On learning Whittle index policy for restless bandits with scalable regret

Nima Akbarzadeh, Aditya Mahajan

***Abstract*— Reinforcement learning is an attractive approach to learn good resource allocation and scheduling policies based on data when the system model is unknown. However, the cumulative regret of most RL algorithms scales as $\tilde{O}(S\sqrt{AT})$, where S is the size of the state space, A is the size of the action space, T is the horizon, and the $\tilde{O}(\cdot)$ notation hides logarithmic terms. Due to the linear dependence on the size of the state space, these regret bounds are prohibitively large for resource allocation and scheduling problems. In this paper, we present a model-based RL algorithm for such problems which has scalable regret. In particular, we consider a restless bandit model, and propose a Thompson-sampling based learning algorithm which is tuned to the underlying structure of the model. We present two characterizations of the regret of the proposed algorithm with respect to the Whittle index policy. First, we show that for a restless bandit with n arms and at most m activations at each time, the regret scales either as $\tilde{O}(mn\sqrt{T})$ or $\tilde{O}(n^2\sqrt{T})$ depending on the reward model. Second, under an additional technical assumption, we show that the regret scales as $\tilde{O}(n^{1.5}\sqrt{T})$ or $\tilde{O}(\max\{m\sqrt{n}, n\}\sqrt{T})$. We present numerical examples to illustrate the salient features of the algorithm.**

***Index Terms*— Restless bandits, Thompson sampling, reinforcement learning, Whittle index**

I. Introduction

Resource allocation and scheduling problems arise in control of networked systems. Examples include opportunistic scheduling in networks [1]–[3], link scheduling in machine type communication [4], user allocation in mmWave networks [5], channel allocation in networks [6], source selection in peer-to-peer networks [7], opportunistic spectrum access [8]–[10], demand response in smart grids [11], [12], dynamic routing in multi-UAVs [13], operator allocation in multi-robot systems [14], etc.

Due to the curse of dimensionality, finding an optimal solution in such resource allocation and scheduling problems is computationally prohibitive [15]. Restless bandits (RBs) [16] have emerged as a popular solution heuristic for such problems. The RB framework is motivated by the *rested* multi-armed bandit problem considered in the seminar work of Gittins [17], who showed that the optimal strategy for the rested multi-armed bandit problem is of the *index type*: one can compute an index for each state of each alternative (also called an arm), and choose the alternative with the highest index. In general, such index-type policies are not optimal for RBs. In fact, computing the optimal policy for RBs is PSPACE hard [15]. However, as argued in [16], an index-type policy (now known as the Whittle index) can be a useful heuristic if a technical condition known as indexability is satisfied. The Whittle index policy is optimal for some specific models [6], [17], [18]. There is also a strong empirical evidence to suggest that the Whittle index policy performs close to optimal in various settings [19]–[24]. For these reasons, the RB framework has been applied in a variety of resource allocation and scheduling applications referenced above.

In all the above references, it is assumed that the system model is known perfectly. In many real-world applications, there is often uncertainty about the system model. In such situations, RL (reinforcement learning) is an attractive alternative. In recent years, there are many papers which investigate RL for RBs [25]–[29]. Most of these learn the Q-function associated with the average reward/cost optimality equation parameterized by the activation cost λ and use it to asymptotically learn the Whittle index.

A common measure of performance of an RL algorithm is *regret*, which measures the difference in performance of the learning algorithm that doesn't a priori know the system model with the performance of a baseline policy that knows the model. However, the regret is not characterized in the existing literature on RL for RBs [25]–[29].

There are some results on characterizing regret for some specific instances of RBs: a model of multi-class queues arising in mobile edge computing [30] and a model for scheduling when to observe uncontrolled Markov chains arising in opportunistic spectrum access in cognitive radios [31]–[36]. However, the regret analyses in these papers exploit specific features of the model and are not applicable to general models. The main contribution of this paper is to characterize the regret of a general RL algorithm for general RBs.

It is not possible to directly use existing RL algorithms that achieve near optimal regret in RBs. To explain why this is the case, we provide a short overview of characterizing regret in RL. Consider a general MDP (Markov decision process) with finite state space of size S and finite action space of size A. It is shown in [37] that no learning algorithm can achieve a regret of less than $\tilde{\Omega}(\sqrt{SADT})$, where D is the diameter of the underlying MDP and T is the time horizon for which the system runs. Several classes of algorithms have been proposed

This research was funded in part by the Innovation for Defence Excellence and Security (IDEaS) Program of the Canadian Department of National Defence through grant CFPMN2-037, and Fonds de Recherche du Quebec-Nature et technologies (FRQNT).

