Planning and Learning in Risk-Aware Restless Multi-Arm Bandits

Nima Akbarzadeh[1,2] **Yossiri Adulyasak**[1] **Erick Delage**[1,2]
[1]GERAD & HEC Montréal, [2]Mila - Quebec AI Institute

Abstract

In restless multi-arm bandits, a central agent is tasked with optimally distributing limited resources across several bandits (arms), with each arm being a Markov decision process. In this work, we generalize the traditional restless multi-arm bandit problem with a risk-neutral objective by incorporating risk-awareness. We establish indexability conditions for the case of a risk-aware objective and provide a solution based on Whittle index. In addition, we address the learning problem when the true transition probabilities are unknown by proposing a Thompson sampling approach and show that it achieves bounded regret that scales sublinearly with the number of episodes and quadratically with the number of arms. The efficacy of our method in reducing risk exposure in restless multi-arm bandits is illustrated through a set of numerical experiments in the contexts of machine replacement and patient scheduling applications under both planning and learning setups.

1 INTRODUCTION

The restless multi-arm bandit (RMAB) problem is a class of sequential stochastic control problems for dynamic decision-making under uncertainty. In RMABs, a central agent confronts the challenge of allocating limited resources over time among competing options, which we refer to as *arms*, each characterized by a Markov decision process (MDP). Such a framework has numerous applications in scheduling problems that appear in machine maintenance Glazebrook et al. (2005); Akbarzadeh and Mahajan (2019b), healthcare Deo et al. (2013), finance Glazebrook et al. (2013), power management in smart grids Wang et al. (2014); Abad and Iyengar (2016), opportunistic scheduling in networks Liu and Zhao (2010); Nino-Mora (2009); Ouyang et al. (2015); Borkar et al. (2018); Wang et al. (2019), and operator allocation in multi-robot systems Dahiya et al. (2022).

The studies of RMABs have primarily focused on risk-neutral/expected value objectives under reward maximization or cost minimization. The assumption of risk neutrality, however, is not always suitable in practice, as the level of risk is an essential aspect to consider in real-world applications Xu et al. (2021); Mate et al. (2021). Such applications encompass various domains, including preventive maintenance Amiri et al. (2018), surgery and medical scheduling in healthcare He et al. (2019); Najjarbashi and Lim (2019), financial portfolio management, and production lot-sizing Long et al. (2023), where risk-neutral solutions can be impracticable and lead to risky outcomes. In such circumstances, risk-aware policies accounting for potential risks offer resilient solutions, designed to reduce adverse effects of uncertain outcomes, guaranteeing the allocation policy remains effective even in adverse conditions Rausand (2013).

In the traditional RMAB with the risk-neutral objective, the key challenge that prevents applying traditional stochastic control methods is the curse of dimensionality due to the state space. As the number of arms increases, the computational complexity of identifying the optimal policy grows exponentially. This poses a formidable obstacle in real-world implementations. Whittle (1988) introduced a scalable and computationally tractable index policy as a heuristic for RMABs. The Whittle index acts as a priority index that highlights the urgency of selecting an arm. In what follows, we refer to this policy as the *Whittle index policy* Niño-Mora (2023).

While the Whittle index policy requires a technical condition, known as *indexability*, to be satisfied, most of the studies in the literature of RMABs implement Whittle index policy. In these problems, either the problem structure is such that the indexability is satisfied Jacko and Villar (2012); Borkar et al. (2018); Yu

Proceedings of the 28^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

et al. (2018); Wang et al. (2019), or sufficient conditions under which the problem is indexable are specified Glazebrook et al. (2006); Niño-Mora (2007); Akbarzadeh and Mahajan (2022). The literature suggests that Whittle index policy works well in practice Glazebrook et al. (2006); Niño-Mora (2007); Akbarzadeh and Mahajan (2022); Avrachenkov et al. (2013); Wang et al. (2020).

Risk-aware objectives have been studied extensively for MDPs and RL Le Tallec (2007); Osogami (2012); Bäuerle and Rieder (2014); Chow et al. (2015); Mannor et al. (2016); Jaimungal et al. (2022); Xu et al. (2023). To the best of our knowledge, only Mate et al. (2021) have considered risk aware RMABs and do so through numerical experiments by using a specific utility function for a binary state partially-observable MDP.

In contrast with Mate et al. (2021), our research provides a comprehensive analysis of risk-awareness across more general state spaces, dynamics, and utility functions. This broader approach allows us to uncover deeper insights into risk-aware decision-making in RMABs. The contribution of our work is fourfold. First, we generalize the traditional RMAB with a risk-neutral objective by incorporating risk-awareness that aims to optimize decision-making with respect to the risk criterion. Second, we then provide conditions under which an arm with a risk-aware objective is indexable, enabling the derivation of the Whittle index policy. Third, we address the learning problem when the true transition probabilities are unknown by proposing a Thompson sampling approach, which samples from posterior distributions over the unknown parameters Osband et al. (2013); Russo et al. (2018).

A Recent study, Akbarzadeh and Mahajan (2023), has focused on RMABs with unknown transition dynamics, broadening the scope of applications and solution approaches. Note that papers by Liu et al. (2012); Khezeli and Bitar (2017); Xu et al. (2021) adopt another viewpoint toward restless bandits which is not excatly the same as our problem of interest and it is not based on Whittle index policy. They study a variant of RMABs in which there is a single “best” arm delivering the highest stationary reward; the objective is to learn this arm and pull it indefinitely to maximize long-run return. Although such a strategy is computationally tractable, it differs structurally from our optimization strategy.

We adopts the solution proposed in Akbarzadeh and Mahajan (2023) for our finite-horizon risk-aware setup and derive a regret bound that scales sublinearly in the number of episodes and quadratically with the number of arms. It should be noted that applying a conventional reinforcement learning algorithm to the RMAB in a naive manner is inefficient due to the linear growth of regret in the state space of Markov decision processes, which implies that for RMABs, regret grows exponentially with the number of arms Akbarzadeh and Mahajan (2023). Finally, we numerically illustrate the efficacy of our methodology in reducing risk exposure in RMABs in our experiments in the contexts of machine replacement and patient scheduling applications under both planning and learning setups.

In Section 2, we present the notation, problem formulation for planning and learning setups. Section 3 describes the Whittle index solution concept and a class of indexable RMAB under a risk-aware objective and how Whittle indices can be computed. In Section 4, we address the learning problem. Finally, the numerical analysis is illustrated in Section 5 and the conclusion is discussed in Section 6.

2 PROBLEM DEFINITION

2.1 Notation

All events occur within a finite horizon setup indexed by $t \in \{0, \dots, T-1\} =: \mathcal{T}$. Random variables and their realizations are denoted by capital and lowercase letters, respectively; for example, X_t and x_t. We use calligraphic letters to denote the set of all realizations, such as $\mathcal{X}$. Let $X_{a:b} := \{X_a, \dots, X_b\}$ represent a collection of the random variables from time a to time b, and let $\boldsymbol{X}_t = (X_t^1, \dots, X_t^N)$ represent a collection of the random variable for N processes at time t. The probability and the expected value of random events are denoted by $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$, respectively. Let $\mathbb{I}(\cdot,\cdot)$ be an indicator function for the equality of two terms and $\mathbf{1}_x$ denotes a vector of zeros where only the x-th element is one. The notation $(X)^+$ represents $\max\{0, X\}$. A function f is called superadditive on partially-ordered sets $\mathcal{X}$ and $\mathcal{Y}$ if given $x_1, x_2 \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$ where $x_1 \geq x_2$ and $y_1 \geq y_2$, then $f(x_1, y_1) - f(x_1, y_2) \geq f(x_2, y_1) - f(x_2, y_2)$.

2.2 Restless Multi-arm Bandit

A restless bandit process (arm) is a Markov decision process defined by the tuple $(\mathcal{X}, \mathcal{A}, \{P(a)\}_{a \in \{0,1\}}, r, x_0)$ where $\mathcal{X}$ denotes a finite state space; $\mathcal{A} = \{0, 1\}$ denotes the action space where we call action 0 the *passive* action and action 1 the *active* action; $P(a)$ denotes the transition probability matrix when action $a \in \{0, 1\}$ is chosen; $r : \mathcal{X} \times \{0, 1\} \to \mathbb{R}$ denotes the reward function; and x_0 denotes the initial state of the process. By Markov property we have $\mathbb{P}(X_{t+1} = x_{t+1} | X_{1:t} = x_{1:t}, A_{1:t} = a_{1:t}) := P(x_{t+1} | x_t, a_t)$.

A restless multi-Arm bandit (RMAB) problem consists of a set of N independent arms $(\mathcal{X}^i, \mathcal{A}, \{P^i(a)\}_{a\in\{0,1\}}, r^i, x_0^i)$, $i \in \mathcal{N} := \{1, \ldots, N\}$. An agent observes the state of all arms and may decide to activate up to $M \leq N$ of them. Let $\boldsymbol{\mathcal{X}} := \prod_{i\in\mathcal{N}} \mathcal{X}^i$ denote the joint state space and let $\boldsymbol{\mathcal{A}}(M) := \{\boldsymbol{a} = (a^1, \ldots, a^n) \in \mathcal{A}^n : \sum_{i=1}^N a^i \leq M\}$ denote the action set. The immediate reward realized at time t is $\boldsymbol{r}(\boldsymbol{x}_t, \boldsymbol{a}_t) := \sum_{i\in\mathcal{N}} r^i(x_t^i, a_t^i)$ when the system is in state $\boldsymbol{x}_t$ and the agent chooses action $\boldsymbol{a}_t$. Let $r^i(x_t^i, a_t^i) \in [R_{\min}, R_{\max}]$ where $R_{\min}$ and $R_{\max}$ are finite. Since the arms are independent, the probability of observing state $\boldsymbol{x}_{t+1}$, given the state $\boldsymbol{x}_t$ and the action $\boldsymbol{a}_t$ is denoted by $\boldsymbol{P}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{a}_t) := \prod_{i\in\mathcal{N}} P(x_{t+1}^i|x_t^i, a_t^i)$.

2.3 Planning Problems

Let $\boldsymbol{\pi} = (\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N) : \boldsymbol{\mathcal{X}} \times \mathcal{T} \to \boldsymbol{\mathcal{A}}(M)$ denote a time-dependent deterministic policy for the system where $\boldsymbol{\pi}^i$ defines the action for arm i in the policy of the system, and let $\boldsymbol{\Pi}$ be the set of all such time-dependent deterministic Markov policies.

Assume action A_t^i is prescribed by policy $\boldsymbol{\pi}^i$ at time t. Then, any policy leads to a total reward for the system and for each arm as follows:

$$\boldsymbol{J}_{\boldsymbol{x}_0}(\boldsymbol{\pi}) := \sum_{i\in\mathcal{N}} \sum_{t=0}^{T-1} r^i\left(X_t^i, A_t^i\right) \Big|_{\boldsymbol{\pi}, \boldsymbol{X}_0 = \boldsymbol{x}_0}.$$

We first describe the classical risk-neutral optimization problem as follows (Whittle, 1988).

Problem RMAB: Given time horizon T, and a set of N arms $(\mathcal{X}^i, \mathcal{A}, \{P^i(a)\}_{a\in\{0,1\}}, r^i, x_0^i)$, $i \in \mathcal{N}$ where at most M of them can be activated at a time, find a $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ that maximizes $\mathbb{E}[\boldsymbol{J}_{\boldsymbol{x}_0}(\boldsymbol{\pi})]$.

Problem RMAB is a multi-stage stochastic control problem where the optimal policy is known to be Markovian and presumably obtainable by using dynamic programming (Puterman, 2014). However, as the cardinality of the state space is $\prod_{i\in\mathcal{N}} |\mathcal{X}^i|$, computing the optimal policy is intractable for large N. In Section 3, we describe a well-known heuristic known as *Whittle index* as a solution to tackle this problem.

As discussed earlier, the assumption of risk-neutrality may not be suitable for various practical applications of RMAB. Thus, we generalize the objective to incorporate risk-sensitivity at the level of the total reward generated by each arm. To this end, we leverage an expected utility formulation (see von Neumann and Morgenstern (1947)), which is commonly used in the literature of risk-aware decision-making (von Neumann and Morgenstern, 1947; Fishburn, 1968; Pratt, 1978). More specifically, a concave or convex utility function models a risk averse or seeking behavior respectively. Tversky and Kahneman (1974) further suggest using an S-shaped utility function with inflexion point at a reference value in order to model an attitude that maximizes the chances of reaching such a target.

