\eta} + \eta NH(T+1).$$

for any $z \in \cap_t \Delta^t$, as long as $0 \le \eta\hat{r}_n^t(s,a) \le 1$ for all t, n, s, a.

Proof. Based on (19), we have $\tilde{z}^t$ as

$$\tilde{z}_n^t(s,a,s';h) = z_n^{t-1}(s,a,s';h)\exp\left(\eta\hat{r}_n^t(s,a)\right),$$

and $z_n^t = \arg\min_{z\in\Delta^t} D(z \,\|\, \tilde{z}_n^t)$. According to (11), we have

$$\begin{aligned}&D(z_n \,\|\, z_n^t) - D(z_n \,\|\, \tilde{z}_n^{t+1}) + D(z_n^t \,\|\, \tilde{z}_n^{t+1}) \\ &= \sum_{s,a,s',h} z_n(s,a,s',h)\ln\frac{z_n(s,a,s',h)}{z_n^t(s,a,s',h)} - z_n(s,a,s',h) + z_n^t(s,a,s',h) \\ &\quad - \sum_{s,a,s',h} z_n(s,a,s',h)\ln\frac{z_n(s,a,s',h)}{\tilde{z}_n^{t+1}(s,a,s',h)} + z_n(s,a,s',h) - \tilde{z}_n^{t+1}(s,a,s',h) \\ &\quad + \sum_{s,a,s',h} z_n^t(s,a,s',h)\ln\frac{z_n^t(s,a,s',h)}{\tilde{z}_n^{t+1}(s,a,s',h)} - z_n^t(s,a,s',h) + \tilde{z}_n^{t+1}(s,a,s',h) \\ &= \sum_{s,a,s',h} z_n(s,a,s',h)\ln\frac{\tilde{z}_n^{t+1}(s,a,s',h)}{z_n^t(s,a,s',h)} - \sum_{s,a,s',h} z_n^t(s,a,s',h)\ln\frac{\tilde{z}_n^{t+1}(s,a,s',h)}{z_n^t(s,a,s',h)} \\ &= \eta\langle z_n - z_n^t, \hat{r}_n^t\rangle,\end{aligned} \tag{28}$$

where the last inequality is due to (19). Furthermore, according to (20) generalized Pythagorean theorem, we have $D(z_n \parallel z_n^{t+1}) \leq D(z_n \parallel \tilde{z}_n^{t+1})$. Hence, the following inequality holds

$$\begin{aligned}
\eta \sum_{t=1}^{T}\sum_{n=1}^{N}\langle z_n - z_n^t, \hat{r}_n^t\rangle &\leq \sum_{t=1}^{T}\sum_{n=1}^{N} D(z_n \parallel z_n^t) - D(z_n \parallel z_n^{t+1}) + D(z_n^t \parallel \tilde{z}_n^{t+1}) \\
&= \sum_{n=1}^{N} D(z_n \parallel z_n^1) - D(z_n \parallel z_n^{T+1}) + \sum_{t=1}^{T}\sum_{n=1}^{N} D(z_n^t \parallel \tilde{z}_n^{t+1}) \\
&= \underbrace{\sum_{n=1}^{N}\sum_{s,a,s',h} z_n(s,a,s',h) \ln \frac{z_n^{T+1}(s,a,s',h)}{z_n^1(s,a,s',h)}}_{(a_1)} \\
&\quad + \underbrace{\sum_{n=1}^{N}\sum_{s,a,s',h} z_n^1(s,a,s',h) - z_n^{T+1}(s,a,s',h)}_{(a_2)} + \underbrace{\sum_{t=1}^{T}\sum_{n=1}^{N} D(z_n^t \parallel \tilde{z}_n^{t+1})}_{(a_3)}. \qquad (29)
\end{aligned}$$