Nima Akbarzadeh and Aditya Mahajan are with Department of Electrical and Computer Engineering, McGill University, Montreal. (nima.akbarzadeh@mail.mcgill.ca, aditya.mahajan@mcgill.ca)

TABLE I: A comparison of the regret bounds of various algorithms

Algorithm	Algorithm Type	Regret[a]	Regret Type
UCRL2 [37]	OUU	$\tilde{\mathcal{O}}(D\mathsf{S}\sqrt{\mathsf{AT}})$	Frequentist
REGAL [38]	OUU	$\tilde{\mathcal{O}}(H\mathsf{S}\sqrt{\mathsf{AT}})$	Frequentist
SCAL [39]	OUU	$\tilde{\mathcal{O}}(H\sqrt{\Gamma\mathsf{SAT}})$	Frequentist
[40]	TS	$\tilde{\mathcal{O}}(D\sqrt{\mathsf{SAT}})$[b]	Frequentist
[41]	TS	$\tilde{\mathcal{O}}(\sqrt{H\mathsf{SAT}})$	Frequentist
TSDE [42]	TS	$\tilde{\mathcal{O}}(H\mathsf{S}\sqrt{\mathsf{AT}})$	Bayesian

[a] In the column on regret bounds, Γ is the maximum number of states that can be reached from any state, D is the diameter of the MDP, H is the span of the bias of the MDP. These are related as $\Gamma \leq S$ and $H \leq D$ (established in [38]).

[b] It is pointed out in [41] that there is a mistake in the proof in [40] and it is suggested that the bound of [40] may be loose by a factor of $\sqrt{\mathsf{S}}$.

in the literature which achieve this lower bound up to a factor of $\sqrt{\mathsf{S}}$ and logarithmic terms. Broadly speaking, these regret optimal RL algorithms fall into two classes: optimism-under-uncertainty (OUU) and Thompson sampling (TS). Two types of regret bounds are provided: frequentist regret, which is a bound on the worst case regret with high probability and Bayesian regret, which is a bound on the average regret (with respect to a pre-specified prior). A summary of the regret bounds for various algorithms is shown in Table I.

Each of these state-of-the-art algorithms has a regret that scales approximately as $\tilde{\mathcal{O}}(\mathsf{S}\sqrt{\mathsf{AT}})$, which is prohibitively large when translated to the RB setting for reasons explained below. Consider a RB with n arms where at most m arms can be activated at a time. Let S_i denote the size of the state space of arm $i \in \{1, \dots, n\}$. Such an RB can be modeled as an MDP where the size of the state space is $\prod_{i=1}^{n} \mathsf{S}_i$ and the size of the action space is $\binom{n}{m}$. Thus, the regret of using any of the algorithms described in Table I on RBs will be $\tilde{\mathcal{O}}\left(\prod_{i=1}^{n} \mathsf{S}_i \sqrt{\binom{n}{m} T}\right)$, which grows exponentially with the number n of arms. In this paper, we provide a more nuanced characterization of the scaling of the regret with the number of alternatives.

In particular, in this paper we propose a Thompson-sampling based learning algorithm for RB, which we call as `RB-TSDE`. This algorithm is inspired from the TSDE (Thompson sampling with dynamic episodes) algorithm [42]. We show that for a RB with n arms where m of them can be chosen at a time, `RB-TSDE` has a Bayesian regret (with respect to the Whittle index policy with known dynamics) of $\tilde{\mathcal{O}}(n^2\sqrt{T})$ or $\tilde{\mathcal{O}}(nm\sqrt{T})$ depending on the assumptions on the per-step reward. Under an additional technical assumption, we obtain an alternative regret bound of $\tilde{\mathcal{O}}(n^{1.5}\sqrt{T})$ or $\tilde{O}(\max\{m\sqrt{n}, n\}\sqrt{T})$.

The rest of the paper is organized as follows. In Sec. II, we formulate the learning problem for RB when the state transition probabilities of all arms are unknown and present the main results. In Sec. III, we present the Thompson sampling with dynamic episodes for RB and provide an upper bound on the regret. In Sec. IV, we provide the proof outline and defer the details to the Appendix. In Sec. V, we demonstrate a numerical example of the regret of our algorithm. In Sec. VI, we discuss relaxation and sufficient conditions of some of the assumptions, in addition to comparison with the optimal policy. Finally, we conclude in Sec. VII.

Notation.: We use upper case variables S, A, etc., to denote random variables, the corresponding lower variables (s, a, etc.) to denote their realizations, and corresponding calligraphic letters ($\mathcal{S}$, $\mathcal{A}$, etc.) to denote set of realizations. Subscripts denote time and superscript denotes arm. Thus S_t^i is the state of arm i at time t. Bold letters denote collection of variables across all arms. Thus, $\boldsymbol{S}_t = (S_t^1, \dots, S_t^n)$ is the set of states of all arms at time t. $S_{0:t}$ is a shorthand for $(S_0, \dots, S_t)$. We use $\mathbb{E}[\cdot]$ to denote expectation of a random variable, $\mathbb{P}(\cdot)$ to denote probability of an event and $\mathbb{1}(\cdot)$ to denote the indicator of an event.

For a given function $f\colon \mathcal{X} \to \mathbb{R}$, the span-norm of f is defined as $\operatorname{span}(f) = \max_{x\in\mathcal{X}} f(x) - \min_{x\in\mathcal{X}} f(x)$. Given two metric spaces $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$, the Lipschitz constant of function $f : \mathcal{X} \to \mathcal{Y}$ is defined by

$$L_f = \sup_{\substack{x_1,x_2\in\mathcal{X} \\ x_1\neq x_2}} \frac{d_Y(f(x_1), f(x_2))}{d_X(x_1, x_2)}.$$

Let ζ_1 and ζ_2 denote probability measures on $(\mathcal{X}, d_X)$. Then, the Kantorovich distance between them is defined as

$$\mathcal{K}(\zeta_1, \zeta_2) = \sup_{f: L_f \leq 1} \left| \sum_{x\in\mathcal{X}} f(x)\zeta_1(x) - \sum_{x\in\mathcal{X}} f(x)\zeta_2(x) \right|.$$

Given a metric space $(\mathcal{X}, d_X)$, $\operatorname{diam}(\mathcal{X}) = \sup\{d_X(x_1, x_2) : x_1, x_2 \in \mathcal{X}\}$ denotes the diameter of the set $\mathcal{X}$.