Problem RRMAB: (Risk-Aware RMAB) Given time horizon T, a set of nondecreasing utility functions U^i, $i \in \mathcal{N}$, and a set of N arms $(\mathcal{X}^i, \mathcal{A}, \{P^i(a)\}_{a\in\{0,1\}}, r^i, x_0^i)$, $i \in \mathcal{N}$, where at most $M \leq N$ of arms can be activated at a time, find a $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ that maximizes $\mathbb{E}[\boldsymbol{D}_{\boldsymbol{x}_0}(\boldsymbol{\pi})]$ defined below

$$\boldsymbol{D}_{\boldsymbol{x}_0}(\boldsymbol{\pi}) := \sum_{i\in\mathcal{N}} U^i\left(\sum_{t=0}^{T-1} r^i\left(X_t^i, A_t^i\right)\right) \Big|_{\boldsymbol{\pi}, \boldsymbol{X}_0 = \boldsymbol{x}_0}.$$

Some notable risk-aware utility functions are described in Section 5.

Problem RRMAB highlights risk-awareness for each arm, which aligns with scalarization methods for multi-objective setups Marler and Arora (2010); Gunantara (2018). Note that solving the risk-aware problem (Problem RRMAB) is more difficult than the risk-neutral problem (Problem RMAB) as already in the case $N = 1$, the dynamic programming equation must be written on a state space augmented by one continuous state capturing the cumulative reward so far (Bäuerle and Rieder, 2014). In section 3, we present how the *Risk-Aware* Whittle index for the risk-aware problem (Problem RRMAB) can be obtained.

2.4 Learning Problem

The parameters of arms in Problem RRMAB may be unknown in various practical applications. Notable examples include a drug discovery problem when a new drug is discovered in a clinical setup Ribba et al. (2020) and a machine maintenance problem for new machines where state transition functions are unknown Ogunfowora and Najjaran (2023). As a result, the objective in such contexts is to determine a learning policy that converges to the ideal policy (the solution to Problem RRMAB) as quickly as possible. Let us assume that the agent interacts with the system for K episodes and let $\boldsymbol{\pi}_k$ denote a time-dependent learning policy in episode k. The performance of a learning policy is measured by Bayesian regret, which quantifies the difference between the policy's performance and that of an oracle who possesses complete knowledge of the environment and executes an optimal policy $\boldsymbol{\pi}^\star$, i.e.,

$$\mathcal{R}(K) = \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}\left[\boldsymbol{D}_{\boldsymbol{x}_0}(\boldsymbol{\pi}^\star)\right] - \mathbb{E}\left[\boldsymbol{D}_{\boldsymbol{x}_0}(\boldsymbol{\pi}_k)\right]\right] \tag{1}$$

where the first expectation is calculated from the prior distribution on $\{P^i(\cdot|x^i, a)\}_{i\in\mathcal{N}, x^i\in\mathcal{X}^i, a\in\{0,1\}}$, while the second expectation is calculated based on the initial states $\boldsymbol{x}_0$ and the learning policy. Bayesian regret is a widely-adopted metric in numerous studies, e.g., Rusmevichientong and Tsitsiklis (2010); Agrawal and Goyal (2013); Russo and Van Roy (2014); Ouyang et al. (2017); Akbarzadeh and Mahajan (2023). As a result, we define the following problem.

Problem LRRMAB: (Learning Risk-Aware RMAB) Given time horizon T, and a set of N arms $(\mathcal{X}^i, \mathcal{A}, r^i, x_0^i)$, $i \in \mathcal{N}$ where at most $M < N$ of them can be activated at a time, find a sequence of $\{\boldsymbol{\pi}_k\}_{k\geq 1}$ that minimizes $\mathcal{R}(K)$.

We present a Thompson sampling algorithm to tackle the learning problem and prove a regret bound of $\mathcal{O}(N^2\sqrt{KT})$. Practically, when the optimal policy of RRMAB is computationally intractable, we can use the Whittle index policy as a proxy for the optimal policy. The effectiveness of this approach is demonstrated in our numerical experiments.

3 INDEXABILITY AND WHITTLE INDEX

3.1 Overview of the Risk-neutral RMAB

Whittle (1988) introduced a priority index policy as a heuristic for Problem RMAB, which has become widely accepted as the conventional method for solving the RMAB Niño-Mora (2023). This policy is a result of a relaxation of the original optimization problem. Particularly, the (hard) constraint of activating at most M arms at a time is replaced by a relaxation where the expected average number of activated arms per time period is at most M, i.e.,

$$\max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \mathbb{E}\left[\boldsymbol{J}_{\boldsymbol{x}_0}(\boldsymbol{\pi})\right] \text{ s.t. } \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\boldsymbol{A}_t\|_1 \middle| \boldsymbol{X}_0 = \boldsymbol{x}_0\right] \leq M.$$

The problem is then decoupled into N independent optimization problem by exploiting a Lagrangean relaxation, parameterized by a multiplier $\lambda \in \mathbb{R}$:

$$\max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \mathbb{E}\left[\boldsymbol{J}_{\boldsymbol{x}_0}(\boldsymbol{\pi})\right] - \lambda\left(\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\boldsymbol{A}_t\|_1 \middle| \boldsymbol{X}_0 = \boldsymbol{x}_0\right] - M\right)$$
$$= \sum_{i=1}^{N}\max_{\bar{\pi}^i\in\bar{\Pi}^i} \bar{J}^i_{\lambda,x_0^i}(\pi^i) - M\lambda \quad (2)$$

where the policy functions captured in $\bar{\Pi}^i$ are Markovian and of the form $\bar{\pi}^i : \mathcal{X}^i \times \mathcal{T} \to \mathcal{A}^i$, and where

$$\bar{J}^i_{\lambda,x_0^i}(\bar{\pi}^i) := \mathbb{E}\left[\sum_{t=0}^{T-1} r^i\left(X_t^i, A_t^i\right) - \frac{\lambda}{T}\sum_{t=0}^{T-1} A_t^i \;\middle|\; X_0^i = x_0^i\right].$$

For each arm i, let $\bar{\pi}_\lambda^{i*}$ denote the optimal policy obtained by dynamic programming on an MDP parametrized by $(\mathcal{X}^i, \mathcal{A}, \{P^i(a)\}_{a\in\{0,1\}}, \bar{r}^i_\lambda, x_0^i)$, with $\bar{r}^i_\lambda(x,a) := r^i(x,a) - \lambda a/T$. The policies π_λ^{i*} can be assembled to obtain a solution of (2) using $\bar{\boldsymbol{\pi}}^*_\lambda := (\bar{\pi}_\lambda^{1*}, \ldots, \bar{\pi}_\lambda^{N*})$.

Next, we define the indexability and Whittle index.

Definition 1 (Indexability and Whittle index)**.** *Given any optimal Markov policy π_λ^{i*}, let the* passive set *be $\mathcal{W}^i_\lambda := \left\{(x,t) \in \mathcal{X}^i \times \mathcal{T} : \pi_\lambda^{i*}(x,t) = 0\right\}$. An RMAB is said to be* indexable *if for all $i \in \mathcal{N}$, $\mathcal{W}^i_\lambda$ is non-decreasing in λ, i.e., for any $\lambda_1, \lambda_2 \in \mathbb{R}$ such that $\lambda_1 \leq \lambda_2$, we have $\mathcal{W}^i_{\lambda_1} \subseteq \mathcal{W}^i_{\lambda_2}$. For an indexable RMAB, the* Whittle index *$w^i(x,t)$ of state $x \in \mathcal{X}^i$ at time t is the smallest value of λ for which state x is part of the passive set $\mathcal{W}^i_\lambda$ at time step t, i.e., $w^i(x,t) = \inf\left\{\lambda \in \mathbb{R}^+ : (x,t) \in \mathcal{W}^i_\lambda\right\}$.*

The *Whittle index policy* activates the arms with the M largest Whittle indices at each time t. By definition, the policy adheres to the M activation limit.

Determining whether a problem is indexable is not immediately apparent, hence researchers have examined various sufficient conditions for indexability (Glazebrook et al., 2006; Niño-Mora, 2007; Akbarzadeh and Mahajan, 2022). Under certain conditions, the Whittle index policy is optimal (Gittins, 1979; Weber and Weiss, 1990; Lott and Teneketzis, 2000) and in other cases, the Whittle index policy is close to optimal (Glazebrook et al., 2006; Niño-Mora, 2007; Avrachenkov et al., 2013; Wang et al., 2020; Akbarzadeh and Mahajan, 2022). General algorithms for computing the Whittle indices are proposed in (Niño-Mora, 2007; Akbarzadeh and Mahajan, 2022), while a numerical search is an alternative for computing the indices (Avrachenkov and Borkar, 2018; Akbarzadeh and Mahajan, 2019a).

3.2 Solution to Problem RRMAB

To establish indexability conditions for Problem RRMAB, we first apply the relaxation and decomposition approach described in Section 3.1. Hence, let π_λ^{i*} denote the optimal policy of an arm that maximizes $\bar{D}^i_{\lambda,x_0^i}(\bar{\pi}^i)$ which is

$$\mathbb{E}\left[U^i\left(\sum_{t=0}^{T-1} r^i\left(X_t^i, A_t^i\right)\right) - \frac{\lambda}{T}\sum_{t=0}^{T-1} A_t^i \middle| X_0^i = x_0^i\right].$$

We follow the steps presented in Bäuerle and Ott (2011) to introduce a new *augmented arm* risk-neutral MDP, which is equivalent to the risk-aware MDP above and can be solved using dynamic programming. Specifically, for each arm

$i \in \mathcal{N}$, we construct a time-dependent MDP $(\bar{\mathcal{X}}^i, \mathcal{A}, \{\bar{P}^i(a)\}_{a\in\{0,1\}}, \{\bar{r}^i_t\}_{t=0}^T, \bar{x}^i_0)$ with $\bar{\mathcal{X}}^i := \mathcal{X}^i \times \mathcal{S}^i$, $\bar{P}^i(x', s'|x, s, a) := P^i(x'|x, a)\mathbb{I}(s', s + r^i(x, a))$, $\bar{r}^i_t(x, s, a) := \mathbb{I}(t, T-1)U^i(s + r^i(x, a)) - (\lambda/T)a$ and $\bar{x}^i_0 := (x^i_0, 0)$. The following result can be immediately derived from Theorem 1 of Bäuerle and Rieder (2014).

Proposition 2. *Let $\tilde{\pi}^{i*}_\lambda$ be an optimal Markovian policy for the augmented arm risk-neutral MDP. Then, one can construct an optimal policy for the relaxation of Problem RRMAB using:*

$$\bar{\pi}^{i*}_\lambda(x_{0:t}, a_{0:t-1}, t) := \tilde{\pi}^{i*}_\lambda(x_t, \sum_{t'=0}^{t-1} r^i(x_{t'}, a_{t'}), t).$$

Namely, $\max_{\bar{\pi}^i \in \bar{\Pi}_H} \bar{D}^i_{\lambda, x^i_0}(\bar{\pi}^i) = \bar{D}^i_{\lambda, x^i_0}(\bar{\pi}^{i*}_\lambda)$ *where* $\bar{\Pi}_H$ *is the set of all history-dependent policies.*

3.3 A class of indexable arms

Puterman (2014, Theorem 4.7.4) provides a set of sufficient conditions under which the optimal policy of a risk-neutral MDP is nondecreasing over the state space. The assumptions and the theorem statement for a risk-neutral arm are presented below.

Assumption 3. *Let $\geq$ be some ordering of states in $\mathcal{X}^i$ and $q^i(k|x, a) := \sum_{z \geq k} P^i(z|x, a)$. Assume that $r^i(x, a)$ is nondecreasing in $x, \forall a \in \mathcal{A}$, $q^i(x'|x, a)$ is nondecreasing in $x, \forall x' \in \mathcal{X}^i$ and $a \in \mathcal{A}$, $r^i(x, a)$ is super-additive on $\mathcal{X}^i \times \mathcal{A}$, and $q^i(x'|x, a)$ is super-additive on $\mathcal{X}^i \times \mathcal{A}$ for every $x' \in \mathcal{X}^i$.*

Lemma 4. *Under Assumption 3, the optimal policy of an arm's MDP $(\mathcal{X}^i, \mathcal{A}, P^i, r^i, x^i_0)$ is nondecreasing over the state space.*

Theorem 5. *For each i, the augmented arm risk neutral MDP is indexable if the arm's original MDP satisfies Assumptions 3, that $q^i(x'|x, a)$ is nondecreasing in a, that $r^i(x, a) = r^i(x)$, for all $x, x' \in \mathcal{X}^i$, and $a \in \mathcal{A}$, and that the utility function U^i is a nondecreasing function.*

Proof Sketch. By definition, an arm is indexable if the passive set is nondecreasing in λ for every state. Thus, it suffices to show that the optimal policy of an arm is nondecreasing in λ. Lemma 4 provides a set of sufficient conditions under which the optimal policy is nondecreasing in $\mathcal{X}$. We adapt these conditions to ensure that this property extends over all possible values of λ.