Next, we bound each term in (29). (a_1) is bounded as

$$(a_1) \leq \sum_{n=1}^{N}\sum_{s,a,s',h} z_n(s,a,s',h) \ln(|\mathcal{S}|^2|\mathcal{A}|) \leq NH \ln(|\mathcal{S}|^2|\mathcal{A}|),$$

due to the initialization of $z_n^1(s,a,s',h), \forall s,a,h$. Similarly, (a_2) is bounded as $(a_2) \leq NH$. (a_3) is bounded as follows,

$$\begin{aligned}
D(z_n^t \parallel \tilde{z}_n^{t+1}) &= \sum_{s,a,s',h} z_n^t(s,a,s',h) \ln \frac{z_n(s,a,s',h)}{\tilde{z}_n^{t+1}(s,a,s',h)} - z_n(s,a,s',h) + \tilde{z}_n^{t+1}(s,a,s',h) \\
&= \sum_{s,a,s',h} -\eta \hat{r}_n^t(s,a) z_n^t(s,a,s',h) - z_n^t(s,a,s',h) + z_n^t(s,a,s',h) \exp(\eta \hat{r}_n^t(s,a)) \\
&\leq \eta^2 \sum_{s,a,s',h} z_n^t(s,a,s',h) \leq H\eta^2,
\end{aligned}$$

where the first inequality is due to the fact $e^z - 1 - z \leq z^2$ for all $z \in [0,1]$, and the second inequality is due to the property of occupancy measure. Substituting $(a_1) - (a_3)$ back to (29), we have

$$\sum_{t=1}^{T}\sum_{n=1}^{N}\langle z_n - z_n^t, \hat{r}_n^t\rangle \leq \frac{NH \ln(|\mathcal{S}|^2|\mathcal{A}|)}{\eta} + \eta NH(T+1).$$

This completes the proof. □

According to the definition of π^{μ^*} and $\tilde{\pi}^t$ and the fact that they are working on the problem with relaxed activation constraint, we can transform $Term_1$ into the following form:

$$\begin{aligned}
Term_1 &= R(\mathbf{s}_1, T, \pi^{\mu^*}, \{\hat{r}_n^t, \forall n, t\}) - R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{\hat{r}_n^t, \forall n \in [N], t \in [T]\}) \\
&= \sum_{t=1}^{T}\sum_{n=1}^{N}\langle \mu_n^*, \hat{r}_n^t\rangle - \sum_{t=1}^{T}\sum_{n=1}^{N}\langle z_n^t, \hat{r}_n^t\rangle \\
&= \sum_{t=1}^{T}\sum_{n=1}^{N}\langle \mu_n^* - z_n^t, \hat{r}_n^t\rangle.
\end{aligned}$$

Since with probability $1 - 2\epsilon$, we have $\mu_n^* \in \cap_t \mathcal{Z}^t$ according to Lemma B.1. Hence, Lemma 5.4 directly follows Lemma B.2.

B.4. Proof of Lemma 5.5

According to the definition, we can rewrite $Term_2$ as

$$\begin{aligned} Term_2 &= R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{\hat{r}_n^t, \forall n \in [N], t \in [T]\}) - R(\mathbf{s}_1, T, \{\tilde{\pi}^t, \forall t\}, \{r_n^t, \forall n \in [N], t \in [T]\}) \\ &= \sum_{t=1}^{T} \sum_{n=1}^{N} \langle z_n^t, \hat{r}_n^t - r_n^t \rangle. \end{aligned}$$

The key is to characterize the gap between $\hat{r}_n^t - r_n^t$. We first present the widely used Hoeffding inequality in the following lemma.

Lemma B.3 (Hoeffding inequality (Bercu et al., 2015))**.** *Let $X_1, X_2, \ldots$ be independent random variables with $b \leq |X_i| \leq c$ for all i. Define $S_n = X_1 + \ldots + X_n$. Then for all $\epsilon > 0$*

$$Pr\left(S_n - \mathbb{E}[S_n] > \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2n(c-b)^2}\right).$$

Notice that $\langle z_n^t, \hat{r}_n^t \rangle \leq HN$ due to the overestimation of $\hat{r}_n^t$ in (14), according the Hoeffding inequality in Lemma B.3, we have that with probablity at least $1 - \epsilon$, such taht

$$\sum_{t=1}^{T} \sum_{n=1}^{N} \langle z_n^t, \hat{r}_n^t - \mathbb{E}[\hat{r}_n^t] \rangle \leq HN\sqrt{TN \ln 1/\epsilon}.$$