II. Model, Problem Formulation and Results

A. Restless Bandits

Restless bandits (RB) are a class of resource allocation problems where, at each time instant, a decision maker has to select m out of n available alternatives. Each alternative, which is also called an arm, is a controlled Markov process $\langle \mathcal{S}^i, \mathcal{A}^i = \{0, 1\}, P^i, r^i \rangle$ where $\mathcal{S}^i$ is the state space, $\mathcal{A}^i = \{0, 1\}$ is the action space, $P^i : \mathcal{S}^i \times \mathcal{A}^i \to \Delta(\mathcal{S}^i)$ is the controlled transition matrix, and $r^i : \mathcal{S}^i \times \mathcal{A}^i \to [0, R_{\max}]$ is the per-step reward. The action $A_t^i = 1$ means the decision maker selects arm i at time t. The arms for which $A_t^i = 1$ are called *active* arms and the arms for which $A_t^i = 0$ are called *passive* arms.

Let $\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{S}}^1 \times \cdots \times \boldsymbol{\mathcal{S}}^n$ denote the joint state space and $\boldsymbol{\mathcal{A}}(m) = \left\{\boldsymbol{a} \in \{0,1\}^n : \|\boldsymbol{a}\|_1 = m\right\}$ denote the feasible action space. We assume that the initial state $\boldsymbol{S}_0 = (S_0^1, \dots, S_0^n)$ is a random variable which is independent across arms and has a known initial distribution. Moreover, the arms evolve independently, i.e., for any $\boldsymbol{s}_{0:t} = (s_{0:t}^1, \dots, s_{0:t}^n)$ and $\boldsymbol{a}_{0:t} = (a_{0:t}^1, \dots, a_{0:t}^n)$, we have

$$\mathbb{P}\left(\boldsymbol{S}_{t+1} = \boldsymbol{s}_{t+1} | \boldsymbol{S}_{0:t} = \boldsymbol{s}_{0:t}, \boldsymbol{A}_{0:t} = \boldsymbol{a}_{0:t}\right) = \prod_{i=1}^{n} P^i\left(s_{t+1}^i | s_t^i, a_t^i\right) := \boldsymbol{P}(\boldsymbol{s}_{t+1} | \boldsymbol{s}_t, \boldsymbol{a}_t).$$

We consider two reward models:

- **Model A:** All arms, whether active or not, yield rewards, i.e., the aggregated per-step reward is given by $\boldsymbol{r}(\boldsymbol{s}_t, \boldsymbol{a}_t) = \sum_{i=1}^{n} r^i(s_t^i, a_t^i)$.

- **Model B:** Only the activated arms yield rewards. The state of the passive arms evolves, but the arms do not yield reward. Thus, the aggregated per-step reward is given by $\boldsymbol{r}(\boldsymbol{s}_t, \boldsymbol{a}_t) = \sum_{i=1}^{n} r^i(s_t^i, a_t^i)\mathbb{1}(\{a_t^i = 1\})$.

Note that Model B is same as Model A under the assumption that $r^i(\cdot, 0) = 0$ for all arms $i \in [n]$. For that reason, for most of the paper, we will take $\boldsymbol{r}(\boldsymbol{s}_t, \boldsymbol{a}_t) = \sum_{i=1}^{n} r^i(s_t^i, a_t^i)$ and assume $r^i(\cdot, 0) = 0$ when specializing for results of Model B.

Remark 1: Both Models A and B arise in different applications. Examples of Model A include queuing networks [2], where all queues incur holding cost; and machine maintenance [21], where all machines incur a cost when run in a faulty state. Examples of Model B include cognitive radios [8], where the reward depends only on the state of the selected channels.

Let $\boldsymbol{\Pi}$ denote the family of all possible (potentially history dependent and randomized) policies for the decision maker (who observes the state of all arms). The performance of any policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ is given by

$$J(\boldsymbol{\pi}) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{r}(\boldsymbol{S}_t, \boldsymbol{A}_t)\right], \tag{1}$$

where the expectation is taken with respect to the initial state distribution and the joint distribution induced on all system variables.

The objective of the decision maker is to choose a policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ to maximize the total expected reward $J(\boldsymbol{\pi})$. This objective is a MDP but computing an optimal policy using dynamic program suffers from the curse of dimensionality. For example, if $|\mathcal{S}^i| = \mathsf{S}$ for each $i \in [n]$, then $|\boldsymbol{\mathcal{S}}| = \mathsf{S}^n$ and $|\boldsymbol{\mathcal{A}}(m)| = \binom{n}{m}$. Then, the computational complexity of each step of value iteration is $|\boldsymbol{\mathcal{A}}(m)||\boldsymbol{\mathcal{S}}|^2 = \binom{n}{m}\mathsf{S}^{2n}$, which is prohibitively large for even moderate values of S and n. For this reason, most of the RB literature focuses on a computationally tractable but sub-optimal approach known as the Whittle index policy.