Theorem 5 indicates that if an arm's MDP satisfies conditions for its optimal risk-neutral policy to be nondecreasing, then its risk-aware policy version is indexable on an augmented state under minor additional assumptions. In Appendix F, we provide a set of simple structured MDPs that satisfy the assumptions made in Theorem 5.

There are notable practical implications with respect to these assumptions. In the machine maintenance application Glazebrook et al. (2006); Akbarzadeh and Mahajan (2019b); Abbou and Makis (2019), they imply that the state of each machine has a direct impact on its performance and so, a better state results in a higher reward. In addition, the machines deteriorate stochastically if not maintained, with a chance of worsening to any state (generalizing the right skip-free transitions of Glazebrook et al. (2006)). However, the chance of visiting a worse state is higher if the current state is worse, which is often the case in practice. Also, on the other hand, repairs may not always fully fix the issue (generalizing the deterministic transition of Glazebrook et al. (2006) under repairs) but the chance is higher in better states.

The same model also works for patient scheduling in hospitals. These assumptions imply that patients' conditions can deteriorate stochastically if not attended to, with worse conditions increasing the urgency for intervention. Treatments, while beneficial, may not always fully restore health immediately, reflecting a probabilistic transition towards recovery. This probability is higher if the patient is in a better state. This reflects realistic clinical scenarios where early and continuous care increases the likelihood of reaching an optimal health state Starfield et al. (2005)

Remark 6. *All the results obtained in the upcoming sections are applicable for any indexable augmented arm risk neutral MDP. Theorem 5 presents only a set of sufficient conditions (and not necessary conditions) for the indexability of the problem.*

3.4 Computation of Whittle indices

Primarily, there are three approaches to consider. One approach is problem-specific where the Whittle index formula is computed exactly (Jacko and Villar, 2012; Borkar et al., 2018; Yu et al., 2018; Wang et al., 2019). The other one is the *modified adaptive greedy algorithm* (Akbarzadeh and Mahajan, 2022) which works for any indexable RMAB in discounted and average reward setups. And the last one is numerical search which is either through *adaptive greedy* (Niño-Mora, 2007), binary search (Qian et al., 2016; Akbarzadeh and Mahajan, 2019a), or brute-force search.

In this work, we adapt the numerical search approach described in Akbarzadeh and Mahajan (2019a) to compute the Whittle indices using binary search. The procedure of the binary search algorithm is described below and also presented in Appendix C in more detail including its computational complexity.

Let us assume that penalties $\mathtt{LB} \in \mathbb{R}^+$ and $\mathtt{UB} \in \mathbb{R}^+$ are such that $\pi^{i*}_{\mathtt{LB}}(x,s,t) = 1$ and $\pi^{i*}_{\mathtt{UB}}(x,s,t) = 0$ for all $(x,s,t) \in \mathcal{X}^i \times \mathcal{S}^i \times \mathcal{T}$. Then, we construct the sequence of passive sets from an empty set to the whole state space by a binary search.

The algorithm starts by initializing λ_l to the lower bound and an empty passive set $\mathcal{W}_{\lambda_l}$. It then enters a loop where it computes the critical penalty $\lambda_c := \inf\{\lambda : \pi_\lambda \neq \pi_{\lambda_l}\}$ using a separate algorithm, Algorithm 3, which performs a binary search between λ_l and the upper bound UB. Next, the algorithm identifies the set $\mathcal{M}$ of states (x,s,t) where the policy changes from active to passive as the penalty increases from λ_l to λ_c. For each state in $\mathcal{M}$, the Whittle index is set to λ_c. The lower bound λ_l, the policy π_{λ_l}, and the passive set $\mathcal{W}_{\lambda_l}$ are then updated to reflect the new penalty. This process continues until the passive set $\mathcal{W}_{\lambda_l}$ encompasses the entire state space $\mathcal{X} \times \mathcal{S} \times \mathcal{T}$. At this point, the algorithm returns the computed Whittle indices.

4 LEARNING POLICY

In this section, we propose a learning algorithm, called *Learning Risk-Aware Policy using Thompson Sampling* (LRAP-TS), to address Problem LRRMAB. In this setup, a learning agent aims to minimize the regret in (13) during K episodes of interaction with the system by balancing the *exploration-exploitation trade-off* (Auer et al., 2002).

We first introduce additional parameters and variables. Let $\theta^{\star i} \in \Theta^i$ represent the unknown parameters of arm $i \in \mathcal{N}$, where Θ^i is a compact set. Additionally, we assume that $\theta^{\star i}$ are independent of each other and $\boldsymbol{\theta}^\star = (\theta^{\star 1}, \ldots, \theta^{\star N})$. Let ϕ^i_1 capture the prior on θ^i for each arm $i \in \mathcal{N}$. Furthermore, we let h^i_k be the history of states and actions of arm i during episode k, and ϕ^i_k be the posterior distribution on $\theta^{\star i}$ given h^i_k. Then, upon applying action a at state x and observing the next state x' for arm i, the posterior distribution ϕ^i_{k+1} can be computed using Bayes rule as

$$\phi^i_{k+1}(d\theta) = \frac{P^i_\theta(x'|x,a)\phi^i_t(d\theta)}{\int P^i_{\tilde{\theta}}(x'|x,a)\phi^i_t(d\tilde{\theta})}. \tag{3}$$

If the prior is a conjugate distribution on Θ^i, then the posterior can be updated in closed form. We note that our algorithm and regret analysis do not depend on the specific structure of the prior and posterior update rules.

4.1 LRAP-TS Algorithm

Algorithm LRAP-TS operates in episodes each with length T. It maintains a posterior distribution ϕ^i_k on the dynamics of arm i and keeps track of $N^i_{k,t}(x,a) = \sum_{\kappa=1}^k \sum_{\tau=0}^t \mathbb{1}\{(X^i_{\kappa,\tau}, A^i_{\kappa,\tau}) = (x,a)\}$ and $N^i_{k,t}(x,a,x_+) = \sum_{\kappa=1}^k \sum_{\tau=0}^t \mathbb{1}\{(X^i_{k,t}, A^i_{k,t}, X^i_{k,t+1}) = (x,a,x_+)\}$.

At the beginning of episode k, one starts by sampling for each arm $i \in \mathcal{N}$ a set of parameters θ^i_k from the posterior distribution ϕ^i_k. The optimal policy for problem LRRMAB under $\{\theta^i_k\}_{i\in\mathcal{N}}$ is then identified and implemented until the end of the horizon. The transitions and rewards observed along the trajectories are finally used to update the posteriors ϕ^i_{k+1} for each arm. The algorithm is described in Algorithm 1.

Algorithm 1 LRAP-TS

1: Input: $x^i_0, \phi^i_1, \forall i \in \mathcal{N}$.
2: **for** $k = 1, 2, \ldots, K$ **do**
3: Sample $\theta^i_k \sim \phi^i_k$ for arm $i \in \mathcal{N}$ and compute the estimated risk-aware policy.
4: **for** $t \in \mathcal{T}$ **do**
5: Implement actions based on the estimated risk-aware policy.
6: **end for**
7: Update ϕ^i_{k+1} according to (3) for arm $i \in \mathcal{N}$.
8: **end for**

4.2 Regret bound

With a slight abuse of notation, let $(\boldsymbol{P}^\star, \boldsymbol{V}^\star, \boldsymbol{\pi}^\star)$ denote the transition probability matrix, the optimal value function, and the optimal policy for the overall system parameterized by the true parameters $\theta^\star$ and let $(\boldsymbol{P}_k, \boldsymbol{V}_k, \boldsymbol{\pi}_k)$ denote the transition probability matrix, the optimal value function, and the optimal policy for the overall system parameterized by the estimated parameters θ_k in episode k.

We first bound the expected error in estimation of the unknown transition probabilities over all arms and rounds.

Lemma 7. *Let $|\bar{\mathcal{X}}| = \max_i |\mathcal{X}^i|$. Then we have*

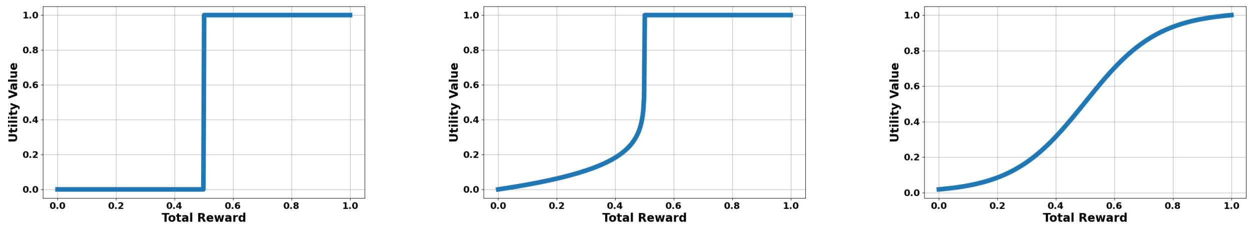
$$\sum_{k=1}^{K} \sum_{t=0}^{T-1} \mathbb{E}\Big[\big\|\boldsymbol{P}^\star(\cdot\,|\boldsymbol{X}_{k,t}, \boldsymbol{A}_{k,t}) - \boldsymbol{P}_k(\cdot\,|\boldsymbol{X}_{k,t}, \boldsymbol{A}_{k,t})\big\|_1\Big] \leq 12N|\bar{\mathcal{X}}|\sqrt{KT(1+\log KT)}.$$

Finally, we bound the expected regret defined in (1).

Theorem 8.

$$\mathcal{R}(K) \leq 12N^2 T R_{\max}|\bar{\mathcal{X}}|\sqrt{KT(1+\log KT)}.$$

The result shows the regret bound is sublinear in the number of episodes K and quadratic in the number of arms N.



(a) $(\alpha, \tau) = (1, 0.5)$ (b) $(\alpha, \tau, o) = (2, 0.5, 8)$ (c) $(\alpha, \tau, o) = (3, 0.5, 8)$

Figure 1: Sample plots of the three utility functions.

5 NUMERICAL ANALYSIS

We evaluate the risk-aware Whittle index policy through numerical experiments to validate the robustness and efficacy of our models in both planning and learning contexts. We consider two applications, *machine maintenance*, and *patient scheduling*.

We first describe planning experiments assessing our model's performance under various parameter configurations for a machine maintenance problem instance which is generated synthetically and aligns with Akbarzadeh and Mahajan (2019b). In the second set of experiments, we evaluate the learning algorithm for LRRMAB, based on both the synthetic machine maintenance model and patient scheduling for the advanced breast cancer models created from real-world data in Le (2016).

Within these experiments, we consider three utility functions for arm i, also shown in Fig. 1. For arm i, they are

$$U^i_{\alpha,\tau,o}(J^i) = \begin{cases} \mathbb{I}(J^i - \tau, 0) & \text{if } \alpha = 1 \\ 1 - \tau^{-1/o} \max(0, \tau - J^i)^{1/o} & \text{if } \alpha = 2 \\ (1 + e^{-o(1-\tau)})/(1 + e^{-o(J^i - \tau)}) & \text{if } \alpha = 3 \end{cases}$$

where the risk parameters are the target value τ, and the order o. The first function maximizes the probability of being above τ. The second shows a risk-seeking tendency for total rewards below the threshold. The third combines risk-seeking behavior below the threshold with risk aversion above it.

The code for all experiments is available on GitHub.[1]

Planning: We consider a machine maintenance scheduling problem with a fleet of non-homogeneous machines (arms) where all of them have the same number of states, generate similar rewards, but differ in Markov dynamics.

The per-step reward of each arm is a linear, non-decreasing function of the state, ranging from 0 to $1/T$. This assumption makes the total reward of a round be bounded by 1. State transitions are chosen from structured matrices $\{\mathcal{P}_a(p)\}_{a\in\{0,1\}}$ parameterized by p which are linearly spaced within $[0.1/|X|, 1/|X|]$ across different arms adhering to the assumptions in Theorem 5 and reflecting the type of dynamics encountered in a machine maintenance problem (See Model 4 in Appendix F).