Thus, we have at least probability $1 - 3\epsilon$ that

$$Term_2 \leq \sum_{t=1}^{T} \sum_{n=1}^{N} \langle z_n^t, \mathbb{E}[\hat{r}_n^t] - r_n^t \rangle + HN\sqrt{TN \ln 1/\epsilon} \leq \sum_{t=1}^{T} \sum_{n=1}^{N} \langle z_n^t, \delta_n^t \rangle + HN\sqrt{TN \ln 1/\epsilon}, \tag{30}$$

where the inequality holds due to the definition of $\hat{r}_n^t(s,a)$ and the fact that $\frac{r_n(s,a)\mathbb{1}(\exists h, S_n^t(h)=s, A_n^t(h)=a)}{\max\{c_n^t(s,a),1\}/H}$ is an unbiased estimator of $r_n^t(s,a)$.

Next, we bound $\sum_{t=1}^{T} \sum_{n=1}^{N} \langle z_n^t, \delta_n^t \rangle$ in the following lemma.

Lemma B.4. *The following inequality holds*

$$\sum_{t=1}^{T} \sum_{n=1}^{N} \langle z_n^t, \delta_n^t \rangle \leq \sqrt{2}H\sqrt{\ln \frac{4SANT}{\epsilon}} \cdot \sqrt{NT}.$$

Proof. The proof goes as follows.

$$\begin{aligned} \sum_{t=1}^{T} \sum_{n=1}^{N} \langle z_n^t, \delta_n^t \rangle &\overset{(a)}{\leq} H \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{s,a} \delta_n^t(s,a) \\ &\leq H \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{s,a} \sqrt{\frac{1}{2C_n^t(s,a)} \ln \frac{4SANT}{\epsilon}} \\ &\leq H\sqrt{\ln \frac{4SANT}{\epsilon}} \sum_{n=1}^{N} \sum_{(s,a)} \sqrt{C_n^T(s,a)} \\ &\overset{(b)}{\leq} \sqrt{2}H\sqrt{\ln \frac{4SANT}{\epsilon}} \sum_{n=1}^{N} \sqrt{\sum_{(s,a)} C_n^T(s,a)} \\ &\overset{(c)}{\leq} \sqrt{2}H\sqrt{\ln \frac{4SANT}{\epsilon}} \cdot \sqrt{NT}, \end{aligned}$$

where (a) follows since z_n^t is a probability measure, (b) follows Cauchy-Schwartz inequality and (c) uses the fact that $\sum_{n=1}^{N} \sum_{(s,a)} C_n^T(s,a) \leq NT$. This completes the proof. □

Bound on $Term_2$. Combining the results in (30) and Lemma B.4, we can bound $Term_2$ as

$$Term_2 \leq \sqrt{2}H\sqrt{\ln\frac{4SANT}{\epsilon}}\cdot\sqrt{NT} + HN\sqrt{TN\ln 1/\epsilon}.$$

B.5. Proof of Lemma 5.6

According to Lemma B.1, with probability at least $1-2\epsilon$, the true transition kernel $\mathcal{P}$ is inside the confidence ball as defined in (27). Due to the optimism of the confidence ball, in each episode, we have that $\tilde{\pi}^t$ under an optimisitic MDP with transition $\tilde{\mathcal{P}}$ achieves no worse performance than π^t under the true MDP with transition $\mathcal{P}$. Hence, we decompose the regret in $Term_3$ into episodic regrets. For simplicity, we denote $c_n^t(s,a) := \sum_{h=1}^{H}\mathbb{1}(S_n^{t,h}=s, A_n^{t,h}=a)$ as the state-action counts for (s,a) in episode t, and $\gamma^{t,*}$ as the average reward achieved per decision epoch by policy $\tilde{\pi}^t$. Then, we define the regret during episode t as follows,

$$\Delta_t := H\gamma^{t,*} - \sum_{(s,a)}\sum_{n} c_n^t(s,a)r_n^t(s,a). \tag{31}$$

The relation between the total regret in $Term_3$ and the episodic regrets Δ_k is as follows.