B. Whittle index policy

The Whittle index policy is motivated by the solution of a relaxation of the original optimization problem. Instead of the hard constraint of activating exactly m arms at a time, consider a relaxation where m arms have to be activated on average, i.e.,

$$\max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{r}(\boldsymbol{S}_t, \boldsymbol{A}_t)\right],$$
$$\text{s.t. } \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \|\boldsymbol{A}_t\|_1\right] = m. \tag{2}$$

Note that this relaxation is simply used to obtain a decomposition to define Whittle indices. The Whittle index policy, which is stated at the end of this section, picks exactly m arms at each time step.

The Lagrangian relaxation of (2) is given by

$$\max_{\lambda\geq 0} \max_{\boldsymbol{\pi}_\lambda\in\boldsymbol{\Pi}} \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\Big[\boldsymbol{r}(\boldsymbol{S}_t, \boldsymbol{A}_t) - \lambda\|\boldsymbol{A}_t\|_1\Big]\right]. \tag{3}$$

Note that the Lagrangian relaxation (3) is decoupled across arms because the per-step reward is decoupled:

$$\boldsymbol{r}(\boldsymbol{S}_t, \boldsymbol{A}_t) - \lambda\|\boldsymbol{A}_t\|_1 = \sum_{i\in[n]}\Big[r^i(S_t^i, A_t^i) - \lambda A_t^i\Big].$$

Therefore, for a given λ, maximizing the Lagrangian relaxation (3) over $\boldsymbol{\pi} : \mathcal{S} \to \{0, 1\}^n$ is equivalent to the following n decoupled optimization problems: for all $i \in [n]$,

$$\max_{\pi^i:\mathcal{S}^i\to\{0,1\}} \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\Big[r^i(S_t^i, A_t^i) - \lambda A_t^i\Big]\right]. \tag{4}$$

Let π_λ^i denote the optimal policy for Problem (4). Define the *passive set* $\mathcal{W}_\lambda^i$ as the set of states for which the optimal policy π_λ^i prescribes passive action, i.e., $\mathcal{W}_\lambda^i := \big\{s \in \mathcal{S}^i : \pi_\lambda^i(s) = 0\big\}$.

Definition 1 (Indexability and Whittle index): A RB is said to be **indexable** if $\mathcal{W}_\lambda^i$ 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 $\mathcal{W}_{\lambda_1}^i \subseteq \mathcal{W}_{\lambda_2}^i$. For an indexable RB, **the Whittle index** $w^i(s)$ of state $s \in \mathcal{S}^i$ is the smallest value of λ for which state s is part of the passive set $\mathcal{W}_\lambda^i$, i.e.,

$$w^i(s) = \inf\left\{\lambda \in \mathbb{R} : s \in \mathcal{W}_\lambda^i\right\}.$$

Note that if the penalty $\lambda = w^i(s)$, then the policy π_λ^i is indifferent between taking passive or active actions at state s.

The Whittle index policy is a feasible policy for the original optimization problem and is given as follows: *At each time, activate the arms with the m largest values of the Whittle index at their current state.*

As argued in [16], the Whittle index policy is meaningful only when all arms are indexable. Various sufficient conditions for indexability are available in the literature [21], [22], [24]. In some settings, the Whittle index policy is optimal [6], [17], [18]. For general models, there is also strong evidence to suggest that the Whittle index policy performs close to optimal [21], [22], [24], [43], [44]. Algorithms to efficiently compute Whittle indices are presented in [24], [45].

C. The learning problem

Let $\boldsymbol{\mu}$ denote the Whittle index policy and $\boldsymbol{J}(\boldsymbol{\mu})$ denote its performance. We are interested in a setting where the transition matrices $\{P^i\}_{i\in[n]}$ of the arms are unknown but the decision maker has a prior on them. In this setting, the performance of a policy $\boldsymbol{\pi}$ operating for horizon T is characterized by the Bayesian regret given by

$$\mathcal{R}(T; \boldsymbol{\pi}) = \mathbb{E}^{\boldsymbol{\pi}}\left[T\boldsymbol{J}(\boldsymbol{\mu}) - \sum_{t=1}^{T} \boldsymbol{r}(\boldsymbol{S}_t, \boldsymbol{A}_t)\right], \tag{5}$$

where the expectation is taken with respect to the prior distribution on $\{P^i\}_{i\in[n]}$, the initial condition, and the potential randomization done by the policy $\boldsymbol{\pi}$. Bayesian regret is a well-known metric used in various setting [42], [46]–[48]. An alternative method to quantify the performance is the frequentist regret but we focus on Bayesian regret for a comparison of the two notions of the regret, we refer the reader to [49], [50].

Remark 2: We measure the regret with respect to the Whittle index policy. In contrast, in most of the existing research, regret is defined with reference to the optimal policy. In principle, the results presented in this paper are also applicable to regret defined with respect to the optimal policy provided it is possible to compute the optimal policy for a given model. See Sec. VI-C for details.