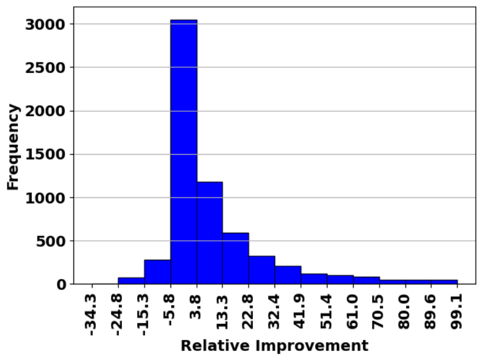
We explored 6804 setups to analyze the behavior of our risk-aware policy. The instances have been created out of all combinations of the following parameters: Time horizon $T \in \{3, 4, 5\}$, State space size $X \in \{2, 3, 4, 5\}$, Number of arms $N \in \{3X, 4X, 5X\}$, Utility functions in $\{(\alpha = 1), (\alpha = 2, o = 4), (\alpha = 2, o = 8), (\alpha = 2, o = 16), (\alpha = 3, o = 4), (\alpha = 3, o = 8), (\alpha = 3, o = 16)\}$, Threshold $\tau \in \{0.1, 0.2, \ldots, 0.8, 0.9\}$, Number of arms to be activated in $\{\lfloor .3N \rfloor, \lfloor .4N \rfloor, \lfloor .5N \rfloor\}$.

In all experiments, each policy's performance was evaluated using Monte Carlo simulations, averaging over 100 sample paths. Given the novelty of our approach, we compare the performance of our proposed risk-aware Whittle index policy against the risk-neutral version and provide a relative improvement of the proposed method with respect to the risk-neutral version measured in percentages.

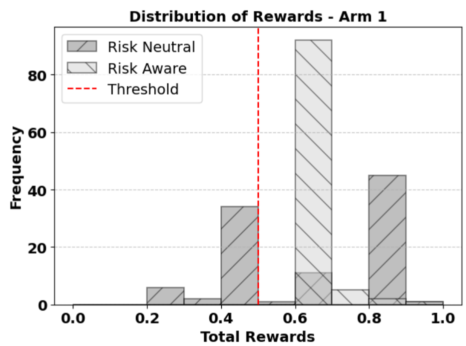
Fig. 2(a) summarizes the relative improvement in the objective function achieved by our policy in a histogram. Some statistics are reported in Table 1. Fig. 2(b) shows the distribution of total rewards under both policies for a specific setup. The risk-aware policy achieved 100% positive mass compared to 58% for the risk-neutral policy.

Mate et al. (2021) addressed a binary state partially-observable restless bandit problem, while our work considers a fully-observable setup. Although the two setups differ fundamentally, we include a numerical baseline inspired by Mate et al.'s who maximize the expected sum of stage-wise utilities instead of the expected utility of the sum of stage-wise performances. To this end, we will include a baseline (named *Sum of Stage-wise Utilities Policy* (SSUP)) where the ob-

[1] `https://github.com/nima-akbarzadeh/AISTATS25`

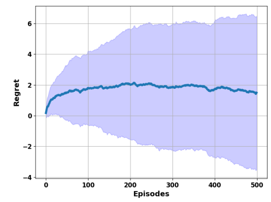


(a)

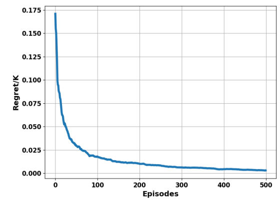


(b)

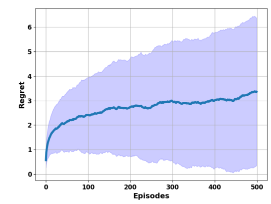
Figure 2: Figure (a) shows the distribution of relative improvements in the objective function achieved by our proposed policy compared to the risk-neutral one in 6804 different setups. This histogram is limited to range of values up to 100. Figure (b) illustrates the distribution of total rewards under both risk-aware and risk-free policies for for one of the arms with the setup $T = 5$, $|\mathcal{X}| = 5$, $N = 25$, $M = 7$, $\alpha = 1$, $\tau = 0.5$. The red line is set at the target τ.



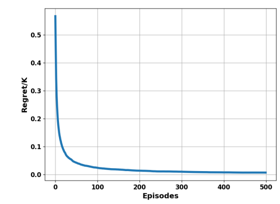
(a) $(|\mathcal{X}|, N) = (3, 4)$



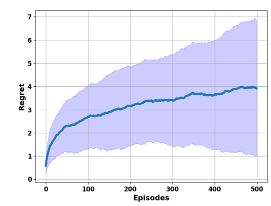
(b) $(|\mathcal{X}|, N) = (3, 4)$



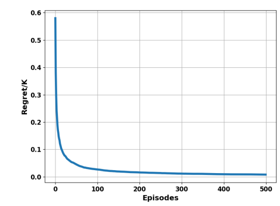
(c) $(|\mathcal{X}|, N) = (5, 4)$



(d) $(|\mathcal{X}|, N) = (5, 4)$



(e) $(|\mathcal{X}|, N) = (5, 5)$



(f) $(|\mathcal{X}|, N) = (5, 5)$

Figure 3: The plot shows $\mathcal{R}(k)$ of RAWIP on the left and $\mathcal{R}(k)/K$ on the right for three different setups when the utility function is $\alpha = 1$. For all these experiments, $T = 5$ and $M = 1$, $\tau = 0.5$. These plots are averaged over 100 sample paths. The transition model for figures $(a)\&(b)$ is according to model 3 of Le (2016) and $(c)\&(d)$ and $(e)\&(f)$ are according to $\mathcal{P}$.

jective is modified to: $\sum_{t=0}^{T-1} \frac{1}{T} U^i_{\alpha,\tau,o}(Tr^i_t)$. Similar to RAWIP, we compute the relative improvement of this baseline with respect to the risk-neutral Whittle index policy. Results (Table 1) indicate that our risk-aware Whittle index policy outperforms the baseline in all the provided statistics, highlighting its superior performance.

To provide better clarity, we plan to include tables illustrating performance variations with respect to both the threshold (τ) and the utility functions considered in the study. This will help highlight how different utility functions affect performance under various parameter configurations. Please refer to the table below.

The table illustrates the average performance of our algorithm (computed across multiple problem setups and iterations) as parameters change. For utility functions 2 and 3, the results show that increasing the order—moving further from the linear case—highlights the superior performance of the risk-aware policy. This is expected, as risk-aware and risk-neutral policies perform identically under a linear utility function.

Learning: This section outlines the experiments conducted for LRRMAB. Since computing the optimal policy is infeasible for large numbers of arms, even with estimated or true parameters, we use the risk-aware Whittle index policy (Section 3.4, which we refer to as RAWIP) as a proxy. We refer to the algorithm as *Risk-Aware Whittle Index Policy using Thompson Sampling* (RAWIP-TS). We consider three models: one for patient scheduling and two for machine maintenance scheduling.

For patient scheduling problem, each arm represents a patient with Advanced Breast Cancer, Le (2016), where states represent disease progression (DEAD: 0, PROGRESSING: 1, STABLE: 2). Patients are treated with either Capecitabine or Lapatinib+Capecitabine. Due to limited hospital resources, the problem is modeled as a restless bandit problem with a fixed bud-

Policy	Min (%)	Max (%)	Mean (%)	% Above 0
RAWIP	-40.82	4842.0	37.9	84.9
SSUP	-45.84	560.1	3.46	64.1

Table 1: Relative improvement of policies compared to the risk-neutral Whittle index policy.

Parameter Value	**Relative Improvement**
$\alpha = 1$	120.39%
$(\alpha, o) = (2, 4)$	6.66%
$(\alpha, o) = (2, 8)$	16.89%
$(\alpha, o) = (2, 16)$	34.04%
$(\alpha, o) = (3, 4)$	3.76%
$(\alpha, o) = (3, 8)$	23.73%
$(\alpha, o) = (3, 16)$	59.61%

Table 2: The effect of the utility function on the relative improvement of RAWIP with respect to the risk-neutral Whittle index policy

get. The rewards for each state are: $r(\text{DEAD}) = 0$, $r(\text{PROGRESSING}) = 1/(2T)$, $r(\text{STABLE}) = 1/(T)$. We use model 3 described in Le (2016) as the state dynamics. According to Table 1 of the paper, there's a range for for each transition probabilities. We first choose a value for each transition probability uniformly random from the given range and assuring the sum of transitions from a given state is 1, and then run the experiments.

Remark 9. *The randomly generated Markov dynamics from Le (2016) may not satisfy our assumptions mentioned in Theorem 5 for indexability. However, the problem instances are indexable. This showcases that our approach works for any indexable RMABs and the assumptions mentioned in Theorem 5 are only a set of sufficient conditions.*

For machine maintenance problem, each arm represents a machine with exactly the same setup as the one described in the planning part. However, the true values of p^i are unknown to the learner. We assume $p^i \in [0.1/|\mathcal{X}|, 1/|\mathcal{X}|]$ for arm i, and set the initial estimate $\hat{p}^i_0$ to the median value $0.5/|\mathcal{X}|$.

For the learning problem, the true transition probabilities are unknown, so we set a Dirichlet prior for Thompson sampling.

At the start of each episode, the RAWIP is computed based on estimated parameters. During the episode, the learner observes 10 batches of finite-horizon rounds. State-action-state observations for each arm are counted, and Dirichlet distributions are constructed over the unknown parameters. These parameters are then updated by sampling from the distributions.

In Figure 3, we illustrate $\mathcal{R}(k)$ for a learner implementing RAWIP-TS, compared to an oracle with full knowledge of the true parameters who implements the true RAWIP. The results suggest the learning mechanism is effective, with regret growing sublinearly as $\mathcal{R}(k)/K$ converges to 0. Note as the RAWIP is generally suboptimal, the RAWIP-TS may outperform it and hence, the slope of cumulative regret becomes negative (as in case (a)). However, the learner's policy eventually converges to the RAWIP (as shown in (b)).

6 CONCLUSION

Our study extends the traditional RMAB by incorporating risk-awareness, providing a robust framework for risk-aware decision-making. We establish indexability conditions for risk-aware objectives and propose a Thompson sampling approach that achieves bounded regret, scaling sublinearly with episodes and quadratically with arms. Rigorous experiments on numerous setups affirm the potential of our methodology in practical applications to effectively control risk exposure.

Acknowledgements

Nima Akbarzadeh was partially funded by GERAD, FRQNT, and IVADO. Erick Delage was partially supported by the Canadian Natural Sciences and Engineering Research Council [Grant RGPIN-2022-05261] and by the Canada Research Chair program [950-230057]. Yossiri Adulyasak was partially supported by the Canadian Natural Sciences and Engineering Research Council [Grant RGPIN-2021-03264] and by the Canada Research Chair program [CRC-2022-00087].

References

Carlos Abad and Garud Iyengar. A near-optimal maintenance policy for automated DR devices. 7(3): 1411–1419, 2016.

Abderrahmane Abbou and Viliam Makis. Group maintenance: A restless bandits approach. *INFORMS Journal on Computing*, 31(4):719–731, 2019.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Int. Conf. Mach. Learn.*, pages 127–135. PMLR, 2013.

Nima Akbarzadeh and Aditya Mahajan. Dynamic spectrum access under partial observations: A restless bandit approach. In *Canadian Workshop on Information Theory*, pages 1–6. IEEE, 2019a.

Nima Akbarzadeh and Aditya Mahajan. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *Conf. Decision Control*, pages 7294–7300, 2019b.

Nima Akbarzadeh and Aditya Mahajan. Conditions for indexability of restless bandits and an $\mathcal{O}(k^3)$ algorithm to compute Whittle index. *Adv. in Appl. Probab.*, 54(4):1164–1192, 2022.

Nima Akbarzadeh and Aditya Mahajan. On learning Whittle index policy for restless bandits with scalable regret. *IEEE Transactions on Control of Network Systems*, 2023.

S Amiri, M Honarvar, et al. Providing an integrated model for planning and scheduling energy hubs and preventive maintenance. *Energy*, 163:1093–1114, 2018.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Konstantin Avrachenkov, Urtzi Ayesta, Josu Doncel, and Peter Jacko. Congestion control of TCP flows in internet routers by means of index policy. *Comm Com Inf Sc*, 57(17):3463–3478, 2013.

Konstantin E. Avrachenkov and Vivek S. Borkar. Whittle index policy for crawling ephemeral content. *IEEE Transactions on Control of Network Systems*, 5(1):446–455, 2018.

Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74:361–379, 2011.

Nicole Bäuerle and Ulrich Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

Vivek S. Borkar, Gaurav S. Kasbekar, Sarath Pattathil, and Priyesh Y. Shetty. Opportunistic scheduling as restless bandits. 5(4):1952–1961, 2018.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015.

Abhinav Dahiya, Nima Akbarzadeh, Aditya Mahajan, and Stephen L. Smith. Scalable operator allocation for multirobot assistance: A restless bandit approach. 9(3):1397–1408, 2022.

Sarang Deo, Seyed Iravani, Tingting Jiang, Karen Smilowitz, and Stephen Samuelson. Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operat. Res.*, 61(6):1277–1294, 2013.

Peter C Fishburn. Utility theory. *Management science*, 14(5):335–378, 1968.

John C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B*, pages 148–177, 1979.

K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride. Some indexable families of restless bandit problems. *Adv. Appl. Prob.*, 38(3):643–672, 2006.

KD Glazebrook, DJ Hodge, and Christopher Kirkbride. Monotone policies and indexability for bidirectional restless bandits. *Adv. Appl. Prob.*, 45(1):51–85, 2013.

Kevin D. Glazebrook, H. M. Mitchell, and P. S. Ansell. Index policies for the maintenance of a collection of machines by a set of repairmen. *Euro. J. Operat. Res.*, 165(1):267–284, 2005.

Nyoman Gunantara. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, 5(1):1502242, 2018.

Fang He, Thierry Chaussalet, and Rong Qu. Controlling understaffing with conditional value-at-risk constraint for an integrated nurse scheduling problem under patient demand uncertainty. *Operations Research Perspectives*, 6:100119, 2019.

Peter Jacko and Sofía S Villar. Opportunistic schedulers for optimal scheduling of flows in wireless systems with arq feedback. In *International Teletraffic Congress*, page 14, 2012.

Sebastian Jaimungal, Silvana M Pesenti, Ye Sheng Wang, and Hariom Tatsat. Robust risk-aware reinforcement learning. *SIAM Journal on Financial Mathematics*, 13(1):213–226, 2022.

Young Hun Jung, Marc Abeille, and Ambuj Tewari. Thompson sampling in non-episodic restless bandits. *arXiv preprint arXiv:1910.05654*, 2019.

Kia Khezeli and Eilyan Bitar. Risk-sensitive learning and pricing for demand response. *IEEE Transactions on Smart Grid*, 9(6):6000–6007, 2017.

Quang A. Le. Structural uncertainty of markov models for advanced breast cancer: A simulation study of lapatinib. *Medical Decision Making*, 36(5):629–640, 2016.

Yann Le Tallec. *Robust, risk-sensitive, and data-driven control of Markov decision processes.* PhD thesis, Massachusetts Institute of Technology, 2007.

Haoyang Liu, Keqin Liu, and Qing Zhao. Learning in a changing world: Restless multiarmed bandit with unknown dynamics. *IEEE Transactions on Information Theory*, 59(3):1902–1916, 2012.

Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access. 56(11):5547–5567, 2010.

Daniel Zhuoyu Long, Melvyn Sim, and Minglong Zhou. Robust satisficing. *Operations Research*, 71 (1):61–82, 2023.

Christopher Lott and Demosthenis Teneketzis. On the optimality of an index rule in multichannel allocation for single-hop mobile networks with multiple service classes. *Probab. Engrg. Inform. Sci.*, 14(3): 259–297, 2000.

Shie Mannor, Ofir Mebel, and Huan Xu. Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

R Timothy Marler and Jasbir S Arora. The weighted sum method for multi-objective optimization: new insights. *Structural and multidisciplinary optimization*, 41:853–862, 2010.

Aditya Mate, Andrew Perrault, and Milind Tambe. Risk-aware interventions in public health: Planning with restless multi-armed bandits. In *AAMAS*, pages 880–888, 2021.

Amirhossein Najjarbashi and Gino J Lim. A variability reduction method for the operating room scheduling problem under uncertainty using cvar. *Operations Research for Health Care*, 20:25–32, 2019.

José Niño-Mora. Dynamic priority allocation via restless bandit marginal productivity indices. *TOP*, 15 (2):161–198, 2007.

José Nino-Mora. A restless bandit marginal productivity index for opportunistic spectrum access with sensing errors. In *International Conference on Network Control and Optimization*, pages 60–74. Springer, 2009.

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

José Niño-Mora. A $(2/3)n^3$ fast-pivoting algorithm for the gittins index and optimal stopping of a Markov Chain. *INFORMS J. Comput.*, 19(4):596–606, 2007.

Oluwaseyi Ogunfowora and Homayoun Najjaran. Reinforcement and deep reinforcement learning-based solutions for machine maintenance planning, scheduling policies, and optimization. *Journal of Manufacturing Systems*, 70:244–263, 2023.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Takayuki Osogami. Robustness and risk-sensitivity in Markov decision processes. *Advances in Neural Information Processing Systems*, 25, 2012.

Wenzhuo Ouyang, Atilla Eryilmaz, and Ness B Shroff. Downlink scheduling over Markovian fading channels. 24(3):1801–1812, 2015.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown Markov decision processes: A Thompson sampling approach. In *Neural Inf. Process. Syst.*, volume 30. Curran Associates, Inc., 2017.

John W Pratt. Risk aversion in the small and in the large. In *Uncertainty in economics*, pages 59–79. Elsevier, 1978.

Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 123–131, 2016.

Marvin Rausand. *Risk assessment: theory, methods, and applications*, volume 115. John Wiley & Sons, 2013.

Benjamin Ribba, Sherri Dudal, Thierry Lavé, and Richard W Peck. Model-informed artificial intelligence: reinforcement learning for precision dosing. *Clinical Pharmacology & Therapeutics*, 107(4):853–857, 2020.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35 (2):395–411, 2010.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Math. Oper. Res.*, 39(4):1221–1243, 2014.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Barbara Starfield, Leiyu Shi, and James Macinko. Contribution of primary care to health systems and health. *Milbank*, 83(3):457–502, 2005.

A Tversky and D Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157): 1124–1131, September 1974.

J. von Neumann and O. Morgenstern. *Theory of games and economic behavior.* Princeton University Press, 1947.

J. Wang, X. Ren, Y. Mo, and L. Shi. Whittle index policy for dynamic multichannel allocation in remote state estimation. *IEEE Transactions on Automatic Control*, 65(2):591–603, 2020.

Kehao Wang, Jihong Yu, Lin Chen, Pan Zhou, Xiaohu Ge, and Moe Z Win. Opportunistic scheduling revisited using restless bandits: Indexability and index policy. 18(10):4997–5010, 2019.

Qingsi Wang, Mingyan Liu, and Johanna L Mathieu. Adaptive demand response: Online learning of restless and controlled bandits. In *Int. Conf. Smart Grid Commun.*, pages 752–757. IEEE, 2014.

Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *J. Appl. Prob.*, 27(3): 637–648, 1990.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *J. Appl. Prob.*, 25(A):287–298, 1988.

Jianyu Xu, Lujie Chen, and Ou Tang. An online algorithm for the risk-aware restless bandit. *European Journal of Operational Research*, 290(2):622–639, 2021.

Wenhao Xu, Xuefeng Gao, and Xuedong He. Regret bounds for Markov decision processes with recursive optimized certainty equivalents. In *International Conference on Machine Learning*, pages 38400–38427. PMLR, 2023.

Zhe Yu, Yunjian Xu, and Lang Tong. Deadline scheduling as restless bandits. *IEEE Transactions on Automatic Control*, 63(8):2343–2358, 2018.

A PROOF OF LEMMA 4

We start by characterizing the sufficient conditions under which the optimal policy is increasing in the state space of a general finite horizon time-dependent MDP $(\mathcal{X}, \mathcal{A}, P, r, \mu)$, with μ as an initial distribution over $\mathcal{X}$.

Proposition 10. *Suppose for $t \in \mathcal{T}$ that*

1. *$r_t(x, a)$ is nondecreasing in $x, \forall a \in \mathcal{A}$,*
2. *$q_t(x'|x, a)$ is nondecreasing in $x, \forall x' \in \mathcal{X}$ and $\forall a \in \mathcal{A}$,*
3. *$r_t(x, a)$ is superadditive on $\mathcal{X} \times \mathcal{A}$,*
4. *$q_t(x'|x, a)$ is superadditive on $\mathcal{X}^i \times \mathcal{A}$ for every $x' \in \mathcal{X}^i$,*
5. *$r_T(x)$ is nondecreasing in x,*

where $q_t(x'|x, a) := \sum_{z \geq x'} P_t^i(z|x, a)$. Then there exist an optimal policy which is nondecreasing in $x \in \mathcal{X}$, $\forall t \in \mathcal{T}$.

See Puterman (2014, Theorem 4.7.4).

For the risk-neutral arm MDP, the reward and transitions are time-independent. Hence, based on Assumption 3 and Proposition 10, the optimal policy of arm i with MDP $(\mathcal{X}^i, \mathcal{A}, P^i, r^i, x_0^i)$ is nondecreasing in $x \in \mathcal{X}^i$.

B PROOF OF THEOREM 5

For indexability, we need to prove that the passive set is nondecreasing in λ. As a result, if we show that the optimal policy is nonincreasing in λ or equivalently, nondecreasing in ψ when $\lambda := -\psi T$ for every state, then the process is necessarily indexable. In order to make such a demonstration, we introduce a state variable with a constant realization $\psi \in \Psi := \{\psi_1, \psi_2\}$ with $\psi_1 \geq \psi_2$. For simplicity, in what follows we drop superscript i in the proof of the theorem.

Let $\hat{\mathcal{X}} = \mathcal{X} \times \mathcal{S} \times \Psi$, $\hat{r}_t(x, s, \psi, a) := \mathbb{I}(t, T-1)U(s + r(x)) + \psi a$. Let also

$$\begin{aligned}
\hat{P}\left((x', s', \psi')|(x, s, \psi), a\right) &:= \mathbb{I}(\psi' = \psi)\mathbb{I}(s', s + r(x))P(x'|x, a). \\
\hat{q}((x', s', \psi')|(x, s, \psi), a) &:= \sum_{x'' \geq x', s'' \geq s'} \sum_{\psi'' \geq \psi'} \hat{p}\left((x', s', \psi')|(x, s, \psi), a\right) \\
&= \sum_{x'' \geq x', s'' \geq s'} \sum_{\psi'' \geq \psi'} \mathbb{I}(\psi'', \psi)\mathbb{I}(s', s + r(x))P(x''|x, a) \\
&= \mathbb{I}(\psi' \leq \psi, \text{True})\mathbb{I}(s'' \leq s + r(x), \text{True})q(x'|x, a).
\end{aligned}$$

Note that as the state and action spaces are finite and $r(x)$ belongs to a finite set, then $s + r(x)$ belongs to a finite set accordingly. We can therefore verify the conditions of Proposition 10 on the *doubly augmented arm* MDP $(\hat{\mathcal{X}}, \mathcal{A}, \hat{P}, \{\hat{r}_t\}_{t=0}^{T-1}, \hat{\mu})$, with $\hat{\mu}$ equally distributed among ψ_1 and ψ_2.

1. Let $x_1 \geq x_2$, $s_1 \geq s_2$, and $\psi_1 \geq \psi_2$. Then, we have

$$\begin{aligned}
\hat{r}_t((x_1, s_1, \psi_1), a) &= \mathbb{I}(t, T-1)U(s_1 + r(x_1)) + \psi_1 a \\
&\geq \mathbb{I}(t, T-1)U(s_2 + r(x_2)) + \psi_2 a = \hat{r}_t((x_2, s_2, \psi_2), a),
\end{aligned}$$

given that $r(x_1) \geq r(x_2)$ based on Assumption 3 and that $U(\cdot)$ is a nondecreasing function.