Lemma B.5. *The regret in $Term_3$ is upper-bounded by*

$$Term_3 \leq \sum_{t=1}^{T}\Delta_t + \sqrt{\frac{1}{4}T\log\Big(\frac{|\mathcal{S}||\mathcal{A}|NT}{\epsilon}\Big)}, \tag{32}$$

with probability at least $1-\epsilon$.

Proof. Using Azuma-Hoeffding's inequality, we have

$$\begin{aligned}\mathbb{P}\Bigg(R(\mathbf{s}_1,T,\{\pi^t,\forall t\}) \leq \sum_{t=1}^{T}\sum_{n}\sum_{(s,a)} c_n^t(s,a)r_n^t(s,a) - \sqrt{\frac{1}{4}T\log\Big(\frac{|\mathcal{S}||\mathcal{A}|NT}{\epsilon}\Big)}\Bigg)\\ \leq \left(\frac{\epsilon}{|\mathcal{S}||\mathcal{A}|NT}\right)^{1/2}.\end{aligned} \tag{33}$$

Therefore,

$$\begin{aligned}\Delta(T) &= \sum_{t=1}^{T} H\gamma^{t,*} - R(\mathbf{s}_1,T,\{\pi^t,\forall t\})\\ &\leq \sum_{t=1}^{T} H\gamma^{t,*} - \sum_{t=1}^{T}\sum_{n}\sum_{(s,a)} c_n^t(s,a)r_n^t(s,a) + \sqrt{\frac{1}{4}T\log\Big(\frac{|\mathcal{S}||\mathcal{A}|NT}{\epsilon}\Big)}\\ &= \sum_{k=1}^{K}\Delta_k + \sqrt{\frac{1}{4}T\log(\frac{|\mathcal{S}||\mathcal{A}|NT}{\epsilon})}.\end{aligned}$$

This completes the proof. □

The following lemma characterizes the regret under scenario where the true transition $\mathcal{P}$ is outside of the confidence ball, which occurs at a small probability at most 2ϵ.

Lemma B.6. *We have*

$$\sum_{t=1}^{T}\Delta_t\mathbb{1}(\mathcal{P}\notin\mathcal{E}_p^t) \leq B\sqrt{TH}.$$

Proof. By Lemma B.1, we have

$$\begin{aligned}\sum_{t=1}^{T}\Delta_t\mathbb{1}(\mathcal{P}\notin\mathcal{E}_p^t) &\le \sum_{t=1}^{T}\sum_{n}\sum_{(s,a)}c_n^t(s,a)\mathbb{1}(\mathcal{P}\notin\mathcal{E}_p^t)\\ &\le \sum_{t=1}^{T}HB\mathbb{1}(\mathcal{P}\notin\mathcal{E}_p^k)\\ &\le B\sum_{i=1}^{TH}i\mathbb{1}(\mathcal{P}\notin\mathcal{E}_p^t)\le B\sqrt{TH}.\end{aligned}$$

□

We next bound the regrets Δ_k when the true transition lies in the confidence ball, i.e., event $\mathcal{E}_p^t$ occurs.

Lemma B.7. *Under event $\mathcal{E}_p^t$, we have*

$$\begin{aligned}\Delta_t\mathbb{1}(\mathcal{P}\in\mathcal{E}_p^t) \le &\sum_{(s,a)}\sum_{n}c_n^t(s,a)(\gamma^{t,*}/B - r_n^t(s,a))\\ &+\sum_{(s,a)}\sum_{n}c_n^t(s,a)\kappa_1,\end{aligned}\tag{34}$$

where $\kappa_1>0$ is as in Lemma B.8 and $\tilde{\gamma}^t$ is the average reward achieved by an optimistic policy induced from the confidence ball (i.e., $\mathcal{E}_p^t$) when the rewards is shifted by $2\delta_n^t(s,a),\forall s,a,n$, i.e., $\tilde{r}_n^t(s,a):=r_n^t(s,a)+2\delta_n^t(s,a)$.