Remark 3: The rested multi-armed bandit problem is a special case of Model B where passive arms are frozen, i.e., $P^i(s_+|s,0) = \mathbb{1}(\{s_+ = s\})$ for all arms $i \in [n]$. For this model, the Whittle index policy reduces to what is called the *Gittins index policy* and is optimal [17]. Thus, the results obtained in this paper are also applicable to the rested multi-armed bandits.

D. The main results

Our main result is to propose a Thompson-sampling based algorithm, which we call RB-TSDE, and characterize its regret. In particular, let $\mathsf{S}^i = |\mathcal{S}^i|$ denote the size of the state space of arm i and $\bar{\mathsf{S}}_n = \sum_{i\in[n]} \mathsf{S}^i$ denote the sum of the sizes of the state space of all arms. Then, we show the following.

Main Result : The regret of RB-TSDE is bounded by

$$\mathcal{R}(T; \texttt{RB} - \texttt{TSDE}) \le \mathcal{O}\big(\alpha \bar{\mathsf{S}}_n \sqrt{T \log T}\big),$$

where $\alpha = n$ for Model A and $\alpha = m$ for Model B. Under additional assumptions, the bound for both models can be tightened to

$$\mathcal{R}(T; \texttt{RB} - \texttt{TSDE}) \le \mathcal{O}\big(\max\{\alpha \bar{\mathsf{S}}_n, \bar{\mathsf{S}}_n\} \sqrt{T \log T}\big).$$

The detailed characterization of the constants in the $\mathcal{O}(\cdot)$ terms is given in Theorem 1 and Theorem 2 later.

III. Learning Algorithm for RB

A. Assumptions on the unknown parameters.

Let $\theta^i_\star$ denote the unknown parameters of the transition matrices $[P^i(\cdot|\cdot,0)\ P^i(\cdot|\cdot,1)]$, $i \in [n]$. We assume that $\theta^i_\star$ belongs to a compact set Θ^i. We impose the following assumptions on the model.

Assumption 1: For any $i \in [n]$ and $\theta^i \in \Theta^i$, the RB $\langle \mathcal{S}^i, \mathcal{A}^i = \{0,1\}, P^i(\theta^i), r^i \rangle$ is indexable.

Assumption 2: Let $\boldsymbol{\mu}_\theta$ denote the Whittle index policy corresponding to model θ. Let $\boldsymbol{P}_\theta$ denote the controlled transition matrix under policy $\boldsymbol{\mu}_\theta$ and J_θ denote the average reward of policy $\boldsymbol{\mu}_\theta$. We assume that for every $\theta \in \Theta$, J_θ does not depend on the initial state and also assume that there exists a bounded differential value function $\boldsymbol{V}_\theta$ such that $(J_\theta, \boldsymbol{V}_\theta)$ satisfy the average reward Bellman equation:

$$J_\theta + \boldsymbol{V}_\theta(\boldsymbol{s}) = \boldsymbol{r}(\boldsymbol{s}, \boldsymbol{\mu}_\theta(\boldsymbol{s})) + \big[\boldsymbol{P}_\theta \boldsymbol{V}_\theta\big](\boldsymbol{s}), \quad \forall \boldsymbol{s} \in \boldsymbol{\mathcal{S}}. \quad (6)$$

Assumption 1 is necessary for the Whittle index heuristic to be meaningful. Assumption 2 ensures that the average reward of the Whittle index policy is well defined for all models. There are various sufficient conditions for Assumption 2 in the literature. See [51] for an overview.

The average reward Bellman equation (6) has an infinite number of solutions. In particular, if $(J_\theta, \boldsymbol{V}_\theta)$ satisfies (6), then so does $(J_\theta, \boldsymbol{V}_\theta + \text{constant})$. Assumption 2 implies that $\operatorname{span}(\boldsymbol{V}_\theta)$ is bounded. A bound on $\operatorname{span}(\boldsymbol{V}_\theta)$ under a different set of assumptions is presented in [38] but this bound does not suffice for our analysis. See Remark 7 for details. As we want to capture the scaling of $\operatorname{span}(\boldsymbol{V}_\theta)$ with n and m, so we impose an additional assumption on the model.

The ergodicity coefficient of $\boldsymbol{P}_\theta$ is defined as

$$\lambda_{\boldsymbol{P}_\theta} = 1 - \min_{\boldsymbol{s},\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{S}}} \min\{\boldsymbol{P}_\theta(\boldsymbol{z}|\boldsymbol{s}), \boldsymbol{P}_\theta(\boldsymbol{z}|\boldsymbol{s}')\}.$$

We impose the following assumption on the model.

Assumption 3: We assume there exists $\lambda^* < 1$ such that $\sup_{\theta\in\Theta} \lambda_{\boldsymbol{P}_\theta} \le \lambda^*$.

See [51, Sec. 5] for various equivalent characterizations of $\lambda_{\boldsymbol{P}_\theta} < 1$. Assumption 3 is used while analyzing the rate of convergence of relative value iteration [52] to bound the span of the value function. A relaxation of Assumption 3 is presented in Sec. VI-A.