2. Let $x_1 \geq x_2$, $s_1 \geq s_2$, and $\psi_1 \geq \psi_2$. Then

$$\begin{aligned}
&\hat{q}((x', s', \psi')|(x_1, s_1, \psi_1), a) \\
&= \mathbb{I}(\psi' \leq \psi_1, \text{True})\mathbb{I}(s' \leq s_1 + r(x_1), \text{True})q(x'|x_1, a) \\
&\geq \mathbb{I}(\psi' \leq \psi_2, \text{True})\mathbb{I}(s' \leq s_2 + r(x_2), \text{True})q(x'|x_2, a) \\
&= \hat{q}((x', s'\psi')|(x_2, s_2, \psi_2), a).
\end{aligned}$$

3. Let $x_1 \geq x_2$, $s_1 \geq s_2$, $\psi_1 \geq \psi_2$, and $a_1 \geq a_2$. Then

$$\begin{aligned}
&\hat{r}_t((x_1, s_1, \psi_1), a_1) + \hat{r}_t((x_2, s_2, \psi_2), a_2) \\
&\quad = \mathbb{I}(t, T-1)U(s_1 + r(x_1)) + \psi_1 a_1 + \mathbb{I}(t, T-1)U(s_2 + r(x_2)) + \psi_2 a_2 \\
&\quad \geq \mathbb{I}(t, T-1)U(s_1 + r(x_1)) + \psi_1 a_2 + \mathbb{I}(t, T-1)U(s_2 + r(x_2)) + \psi_2 a_1 \\
&\quad = \hat{r}_t((x_1, s_1, \psi_1), a_2) + \hat{r}_t((x_2, s_2, \psi_2), a_1).
\end{aligned}$$

where the inequality holds due to the fact that

$$\psi_1 a_1 + \psi_2 a_2 - \psi_1 a_2 - \psi_2 a_1 = (\psi_1 - \psi_2)(a_1 - a_2) \geq 0.$$

4. Let $x_1 \geq x_2$, $s_1 \geq s_2$, $\psi_1 \geq \psi_2$, and $a_1 \geq a_2$. Then, let

$$\begin{aligned}
(*) &:= \hat{q}((x', s', \psi')|(x_1, s_1, \psi_1), a_1) \\
&\quad + \hat{q}((x', s', \psi')|(x_2, s_2, \psi_2), a_2) \\
&= \mathbb{I}(\psi' \leq \psi_1, \text{True})\mathbb{I}(s' \leq s_1 + r(x_1), \text{True})q(x'|x_1, a_1) \\
&\quad + \mathbb{I}(\psi' \leq \psi_2, \text{True})\mathbb{I}(s' \leq s_2 + r(x_2), \text{True})q(x'|x_2, a_2), \\
(**) &:= \hat{q}((x', s', \psi')|(x_1, s_1, \psi_1), a_2) \\
&\quad + \hat{q}((x', s', \psi')|(x_2, s_2, \psi_2), a_1).
\end{aligned}$$

We consider all possible scenarios for $(*)$:

- $\psi' \leq \psi_2$, $s' \leq s_2 + r(x_2)$:

$$\begin{aligned}
(*) &= q(x'|x_1, a_1) + q(x'|x_2, a_2) \\
&\geq q(x'|x_1, a_2) + q(x'|x_2, a_1) \\
&= \mathbb{I}(\psi' \leq \psi_1, \text{True})\mathbb{I}(s' \leq s_1 + r(x_1), \text{True})q(x'|x_1, a_2) \\
&\quad + \mathbb{I}(\psi' \leq \psi_2, \text{True})\mathbb{I}(s' \leq s_2 + r(x_2), \text{True})q(x'|x_2, a_1) \\
&= (**)
\end{aligned}$$

- $\psi' \leq \psi_1$, $s_2 + r(x_2) < s' \leq s_1 + r(x_1)$ or $\psi_2 < \psi' \leq \psi_1$, $s' \leq s_1 + r(x_1)$:

$$\begin{aligned}
(*) &= q(x'|x_1, a_1) \\
&\geq q(x'|x_1, a_2) \\
&= \mathbb{I}(\psi' \leq \psi_1, \text{True})\mathbb{I}(s' \leq s_1 + r(x_1), \text{True})q(x'|x_1, a_2) \\
&\quad + \mathbb{I}(\psi' \leq \psi_2, \text{True})\mathbb{I}(s' \leq s_2 + r(x_2), \text{True})q(x'|x_2, a_1) \\
&= (**)
\end{aligned}$$

- $s_1 + r(x_1) < s'$ or $\psi_1 < \psi'$:

$$\begin{aligned}
(*) &= 0 \\
&= \mathbb{I}(\psi' \leq \psi_1, \text{True})\mathbb{I}(s' \leq s_1 + r(x_1), \text{True})q(x'|x_1, a_2) \\
&\quad + \mathbb{I}(\psi' \leq \psi_2, \text{True})\mathbb{I}(s' \leq s_2 + r(x_2), \text{True})q(x'|x_2, a_1) \\
&= (**)
\end{aligned}$$

Hence, based on Proposition 10, we have that for any $\psi_1 \geq \psi_2$ the optimal policy of the *doubly augmented arm* MDP is monotone in ψ. We conclude that the passive set is nondecreasing in λ confirming that the process is indexable.

C WHITTLE INDEX COMPUTATION ALGORITHMS

The algorithms are presented in Algorithms 2 and 3. For simplicity, we drop the superscript i in the pseudo-code.

The computational complexity of the Whittle index calculation is primarily determined by the backward induction over T time steps for a Markov decision process with two actions. The state space consists of two dimensions $\mathcal{X}$ and $\mathcal{S}$ with transitions governed by a Markov matrix P of size $|\mathcal{X}| \times |\mathcal{X}|$. As already defined, the action space is denoted by $\mathcal{A}$. Thus, the total complexity is $O(|\mathcal{X}|^2|\mathcal{S}||\mathcal{A}|T)$ for each specific value of λ. Additionally, the critical penalty search contributes a logarithmic factor of $O(\log((\lambda_u - \lambda_l)/\epsilon))$. Thus, the overall complexity of the algorithm is $O\left(|\mathcal{X}|^2|\mathcal{S}||\mathcal{A}|T \log((\lambda_u - \lambda_l)/\epsilon)\right)$.

Algorithm 2 Whittle Index Calculation

Input: $\texttt{LB}$ (lower bound), $\texttt{UB}$ (upper bound), ϵ (tolerance level)
$\lambda_l \leftarrow \texttt{LB}$, $\mathcal{W}_{\lambda_l} = \emptyset$.
while $\mathcal{W}_{\lambda_l} \neq \mathcal{X} \times \mathcal{S} \times \mathcal{T}$ **do**
  Compute $\lambda_c, \pi^*_{\lambda_c} = \Lambda^\epsilon_c(\lambda_l, \texttt{UB})$ using Algorithm 3.
  $\mathcal{M} = \{(s, x, t) : \pi_{\lambda_l}(s, x, t) = 1 \text{ and } \pi_{\lambda_c}(s, x, t) = 0\}$.
  Set $\hat{w}(x, s, t) \leftarrow \lambda_c$ for all $(x, s, t) \in \mathcal{M}$.
  Update $\lambda_l \leftarrow \lambda_c$, $\pi_{\lambda_l} \leftarrow \pi_{\lambda_c}$, $\mathcal{W}_{\lambda_l} \leftarrow \mathcal{W}_{\lambda_l} \cup \mathcal{M}$.
end while
return $\hat{w}$.

Algorithm 3 Critical Penalty Finder: $\Lambda^\epsilon_c(\lambda, \texttt{UB})$

Input: λ (current lower bound), $\texttt{UB}$ (upper bound), ϵ (tolerance level)
$\lambda_l \leftarrow \lambda$, $\lambda_u \leftarrow \texttt{UB}$.
if $\pi^*_{\lambda_l}$ equals $\pi^*_{\lambda_u}$ **then**
  return $\lambda_u, \pi^*_{\lambda_u}$.
else
  while $|\lambda_u - \lambda_l| \geq \epsilon$ **do**
    Compute midpoint $\lambda_c \leftarrow (\lambda_l + \lambda_u)/2$.
    if $\pi^*_{\lambda_c}$ equals $\pi^*_{\lambda_l}$ **then**
      Update $\lambda_l \leftarrow \lambda_c$.
    else
      Update $\lambda_u \leftarrow \lambda_c$.
    end if
  end while
  return $\lambda_u, \pi^*_{\lambda_u}$.
end if

D PROOF OF LEMMA 7

We follow the steps provided in Akbarzadeh and Mahajan (2023) to prove the lemma.

First, we state some basic properties of Thompson sampling algorithm.

Lemma 11 (Lemma 1 of Russo and Van Roy (2014))**.** *Suppose the true parameters $\theta^\star$ and the estimated ones θ_k have the same distribution given the same history $\mathcal{H}$. For any $\mathcal{H}$-measurable function f, we have* $\mathbb{E}[f(\theta^\star)|\mathcal{H}] = \mathbb{E}[f(\theta_k)|\mathcal{H}]$.

With an abuse of notation, let use $N^i_k(x^i, a^i)$ and $N^i_k(x^i, a^i, x'^i)$ instead of $N^i_{k,0}(x^i, a^i)$ and $N^i_{k,0}(x^i, a^i, x'^i)$, respectively.

Let

$$\hat{P}^i_k(x'^i|x^i, a^i) = \begin{cases} \bar{p}^i_0(x'^i|x^i, a^i), & \text{if } N^i_k(x^i, a^i) = 0, \\ N^i_k(x^i, a^i, x'^i)/N^i_k(x^i, a^i), & \text{otherwise,} \end{cases}$$

denote the sample mean estimation of $P^i(x'^i|x^i,a^i)$ based on observations up to end of episode k. For the ease of notation, for a given $\delta \in (0,1)$, we define

$$\epsilon^i_\delta(\ell) = \begin{cases} \sqrt{\frac{2|\mathcal{X}|^i \log(1/\delta)}{\ell}} & \text{if } \ell \geq 1 \\ \sqrt{2|\mathcal{X}|^i \log(1/\delta)} & \text{if } \ell = 0 \end{cases}. \tag{4}$$

Lemma 12. *Consider any arm i, episode k, $\delta \in (0,1)$, $\ell > 1$ and state-action pair (x^i,a^i). Define events $\mathcal{E}^i_{k,\ell}(x^i,a^i) = \{N^i_k(x^i,a^i) = \ell\}$ and $\mathcal{F}^{\star i}_k(x^i,a^i) = \{\|P^{\star i}(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\|_1 \leq \epsilon_\delta(N^i_k(x^i,a^i))\}$, and $\mathcal{F}^i_k(x^i,a^i) = \{\|P^i_k(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\|_1 \leq \epsilon_\delta(N^i_k(x^i,a^i))\}$. Then, we have*

$$\mathbb{P}\Big(\big\|P^{\star i}(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\big\|_1 > \epsilon^i_\delta(\ell) \;\Big|\; \mathcal{E}^i_{k,\ell}(x^i,a^i)\Big) \leq \delta, \tag{5}$$

$$\mathbb{P}\Big(\big\|P^i_k(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\big\|_1 > \epsilon^i_\delta(\ell) \;\Big|\; \mathcal{E}^i_{k,\ell}(x^i,a^i)\Big) \leq \delta, \tag{6}$$

where $\epsilon^i_\delta(\ell)$ is given by (4). The above inequalities imply that

$$\mathbb{E}\Big[\big\|P^{\star i}(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\big\|_1\Big] \leq \mathbb{E}[\epsilon^i_\delta(N^i_k(x^i,a^i))|\mathcal{F}^{\star i}(x^i,a^i)] + 2\delta, \tag{7}$$

$$\mathbb{E}\Big[\big\|P^i_k(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\big\|_1\Big] \leq \mathbb{E}[\epsilon^i_\delta(N^i_k(x^i,a^i))|\mathcal{F}^{\star i}_k(x^i,a^i)] + 2\delta. \tag{8}$$

Proof. Given arm i, and the state-action pair (x^i,a^i) of the arm, we know from Weissman et al. (2003) that for any $\varepsilon > 0$, the $L1$-deviation between the true distribution and the empirical distribution over the finite state space $\mathcal{X}^i$ given $N^i_k(x^i,a^i) = \ell$ samples is bounded by

$$\mathbb{P}\Big(\|P^{\star i}(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\|_1 \geq \varepsilon \;\Big|\; \mathcal{E}^i_{k,\ell}\Big) \leq 2^{|\mathcal{X}^i|}\exp\Big(-\frac{\ell\varepsilon^2}{2}\Big) < \exp\Big(|\mathcal{X}^i| - \frac{\ell\varepsilon^2}{2}\Big).$$

Therefore, setting $\delta = \exp(|\mathcal{X}^i| - \ell\varepsilon^2/2)$, we get

$$\mathbb{P}\left(\big\|P^{\star i}(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\big\|_1 > \sqrt{\frac{2(|\mathcal{X}^i| + \log(1/\delta))}{1 \vee \ell}} \;\Big|\; \mathcal{E}^i_{k,\ell}\right) \leq \delta.$$

This proves (5). Eq. (6) follows from the Thompson sampling Lemma (Lemma 11).