Proof. This result follows directly from the definition of episodic regret in (31) and the fact that the total activated arms count should be less than HB per episode. □

Lemma B.8. *With probability at least $1-2\epsilon$, we have $\tilde{\gamma}^t\ge\gamma^{t,*}-\kappa_1$, where $\kappa_1=\mathcal{O}(\frac{c_1}{\sqrt{tH}})$ with $0<c_1<1$.*

Proof. We denote by S_h^t and $S_h^{t,*}$ the arms that are activated by an optimistic policy and $\tilde{\pi}^t$ under true reward and at the h-th decision epoch of episode t. For abuse of notation, we use $\tilde{r}_n^t(h)$ to denote the reward of a specific state-action pair (s,a) visited at h for arm n. Due to optimism, we have

$$\sum_{n\in S_h^t\setminus S_h^{t,*}}\tilde{r}_n^t(h)\ge\sum_{n\in S_h^{t,*}\setminus S_h^t}\tilde{r}_n^t(h).$$

When conditioned on the event $\mathcal{E}_p^t$ simultaneously, we have

$$\begin{aligned}\sum_{n\in S_h^k\setminus S_h^{k,*}}r_n^t(h)+2\delta_n^k &\ge \sum_{n\in S_h^k\setminus S_h^{k,*}}\tilde{r}_n^t(h)\ge\sum_{n\in S_h^{k,*}\setminus S_h^k}\tilde{r}_n^t(h)\\ &\ge\sum_{n\in S_h^{k,*}\setminus S_h^k}r_n^t(h).\end{aligned}$$

Hence, we have

$$\sum_{n\in S_h^{t,(*)}\setminus S_h^t}r_n^t(h)-\sum_{n\in S_h^t\setminus S_h^{t,*}}r_n^t(h)\le\sum_{n\in S_h^t\setminus S_h^{t,*}}2\delta_n^t\le 2B\delta^t,$$

where $\delta_n^t=\delta^t,\forall n$ and is given in Lemma B.1. Therefore, we have $\gamma^{t,*}-\tilde{\gamma}^t\le 2B\delta=\mathcal{O}\Big(\frac{2B}{\sqrt{tH}}\Big)$. □

Upon combining the results obtained in Lemma B.7 and Lemma B.8, we obtain the following.

Lemma B.9. *We have*

$$\sum_{t=1}^{T} \Delta_t \mathbb{1}(\mathcal{P} \in \mathcal{E}_p^t) \leq \Big(\sqrt{2\log\Big(\frac{4|\mathcal{S}||\mathcal{A}|NTH}{\epsilon}\Big)} + 2B\Big)\sqrt{|\mathcal{S}||\mathcal{A}|NTH}.$$

Proof. From Lemma B.7 and Lemma B.8, we can rewrite the summation over Δ_t as follows:

$$\sum_{t=1}^{T} \Delta_k \mathbb{1}(\mathcal{P} \in \mathcal{E}_p^t) \leq \underbrace{\sum_{t=1}^{T}\sum_{(s,a)}\sum_{n} c_n^t(s,a)(\gamma^{t,*}/B - \tilde{r}_n^t(s,a))}_{(I)}$$
$$+ \underbrace{\sum_{t=1}^{T}\sum_{(s,a)}\sum_{n} c_n^t(s,a)\Big(\tilde{r}_n^t(s,a) - r_n^t(s,a) + \frac{1}{\sqrt{tH}}\Big)}_{(II)}.$$

It is clear that (I) is non-positive. Conditioned on good events, it is clear that $\tilde{r}_n^t(s,a) - r_n^t(s,a) \leq 2\delta_n^t(s,a)$. Therefore, we have

$$\begin{aligned}\sum_{t=1}^{T} \Delta_t \mathbb{1}(\mathcal{P} \in \mathcal{E}_p^k) &\leq \left(\sqrt{2\log\Big(\frac{4|\mathcal{S}||\mathcal{A}|NTH}{\epsilon}\Big)} + 2B\right)\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{(s,a)}\frac{c_n^t(s,a)}{\sqrt{C_n^t(s,a)}} \\ &\leq \left(\sqrt{2\log\Big(\frac{4|\mathcal{S}||\mathcal{A}|NTH}{\epsilon}\Big)} + 2B\right)\sqrt{|\mathcal{S}||\mathcal{A}|NTH},\end{aligned}$$

where last inequality is due to Jensen's inequality and Lemma B.10.