The ergodicity coefficient of $\boldsymbol{P}_\theta$ depends on the Whittle index policy $\boldsymbol{\mu}_\theta$. A policy independent upper bound of the ergodicity coefficient is given by the contraction factor, which is defined as:

$$\lambda' = 1 - \min_{\substack{\boldsymbol{s},\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}, \\ \boldsymbol{a} \in \boldsymbol{\mathcal{A}}(m), \boldsymbol{a}' \in \boldsymbol{\mathcal{A}}(m)}} \sum_{\boldsymbol{z} \in \boldsymbol{\mathcal{S}}} \min\{\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{s},\boldsymbol{a}), \boldsymbol{P}(\boldsymbol{z}|\boldsymbol{s}',\boldsymbol{a}')\}$$

Since the dynamics of the arms are independent, the definition of contraction factor implies that a sufficient condition for Assumption 3 is that for every arm, and every pair of state-action pairs, there exists a next state that can be reached from both state-action pairs with positive probability in one step. Moreover, if for every arm there is a distinguished state which can be reached from all state-action pairs with probability at least ε, then the ergodicity coefficient is less than $1 - \varepsilon$.

B. Priors and posterior updates.

We assume that $\{\theta^i_\star\}_{i\in[n]}$ are independent random variables and use ϕ^i_1 to denote the prior on θ^i for each arm $i \in [n]$. At time t, let $h^i_t = (s^i_1, a^i_1, \dots, s^i_{t-1}, a^i_{t-1}, s^i_t)$ denote the history of states and actions at arm i. Let ϕ^i_t denote the posterior distribution on $\theta^i_\star$ given h^i_t. Then, upon applying action a^i_t and observing the next state s^i_{t+1}, the posterior distribution ϕ^i_{t+1} can be computed using Bayes rule as

$$\phi^i_{t+1}(d\theta) = \frac{P^i(s^i_{t+1}|s^i_t, a^i_t; \theta)\phi^i_t(d\theta)}{\int P^i(s^i_{t+1}|s^i_t, a^i_t; \tilde{\theta})\phi^i_t(d\tilde{\theta})}. \quad (7)$$

If the prior is a conjugate distribution on Θ^i, then the posterior can be updated in closed form. Note that the exact form of the prior and the posterior update rule are not important for the description of the algorithm or its regret analysis.

C. RB-TSDE Algorithm (Distributed implementation)

We propose a Thompson-sampling based algorithm, which we call RB-TSDE. Our algorithm is inspired by the TSDE (Thompson Sampling with Dynamic Episodes) algorithm of [42].

The RB-TSDE algorithm consists of a coordinator and n actors, one for each arm. The coordinators and the actors

Algorithm 1 RB-TSDE

Input: $\{(\Theta^i, \phi_1^i)\}_{i\in[n]}$.
Initialize $t \leftarrow 1$, $t_1 \leftarrow 0$, $T_0 \leftarrow 0$, $k \leftarrow 0$, $N^i(s^i, a^i) = 0, \forall a^i \in \{0,1\}, s^i \in \mathsf{S}^i, \forall i \in [n]$, θ_0, μ_{θ_0}.
for $t = 1, 2, \dots$ **do**
  if $t_k + T_{k-1} < t$ or $2N^i_{t_k}(s,a) < N^i_t(s,a)$ for any $s \in \mathcal{S}^i$, $a \in \{0,1\}$, $i \in [n]$ **then**
    Set $k \leftarrow k+1$, $T_{k-1} \leftarrow t - t_{k-1}$, $t_k \leftarrow t$.
    Actor $i \in [n]$ samples $\theta^i_k \sim \phi^i_{t_k}$ and compute $w^i_{t_k}$.
  end if
  Actor $i \in [n]$ sends the Whittle index $w^i_{t_k}(s^i_t)$ to the coordinator.
  The coordinator sends $a^i_t = 1$ to the arms with the m-highest Whittle index and sends $a^i_t = 0$ to others.
  Actor $i \in [n]$ updates ϕ^i_{t+1} according to (7).
end for

require synchronized communication as described below. The whole algorithm is described in Alg. 1.

As the name suggests, RB-TSDE operates in episodes of dynamic length. The episodes are synchronized for all actors and the coordinator signals the start of episodes to all actors. The actor at arm i maintains a posterior ϕ^i_t distribution on the dynamics of arm i according to (7) and keep track of $N^i_t(s^i, a^i) = \sum_{\tau=1}^{t-1} \mathbb{1}(\{(S^i_\tau, A^i_\tau) = (s^i, a^i)\})$.

Let t_k and T_k denote the start time and length of episode k, respectively. The end of the episode can either be triggered by the coordinator or any of the actors. The coordinator triggers the end of the episode if the length of the episode is one more than the length of the previous episode. The actor for arm i triggers the end of the episode if the number of state-action visits $N^i_t(s^i_t, a^i_t)$ of the current state-action pair are more than double of their value at the beginning of the episode. Thus,

$$t_{k+1} = \min\big\{t > t_k : t - t_k > T_{k-1} \text{ or } N^i_t(s^i, a^i) > 2N_{t_k}(s^i, a^i) \text{ for some } (i, s^i, a^i)\big\}.$$

At the beginning of episode k, the actor for arm $i \in [n]$ samples a parameter θ^i_k from the posterior $\phi^i_{t_k}$ and computes the Whittle index $w^i_{t_k}$ for all states. During episode k, at each time t, the actor at arm i sends the value of $w^i_{t_k}(s^i_t)$ to the coordinator. The coordinator receives $w^i_{t_k}(s^i_t)$ from all arms, sends the active action $a^i_t = 1$ to the arms with the m-highest values of the Whittle index, and sends the passive action $a^i_t = 0$ to the remaining arms. This process continues until a condition for ending the episode is triggered by the coordinator or one of the actors.