To prove (7) and (8), we first show

$$\begin{aligned} P((\mathcal{F}^{\star i}_k(x^i,a^i))^c) &= \mathbb{P}\Big(\big\|P^{\star i}(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\big\|_1 > \epsilon^i_\delta(N^i_k(x^i,a^i))\Big) \\ &= \sum_{\ell=0}^{\infty} \mathbb{P}\Big(\big\|P^{\star i}(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\big\|_1 > \epsilon^i_\delta(\ell) \;\Big|\; \mathcal{E}^i_{k,\ell}\Big)\mathbb{P}(\mathcal{E}^i_{k,\ell}) \\ &\leq \sum_{\ell=0}^{\infty} \delta\mathbb{P}(\mathcal{E}^i_{k,\ell}) = \delta. \end{aligned} \tag{9}$$

Now consider

$$\begin{aligned} &\mathbb{E}\Big[\|P^{\star i}(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\|_1\Big] = \mathbb{E}\Big[\big\|P^{\star i}(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\big\|_1 \;\Big|\; \mathcal{F}^{\star i}_k(x^i,a^i)\Big]\mathbb{P}(\mathcal{F}^{\star i}_k(x^i,a^i)) \\ &\quad + \mathbb{E}\Big[\big\|P^{\star i}(\cdot\,|x^i,a^i) - \hat{P}^i_k(\cdot\,|x^i,a^i)\big\|_1 \;\Big|\; (\mathcal{F}^{\star i}_k(x^i,a^i))^c\Big]\mathbb{P}((\mathcal{F}^{\star i}_k(x^i,a^i))^c) \\ &\quad \overset{(a)}{\leq} 2P((\mathcal{F})^c) + \mathbb{E}[\epsilon^i_\delta(N^i_k(x^i,a^i))|\mathcal{F}^{\star i}_k(x^i,a^i)] \\ &\quad \overset{(b)}{\leq} 2\delta + \mathbb{E}[\epsilon^i_\delta(N^i_k(x^i,a^i))|\mathcal{F}^{\star i}_k(x^i,a^i)], \end{aligned}$$

where (a) uses $\|\cdot\|_1 \leq 2$ and $P(\mathcal{F}) \leq 1$ and (b) uses (9). This proves (7). Eq. (8) follows from a similar argument. $\square$

Lemma 13. *Consider episode k, $\delta \in (0,1)$. Let denote the joint state-action pair of the learning policy for the system at time step t is $(\boldsymbol{X}_{k,t}, \boldsymbol{A}_{k,t})$. Define events $\mathcal{F}_k^{\star i}$ and $\mathcal{F}_k^i$ as in Lemma 12. Then we have*

$$\begin{aligned}\mathbb{E}\Big[\big\|\boldsymbol{P}^\star(\cdot|\boldsymbol{X}_{k,t}, \boldsymbol{A}_{k,t}) - \boldsymbol{P}_k(\cdot|\boldsymbol{X}_{k,t}, \boldsymbol{A}_{k,t})\big\|_1\Big] \\ \le 4N\delta + \sum_{i\in\mathcal{N}} \big(\mathbb{E}\big[\epsilon_\delta^i(N_k^i(X_{k,t}^i, A_{k,t}^i))|\mathcal{F}_k^{\star i}(X_{k,t}^i, A_{k,t}^i)\big] + \mathbb{E}\big[\epsilon_\delta^i(N_k^i(X_{k,t}^i, A_{k,t}^i))|\mathcal{F}_k^{i}(X_{k,t}^i, A_{k,t}^i)\big]\big).\end{aligned}$$

Proof. From (Jung et al., 2019, Lemma 13), we have

$$\mathbb{E}\Big[\big\|\boldsymbol{P}^\star(\cdot|\boldsymbol{x}, \boldsymbol{a}) - \boldsymbol{P}_k(\cdot|\boldsymbol{x}, \boldsymbol{a})\big\|_1\Big] \le \sum_{i\in\mathcal{N}} \mathbb{E}\Big[\big\|P^{\star i}(\cdot\,|x^i, a^i) - P_k^i(\cdot\,|x^i, a^i)\big\|_1\Big].$$

Then, the rest of the proof is as follows:

$$\begin{aligned}&\mathbb{E}\Big[\big\|\boldsymbol{P}^\star(\cdot|\boldsymbol{X}_{k,t}, \boldsymbol{A}_{k,t}) - \boldsymbol{P}_k(\cdot|\boldsymbol{X}_{k,t}, \boldsymbol{A}_{k,t})\big\|_1\Big] \\ &\quad\le \sum_{i=1}^N \mathbb{E}\Big[\big\|P^{\star i}(\cdot\,|X_{k,t}^i, A_{k,t}^i) - P_k^i(\cdot\,|X_{k,t}^i, A_{k,t}^i)\big\|_1\Big] \\ &\quad\le \sum_{i=1}^N \mathbb{E}\Big[\big\|P^{\star i}(\cdot\,|X_{k,t}^i, A_{k,t}^i) - \hat{P}_k^i(\cdot\,|X_{k,t}^i, A_{k,t}^i)\big\|_1 + \big\|P_k^i(\cdot\,|X_{k,t}^i, A_{k,t}^i) - \hat{P}_k^i(\cdot\,|X_{k,t}^i, A_{k,t}^i)\big\|_1\Big] \\ &\quad\le 4N\delta + \sum_{i=1}^N \mathbb{E}\big[\epsilon_\delta^i(N_k^i(X_{k,t}^i, A_{k,t}^i))|\mathcal{F}_k^{\star i}(X_{k,t}^i, A_{k,t}^i)\big] + \mathbb{E}\big[\epsilon_\delta^i(N_k^i(X_{k,t}^i, A_{k,t}^i))|\mathcal{F}_k^{i}(X_{k,t}^i, A_{k,t}^i)\big],\end{aligned}$$

where the second inequality follows from triangle inequality, and the third follows from Lemma 12. □

Finally, to prove the result of Lemma 7, we have the following.

$$\begin{aligned}&\sum_{k=1}^K\sum_{t=0}^{T-1} \mathbb{E}\Big[\big\|\boldsymbol{P}^\star(\cdot\,|\boldsymbol{X}_{k,t}, \boldsymbol{A}_{k,t}) - \boldsymbol{P}_k(\cdot\,|\boldsymbol{X}_{k,t}, \boldsymbol{A}_{k,t})\big\|_1\Big] \\ &\quad\le \sum_{k=1}^K\sum_{t=0}^{T-1} 4n\delta + \sum_{i\in\mathcal{N}} \big(\mathbb{E}\big[\epsilon_\delta^i(N_k^i(X_{k,t}^i, A_{k,t}^i))|\mathcal{F}_k^{\star i}(X_{k,t}^i, A_{k,t}^i)\big] + \mathbb{E}\big[\epsilon_\delta^i(N_k^i(X_{k,t}^i, A_{k,t}^i))|\mathcal{F}_k^{i}(X_{k,t}^i, A_{k,t}^i)\big]\big)\end{aligned}$$

For the first term, $\sum_{k=1}^K \sum_{t=0}^{T-1} 4N\delta = 4N\delta KT$.

For the second term, we follow the steps from Osband et al. (2013). Therefore, by definition of $\epsilon_\delta^i(\cdot)$, we have

$$\begin{aligned}&\sum_{k=1}^K\sum_{t=0}^{T-1}\sum_{i=1}^N \mathbb{E}\left[\sqrt{\frac{(4|\mathcal{X}|^i + \log(1/\delta))}{1\vee N_k^i(X_{k,t}^i, A_{k,t}^i)}}\;\middle|\; \mathcal{F}_k^{\star i}(X_{k,t}^i, A_{k,t}^i)\right] \qquad (10)\\ &\le \sum_{i=1}^N \sqrt{4\left(|\mathcal{X}|^i + \log(1/\delta)\right)} \sum_{k=1}^K\sum_{t=0}^{T-1} \mathbb{E}\left[\sqrt{\frac{1}{1\vee N_k^i(X_{k,t}^i, A_{k,t}^i)}}\;\middle|\; \mathcal{F}_k^{\star i}(X_{k,t}^i, A_{k,t}^i)\right] \\ &\le \sum_{i=1}^N \sqrt{4\left(|\mathcal{X}|^i + \log(1/\delta)\right)} \sum_{k=1}^K\sum_{t=0}^{T-1} \\ &\quad\left(\mathbb{1}(N_k^i(X_{k,t}^i, A_{k,t}^i) \le T) + \mathbb{1}(N_k^i(X_{k,t}^i, A_{k,t}^i) > T)\mathbb{E}\left[\sqrt{\frac{1}{1\vee N_k^i(X_{k,t}^i, A_{k,t}^i)}}\;\middle|\; \mathcal{F}_k^{\star i}(X_{k,t}^i, A_{k,t}^i)\right]\right).\end{aligned}$$

Let assume $(X_{k,t}^i, A_{k,t}^i) = (x, a)$. Consider $N_k^i(x,a) \le T$. This can happen fewer than $2T$ times per state action pair. Therefore, $\sum_{k=1}^K\sum_{t=0}^{T-1} \mathbb{1}(N_k^i(x,a) \le T) \le 2(T2|\mathcal{X}^i|) = 4T|\mathcal{X}^i|$.

Now, consider $N^i_k(x,a) > T$. Then, we have $N^i_{k,t}(x,a) \le N^i_k(x,a) + T \le 2N^i_k(x,a)$. Hence,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{t=0}^{T-1} \mathbb{1}(N^i_k(X^i_{k,t},A^i_{k,t}) > T)\mathbb{E}\Bigg[\sqrt{\frac{1}{1\vee N^i_k(X^i_{k,t},A^i_{k,t})}}\ \Bigg|\ \mathcal{F}^{\star i}_k(X^i_{k,t},A^i_{k,t})\Bigg] \\
&\le \sum_{k=1}^{K}\sum_{t=0}^{T-1} \mathbb{E}\Bigg[\sqrt{\frac{2}{N^i_{k,t}(X^i_{k,t},A^i_{k,t})}}\ \Bigg|\ \mathcal{F}^{\star i}_k(X^i_{k,t},A^i_{k,t})\Bigg] \le \sqrt{2}\sum_{(x^i,a^i)}\sum_{j=1}^{N^i_{K+1}(x^i,a^i)}\frac{1}{\sqrt{j}} \\
&\le \sqrt{4|\mathcal{X}^i|\sum_{(x^i,a^i)} N^i_{K+1}(x^i,a^i)} = \sqrt{4|\mathcal{X}^i|KT}.
\end{aligned}$$

Finally, we get

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{t=0}^{T-1}\sum_{i=1}^{N} \mathbb{E}\Bigg[\sqrt{\frac{\left(4|\mathcal{X}|^i+\log(1/\delta)\right)}{1\vee N^i_k(X^i_{k,t},A^i_{k,t})}}\ \Bigg|\ \mathcal{F}^{\star i}_k(X^i_{k,t},A^i_{k,t})\Bigg] &\le \sum_{i=1}^{N}\sqrt{4\left(|\mathcal{X}|^i+\log(1/\delta)\right)}\sqrt{4|\mathcal{X}^i|KT} \\
&= 4\sum_{i=1}^{N}\sqrt{\left(|\mathcal{X}|^i+\log(1/\delta)\right)|\mathcal{X}^i|KT} \\
&= 4\sum_{i=1}^{N}\sqrt{((|\mathcal{X}|^i)^2KT+\log(1/\delta)|\mathcal{X}^i|KT)} \\
&\le 4\sum_{i=1}^{N}|\mathcal{X}|^i\sqrt{KT(1+\log(1/\delta))} \\
&\le 4N|\bar{\mathcal{X}}|\sqrt{KT(1+\log(1/\delta))} \qquad (11)
\end{aligned}$$

The same approach works for the third term and we get

$$\sum_{k=1}^{K}\sum_{t=0}^{T-1}\sum_{i=1}^{N} \mathbb{E}\Bigg[\sqrt{\frac{4|\mathcal{X}|^i\log(1/\delta)}{1\vee N^i_k(X^i_{k,t},A^i_{k,t})}}\ \Bigg|\ \mathcal{F}^{i}_k(X^i_{k,t},A^i_{k,t})\Bigg] \le 4N|\bar{\mathcal{X}}|\sqrt{KT(1+\log(1/\delta))}. \qquad (12)$$

Finally, by setting $\delta = 1/(KT)$, and substituting the upper-bounds, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\boldsymbol{P}^\star(\cdot|\boldsymbol{X}_{k,t},\boldsymbol{A}_{k,t}) - \boldsymbol{P}_k(\cdot|\boldsymbol{X}_{k,t},\boldsymbol{A}_{k,t})\big\|_1\Big] \\
&\qquad\le \sum_{k=1}^{K}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\Big[\big\|P^{\star i}(\cdot\,|X^i_{k,t},A^i_{k,t}) - P^i_k(\cdot\,|X^i_{k,t},A^i_{k,t})\big\|_1\Big] \\
&\qquad\le 4N + 8N|\bar{\mathcal{X}}|\sqrt{KT(1+\log KT)} \le 12N|\bar{\mathcal{X}}|\sqrt{KT(1+\log KT)}
\end{aligned}$$

Hence, the first term can be upper-bounded by the second term.