Lemma B.10. *For any sequence of numbers $w_1, w_2, ..., w_T$ with $0 \leq w_k$, define $W_k := \sum_{i=1}^{k} w_i$,*

$$\sum_{k=1}^{T} \frac{w_k}{\sqrt{W_k}} \leq (\sqrt{2}+1)\sqrt{W_T}.$$

Proof. The proof follows by induction. When $t = 1$, it is true as $1 \leq \sqrt{2}+1$. Assume for all $k \leq t-1$, the inequality holds, then we have the following:

$$\begin{aligned}&\sum_{k=1}^{T} \frac{w_k}{\sqrt{W_k}} = \sum_{k=1}^{T-1} \frac{w_k}{\sqrt{W_k}} + \frac{w_T}{\sqrt{W_T}} \\ &\leq (\sqrt{2}+1)\sqrt{W_{T-1}} + \frac{w_T}{\sqrt{W_T}} \\ &= \sqrt{(\sqrt{2}+1)^2 W_{T-1} + 2(\sqrt{2}+1)w_T\sqrt{\frac{W_{T-1}}{W_T}} + \frac{{w_T}^2}{W_T}} \\ &\leq \sqrt{(\sqrt{2}+1)^2 W_{T-1} + 2(\sqrt{2}+1)w_T\sqrt{\frac{W_{T-1}}{W_{T-1}}} + \frac{w_T W_T}{W_T}} \\ &= \sqrt{(\sqrt{2}+1)^2 W_{T-1} + (2(\sqrt{2}+1)+1)w_T} \\ &= (\sqrt{2}+1)\sqrt{(W_{T-1}+w_T)} \\ &= (\sqrt{2}+1)\sqrt{W_T}.\end{aligned}$$

□

□

C. Details of the Numerical Case-Study

Continuous Positive Airway Pressure Therapy (CPAP). The CPAP (Kang et al., 2013; Herlihy et al., 2023; Li & Varakantham, 2022; Wang et al., 2024) is a highly effective treatment when it is used consistently during 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 to a two-state system with the clinical adherence criteria. To elaborate, three distinct states are defined to characterize adherence levels: low, intermediate, and acceptable. Patients are categorized into two clusters: "Adherence" and "Non-Adherence." Those within the "Adherence" cluster exhibit a greater likelihood of maintaining an acceptable level of adherence. In particular, the state space is $1, 2, 3$, which represents low, intermediate and acceptable adherence level respectively. The difference between these two groups reflects on their transition probabilities, as in Figures. 4- 5. Generally speaking, the first group has a higher probability of staying in a good adherence level. From each group, we construct 10 arms, whose transition probability matrices are generated by adding a small noise to the original one. Actions such as text patients/ making a call/ visit in person will cause a 5% to 50% increase in adherence level. The budget is set to $B = 10$. The objective is to maximize the total adherence level.

In standard CPAP, which is used for stochastic RMAB setting, the reward is set as the 1 for state "low adherence", 2 for state "intermediate adherence", and 3 for state "high adherence". Based on this, we randomly generate a sequence of coefficients to increase or decrease the reward for each episode. In our setting, we consider an MDP with 50 episode, and each episode contains 50 time steps. The control coefficient for episode i is $0.5 + (i-1)/49$. The whole experiment runs in 1000 Monte Carlo independent rounds.