D. Regret bound

Theorem 1: Under Assumptions 1–3, the regret of RB-TSDE is upper bounded as follows:

$$\mathcal{R}(T; \texttt{RB-TSDE}) < 30\alpha \frac{R_{\max}}{1-\lambda^*} \bar{\mathsf{S}}_n \sqrt{T \log T},$$

where $\alpha = n$ for Model A and $\alpha = m$ for Model B.
See Sec. IV-D for proof.

We derive a tighter regret bound under a stronger assumption. We first assume that the state space of each arm $\mathcal{S}^i$, $i \in [n]$, is a finite subset of $\mathbb{R}$ and use d^i to denote the Euclidean metric on $\mathbb{R}$, i.e., $d^i(s, s') = |s - s'|$. Furthermore, let $\boldsymbol{d}_p(\boldsymbol{s}, \boldsymbol{s}') = \big(\sum_{i\in[n]} d^i(s^i, s'^i)^p\big)^{1/p}$ for any $\boldsymbol{s}, \boldsymbol{s}' \in \boldsymbol{\mathcal{S}}$. We then impose the following assumption.

Assumption 4: For each $\theta \in \Theta$, the value function $\boldsymbol{V}_\theta$ is Lipschitz with a Lipschitz constant upper bounded by $\boldsymbol{L}_v$.

In general, Assumption 4 depends on the specific model being considered. We present one instance where Assumption 4 is satisfied in Sec. VI-B.

Theorem 2: Under Assumptions 1–4, the regret of RB-TSDE for both Model A and Model B is upper bounded as follows:

$$\mathcal{R}(T; \texttt{RB-TSDE}) < 12 \max\{\alpha\sqrt{\bar{\mathsf{S}}_n}, \bar{\mathsf{S}}_n\} \max\Big\{\frac{R_{\max}}{1-\lambda^*}, \mathsf{D}_{\max} \boldsymbol{L}_v\Big\} \sqrt{KT \log(T \max\{1, K'\})}$$

where $\alpha = n$ for Model A and $\alpha = m$ for Model B, K and K' are positive constants independent of n and T, and $\mathsf{D}_{\max} = \max_{i\in[n]} \operatorname{diam}(\mathcal{S}^i)$.
See Sec. IV-D for proof.

Remark 4: If we directly use an existing RL algorithm for RBs, the regret will scale as $\tilde{\mathcal{O}}(2^n\sqrt{T})$ or larger. The results of Theorems 1 and 2 show that the regret scales as either $\tilde{\mathcal{O}}(n^2\sqrt{T})$, $\tilde{\mathcal{O}}(n^{1.5}\sqrt{T})$, or $\tilde{\mathcal{O}}(n\sqrt{T})$ depending on the modeling assumptions. Thus, using a learning algorithm which is adapted to the structure of the models gives a significantly better scaling with the number of arms.

Remark 5: The exact scaling with the number of arms depends on how m scales with n. For example, if m remains constant, then under Assumptions 1–3, the regret for Model A scales as $\tilde{\mathcal{O}}(n^2\sqrt{T})$, while the regret for Model B scales as $\tilde{\mathcal{O}}(n\sqrt{T})$. Under Assumption 4, the regret for Model A scales as $\tilde{\mathcal{O}}(n^{1.5}\sqrt{T})$, while the regret for Model B scales as $\tilde{\mathcal{O}}(n\sqrt{T})$. On the other hand, if m scales as βn, where $\beta < 1$ is a constant[1], then under Assumptions 1–3, the regret for both models scales as $\tilde{\mathcal{O}}(n^2\sqrt{T})$. Under Assumption 4, the regret for both models scales as $\tilde{\mathcal{O}}(n^{1.5}\sqrt{T})$. Thus, for Model A, the regret bound of Theorem 2 is tighter than that of Theorem 1, but for Model B, it depends on the scaling assumptions on m.

IV. Regret Analysis

The high level idea of the proof is similar to the analysis in [42] but we exploit the properties of the RB model while simplifying individual terms. We first start with bounds on the average reward and the differential value function.

A. Bounds on average reward and differential value function.

As mentioned earlier, $\boldsymbol{V}_\theta$ is unique only up to an additive constant. We assume that $\boldsymbol{V}_\theta$ is chosen such that $\boldsymbol{\xi}_\theta \boldsymbol{V}_\theta = 0$, where $\boldsymbol{\xi}_\theta$ is the stationary distribution of $\boldsymbol{P}_\theta$. This ensures that

[1]For this setting, it was shown in [18] that the Whittle index policy is optimal as $n \to \infty$.

$\boldsymbol{V}_\theta$ is equal to the *asymptotic bias* of policy $\boldsymbol{\mu}_\theta$ and is given by

$$\boldsymbol{V}_\theta = \sum_{t=0}^{\infty} \boldsymbol{P}_\theta^t \big[\boldsymbol{r} - J_\theta\big]. \tag{8}$$

See for example [53].