E PROOF OF THEOREM 8

The technique used in Section 3.2 can also be applied to the overall system prior to the decomposition. Let construct a time-dependent MDP for the overall system as $(\bar{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{A}}, \{\tilde{\boldsymbol{P}}(\boldsymbol{a})\}_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}, \{\boldsymbol{r}_t\}_{t=0}^T, \boldsymbol{x}_0)$ with $\bar{\boldsymbol{\mathcal{X}}} := \boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{S}}$, $\tilde{\boldsymbol{P}}(\boldsymbol{x}',\boldsymbol{s}'|\boldsymbol{x},\boldsymbol{s},\boldsymbol{a}) := \boldsymbol{P}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{a})\mathbb{I}(\boldsymbol{s}',\boldsymbol{s}+\boldsymbol{r}(\boldsymbol{x},\boldsymbol{a}))$, $\boldsymbol{r}_t(\boldsymbol{x},\boldsymbol{s},\boldsymbol{a}) := \mathbb{I}(t,T-1)U_{\boldsymbol{\tau}}(\boldsymbol{s}+\boldsymbol{r}(\boldsymbol{x},\boldsymbol{a}))$ and $\bar{\boldsymbol{x}}_0 := (\boldsymbol{x}_{\boldsymbol{0}},\boldsymbol{0})$.

The value function under optimal policy $\boldsymbol{\pi}^\star$ for the overall system with the true parameter set $\boldsymbol{\theta}^\star$ is $\boldsymbol{V}^{\boldsymbol{\pi}^\star,\boldsymbol{\theta}^\star}_{t-1} = \langle\tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^\star,\boldsymbol{\theta}^\star}, \boldsymbol{V}^{\boldsymbol{\pi}^\star,\boldsymbol{\theta}^\star}_t\rangle$. The value function under the estimated policy $\boldsymbol{\pi}^k$ for the overall system with the estimated parameter set $\boldsymbol{\theta}^k$ is $\boldsymbol{V}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}_{t-1} = \langle\tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}, \boldsymbol{V}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}_{k,t}\rangle$. And the value function under the estimated policy $\boldsymbol{\pi}^k$ for the overall system with the true parameter set $\boldsymbol{\theta}^\star$ is $\boldsymbol{V}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}_{t-1} = \langle\tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}, \boldsymbol{V}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}_{k,t}\rangle$.

Note that in the definition of the regret, the objectives are obtained using the parameters of the true MDP but the policy is the estimated one. Hence, we have

$$\begin{aligned}
\mathcal{R}(K) &= \sum_{k=1}^{K} \mathbb{E}\Big[\mathbb{E}\left[\boldsymbol{D}_{\boldsymbol{x}_0}(\boldsymbol{\pi}^*)\right] - \mathbb{E}\left[\boldsymbol{D}_{\boldsymbol{x}_0}(\boldsymbol{\pi}_k)\right]\Big] \\
&= \sum_{k=1}^{K} \mathbb{E}\Big[\boldsymbol{V}_0^{\boldsymbol{\pi}^\star,\boldsymbol{\theta}^\star}(\boldsymbol{x}_0, \boldsymbol{s}_0) - \boldsymbol{V}_0^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}(\boldsymbol{x}_0, \boldsymbol{s}_0)\Big] \\
&= \sum_{k=1}^{K} \mathbb{E}\Big[\boldsymbol{V}_0^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}(\boldsymbol{x}_0, \boldsymbol{s}_0) - \boldsymbol{V}_0^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}(\boldsymbol{x}_0, \boldsymbol{s}_0)\Big] \qquad (13)
\end{aligned}$$

where the last equality holds by Lemma 11. Then, we present the main result of this section.

From Eq. (13), by adding and subtracting $\langle \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}\rangle$ we get

$$\begin{aligned}
\mathcal{R}(K) &= \sum_{k=1}^{K} \mathbb{E}\Big[\boldsymbol{V}_0^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}(\boldsymbol{x}_0, \boldsymbol{s}_0) - \boldsymbol{V}_0^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}(\boldsymbol{x}_0, \boldsymbol{s}_0)\Big] \\
&= \sum_{k=1}^{K} \mathbb{E}\Big[\langle \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}\rangle - \langle \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\rangle \\
&\qquad + \langle \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}\rangle - \langle \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}\rangle\Big] \\
&= \sum_{k=1}^{K} \mathbb{E}\Big[\langle (\tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star})\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}\rangle + \langle \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\rangle\Big].
\end{aligned}$$

Then, by adding and subtracting $\left(\boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\right)(\boldsymbol{X}_{k,1}, \boldsymbol{S}_{k,1})$, we get

$$\begin{aligned}
\mathcal{R}(K) &= \sum_{k=1}^{K} \mathbb{E}\Big[\langle (\tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star})\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}\rangle + \langle \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\rangle \\
&\qquad + \left(\boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\right)(\boldsymbol{X}_{k,1}, \boldsymbol{S}_{k,1}) - \left(\boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\right)(\boldsymbol{X}_{k,1}, \boldsymbol{S}_{k,1})\Big] \\
&= \sum_{k=1}^{K} \mathbb{E}\Big[\langle (\tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star})\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}\rangle + \left(\boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\right)(\boldsymbol{X}_{k,1}, \boldsymbol{S}_{k,1})\Big]
\end{aligned}$$

where we have used the fact that for any arbitrary policy $\boldsymbol{\pi}^k$,

$$\begin{aligned}
&\mathbb{E}\Big[\langle \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}\rangle\Big] = \mathbb{E}[\boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}(\boldsymbol{X}_{k,1}, \boldsymbol{S}_{k,1})], \\
&\mathbb{E}\Big[\langle \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\rangle\Big] = \mathbb{E}[\boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}(\boldsymbol{X}_{k,1}, \boldsymbol{S}_{k,1})] \qquad (14)
\end{aligned}$$

and hence,

$$\mathbb{E}\Big[\langle \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\mathbf{1}_{\{\boldsymbol{x}_0,\boldsymbol{s}_0\}}, \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\rangle - \left(\boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \boldsymbol{V}_1^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}\right)(\boldsymbol{X}_{k,1}, \boldsymbol{S}_{k,1})\Big] = 0.$$

Now, recursively, the second inner term of (14) can be decomposed until the end of the finite time horizon. Therefore, for a horizon of length T, we have

$$\begin{aligned}
\mathcal{R}(K) &= \sum_{k=1}^{K} \mathbb{E}\Big[\boldsymbol{V}_0^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}(\boldsymbol{x}_0, \boldsymbol{s}_0) - \boldsymbol{V}_0^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star}(\boldsymbol{x}_0, \boldsymbol{s}_0)\Big] \\
&= \sum_{k=1}^{K} \sum_{t=0}^{T-1} \mathbb{E}\Big[\langle (\tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k} - \tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star})\mathbf{1}_{\{\boldsymbol{X}_{k,t},\boldsymbol{S}_{k,t}\}}, \boldsymbol{V}_{k,t+1}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}\rangle\Big].
\end{aligned}$$

As the per-step reward is upper-bounded by $R_{\max}$, then $|\boldsymbol{V}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}_{k,t+1}(\boldsymbol{X}_{k,t},\boldsymbol{S}_{k,t})| \leq NTR_{\max}$ for any t,

$$\begin{aligned}
\mathcal{R}(K) &\leq N\sum_{k=1}^{K}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|(\tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^k}-\tilde{\boldsymbol{P}}^{\boldsymbol{\pi}^k,\boldsymbol{\theta}^\star})\mathbf{1}_{\{\boldsymbol{X}_{k,t},\boldsymbol{S}_{k,t}\}}\big\|_1\Big] \\
&= N\sum_{k=1}^{K}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|(\boldsymbol{P}_k^{\boldsymbol{\pi}^k}-\boldsymbol{P}^{\star\boldsymbol{\pi}^k})\mathbf{1}_{\{\boldsymbol{X}_{k,t}\}}\big\|_1\Big] \\
&\leq N\sum_{k=1}^{K}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\boldsymbol{P}^\star(\cdot\,|\boldsymbol{X}_{k,t},\boldsymbol{A}_{k,t})-\boldsymbol{P}_k(\cdot\,|\boldsymbol{X}_{k,t},\boldsymbol{A}_{k,t})\big\|_1\Big] \\
&\leq 12N^2TR_{\max}|\bar{\mathcal{X}}|\sqrt{KT(1+\log KT)}.
\end{aligned}$$

where the last inequality holds by Lemma 7.

F SIMPLE ILLUSTRATIVE MODELS THAT SATISFY ASSUMPTIONS OF THEOREM 5

A class of models which satisfy the assumptions of Theorem 5 is presented below: The state space for each arm is sorted from the worst to the best state. Per-step reward is only a function of the state for each arm and is nondecreasing over the state space. The transition probabilities follow any of the following pairs:

1. Given an arm with n states and a parameter $p \in [0,1]$, let the transition probability matrix under passive action be:

$$\mathcal{P}_0(p) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 1-p & p & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1-p & \dots & 0 & 0 & 0 & 0 & 0 & p & 0 \\ 1-p & \dots & 0 & 0 & 0 & 0 & 0 & 0 & p \end{bmatrix}.$$

   And let the transition probability matrix under active action be:

$$\mathcal{P}_1(p) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

2. Given an arm with n states and parameter $p1 > p2$, let the transition probability matrix under passive action be:

$$\mathcal{P}_0(p) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 1-p2 & p2 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1-p2 & \dots & 0 & 0 & 0 & 0 & 0 & p2 & 0 \\ 1-p2 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & p2 \end{bmatrix}.$$

   And let the transition probability matrix under active action be:

$$\mathcal{P}_1(p) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 1-p1 & p1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1-p1 & \dots & 0 & 0 & 0 & 0 & 0 & p1 & 0 \\ 1-p1 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & p1 \end{bmatrix}.$$

3. Given an arm with n states and a parameter $p \in [0, 0.5]$, let the transition probability matrix under passive action be:

$$\mathcal{P}_0(p) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 1-p & p & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & 1-p & p & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 1-p & p \end{bmatrix}.$$

   And let the transition probability matrix under active action be:

$$\mathcal{P}_1(p) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ p & 1-p & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & p & 1-p & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 1-p & p \end{bmatrix}.$$

4. Given an arm with n states and a parameter $p \in [0, 1/(n-1)]$, let the transition probability matrix under passive action be:

$$\mathcal{P}_0(p) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 1-(n-1)p & (n-1)p & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 1-(n-1)p & p & (n-2)p & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1-(n-1)p & \dots & p & p & p & p & p & 2p & 0 \\ 1-(n-1)p & \dots & p & p & p & p & p & p & p \end{bmatrix}.$$

   And let the transition probability matrix under passive action be:

$$\mathcal{P}_1(p) = \begin{bmatrix} (n-1)p & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1-(n-1)p \\ 0 & (n-2)p & 0 & 0 & 0 & 0 & 0 & \dots & 1-(n-2)p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & p & 1-p \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

We would like to highlight that how the above transition matrices (specifically No. 4) can be used in a machine maintenance problem instance. In the machine maintenance problem, the transition matrices $\mathcal{P}_0(p)$ and $\mathcal{P}_1(p)$ are designed to model the system's state dynamics under two actions: do nothing (0) and perform maintenance (1).

For $\mathcal{P}_0(p)$, the "do nothing" action, the machine tends to remain in its current state or transition to a worse state due to natural degradation. The probability of transitioning to a better state is zero, while the probability of staying in the same state decreases as the machine condition worsens, reflecting the likelihood of continued degradation.

For $\mathcal{P}_1(p)$, the "maintenance" action, the machine improves its condition or stays in the same state. The best state has the highest probability of retention (1), reflecting that once the machine is fully repaired, it cannot be degraded. The probability of transitioning to better states increases as maintenance is applied, highlighting the beneficial effects of the action.