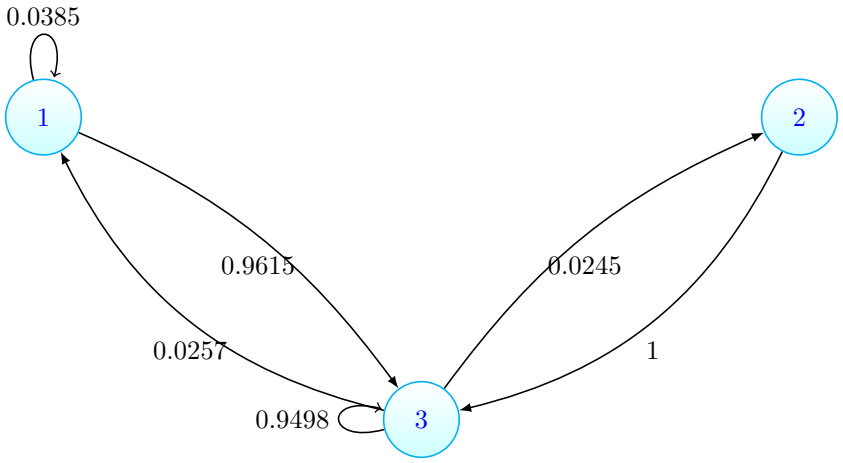


Figure 4: Transition diagram for CPAP Cluster 1

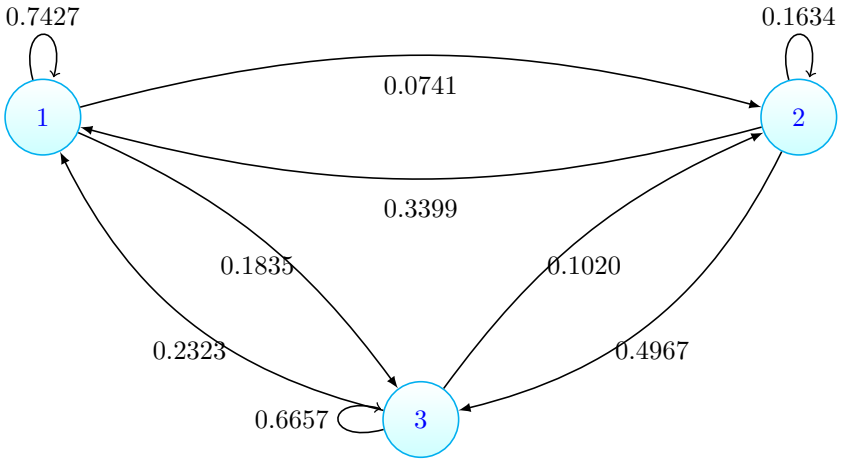


Figure 5: Transition diagram for CPAP Cluster 2

A Deadline Scheduling Problem. We consider the deadline scheduling problem for the scheduling of electrical vehicle charging stations. A charging station (agent) has total N charging spots (arms) and can charge M vehicles in each round. The charging station obtains a reward for each unit of electricity that it provides to a vehicle and receives a penalty (negative reward) when a vehicle is not fully charged. The goal of the station is to maximize its net reward. We use exactly the same setting as in (Xiong et al. 2022a) for our experiment. More specifically, the state of an arm is denoted by a pair of integers $(D; B)$, where B is the amount of electricity that the vehicle still needs and D is the time until the vehicle leaves the station. When a charging spot is available, its state is $(0; 0)$. B and D are upper-bounded by 9 and 12, respectively. Hence, the size of state space is 109 for each arm. The state transition is given by

$$S_i(t+1) = \begin{cases} (D_i(t)-1, B_i(t)-a_i(t)), & \text{if } D_i(t) > 1, \\ (D, B), & \text{with prob. 0.7 if } D_i(t) \leq 1, \end{cases}$$

where (D, B) is a random state when a new vehicle arrives at the charging spot i. Specifically, $a_i(t) = 0$ means being passive and $a_i(t) = 1$ means being active. There are total $N = 100$ charging spots and a maximum $M = 30$ can be served at each time.

The agent receives a base reward at each time from arm i according to

$$r_i(t) = \begin{cases} (1-0.5)a_i(t), & \text{if } B_i(t) > 0, D_i(t) > 1, \\ (1-0.5)a_i(t) - 0.2(B_i(t)-a_i(t))^2, & \\ & \text{if } B_i(t) > 0, D_i(t) = 1, \\ 0, & \text{Otherwise.} \end{cases}$$

Based on this, we randomly generate a sequence of coefficients to increase or decrease the reward for each episode. In our setting, we consider an MDP with 100 episode, and each episode contains 100 time steps. The control coefficient for episode i is $0.5 + (i-1)/99$. The whole experiment runs in 1000 Monte Carlo independent rounds.