Then we have the following bounds.

Lemma 1: Under Assumption 3, for any $\theta \in \Theta$,

$$0 \le J_\theta \le \alpha R_{\max} \text{ and } \operatorname{span}(\boldsymbol{V}_\theta) \le 2\alpha R_{\max}/(1-\lambda^*),$$

where $\alpha = n$ for Model A and $\alpha = m$ for Model B.

Proof: Note that for Model A, $\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a}) \in [0, nR_{\max}]$ while for Model B, $\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a}) \in [0, mR_{\max}]$. Then, the bounds for J_θ follow immediately from definition. The bounds on $\operatorname{span}(\boldsymbol{V}_\theta)$ follow from (8), $\operatorname{span}(\boldsymbol{s}+\boldsymbol{y}) \le \operatorname{span}(\boldsymbol{s}) + \operatorname{span}(\boldsymbol{y})$, Assumption 3, and the fact that for any vector $\boldsymbol{v}$, $\operatorname{span}(\boldsymbol{P}_\theta \boldsymbol{v}) \le \lambda_{\boldsymbol{P}_\theta} \operatorname{span}(\boldsymbol{v})$. ∎

Remark 6: Lemma 1 shows the key difference between Models A and B. When all arms yield rewards, the maximum value of $\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a})$ is $nR_{\max}$ while when only active arms yield rewards, the maximum value of $\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a})$ is $mR_{\max}$. This leads to different bounds on J_θ and $\boldsymbol{V}_\theta$.

Remark 7: An alternative bound on $\operatorname{span}(\boldsymbol{V}_\theta)$ is presented in [38]. Let $\boldsymbol{T}_\theta^{\boldsymbol{s_1}\to\boldsymbol{s_2}}$ denote the expected number of steps to go from state $\boldsymbol{s_1}$ to state $\boldsymbol{s_2}$ under policy $\boldsymbol{\mu}_\theta$ for model θ. Define $\boldsymbol{D}_\theta = \max_{\boldsymbol{s_1},\boldsymbol{s_2}} \boldsymbol{T}_\theta^{\boldsymbol{s_1}\to\boldsymbol{s_2}}$ to be the *one-way diameter*. Then, it is shown in [38] that $\operatorname{span}(\boldsymbol{V}_\theta) \le J_\theta \boldsymbol{D}_\theta$. We do not know of an easy way to characterize the dependence of $\boldsymbol{D}_\theta$ on the number n of arms. That is why we consider an alternative bound on $\operatorname{span}(\boldsymbol{V}_\theta)$.

B. Regret decomposition.

For the ease of notation, we simply use $\mathcal{R}(T)$ instead of $\mathcal{R}(T; \texttt{RB}-\texttt{TSDE})$. We also use $(J_\star, \mu_\star, \boldsymbol{P}_\star, \boldsymbol{V}_\star)$ instead of $(J_{\theta_\star}, \mu_{\theta_\star}, \boldsymbol{P}_{\theta_\star}, \boldsymbol{V}_{\theta_\star})$ and use $(J_k, \mu_k, \boldsymbol{P}_k, \boldsymbol{V}_k)$ instead of $(J_{\theta_k}, \mu_{\theta_k}, \boldsymbol{P}_{\theta_k}, \boldsymbol{V}_{\theta_k})$. Rearranging terms in Bellman equation (6) and adding and subtracting $\boldsymbol{V}_k(\boldsymbol{s}_{t+1})$, we get:

$$\begin{aligned}\boldsymbol{r}(\boldsymbol{s}_t,\boldsymbol{a}_t) = J_k + \boldsymbol{V}_k(\boldsymbol{s}_t) - \boldsymbol{V}_k(\boldsymbol{s}_{t+1}) + \boldsymbol{V}_k(\boldsymbol{s}_{t+1}) \\ - \big[\boldsymbol{P}_k \boldsymbol{V}_k\big](\boldsymbol{s_t}).\end{aligned} \tag{9}$$

Let K_T denote the number of episodes until horizon T. Substituting (9) in (5), we get:

$$\begin{aligned}\mathcal{R}(T) = \quad & \underbrace{\mathbb{E}\Big[T J_\star - \sum_{k=1}^{K_T} T_k J_k\Big]}_{\text{regret due to sampling error} =: \mathcal{R}_0(T)} \\ & + \underbrace{\mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \boldsymbol{V}_k(\boldsymbol{S}_{t+1}) - \boldsymbol{V}_k(\boldsymbol{S}_t)\Big]}_{\text{regret due to time-varying policy} =: \mathcal{R}_1(T)} \\ & + \underbrace{\mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \big[\boldsymbol{P}_k \boldsymbol{V}_k\big](\boldsymbol{S}_t) - \boldsymbol{V}_k(\boldsymbol{S}_{t+1})\Big]}_{\text{regret due to model mismatch} =: \mathcal{R}_2(T)}.\end{aligned} \tag{10}$$

C. Bounding individual terms.

Each term of (10) is bounded as follows.

Lemma 2: Under Assumptions 1–3, we have

1) $\mathcal{R}_0(T) \le 2\alpha R_{\max} \sqrt{\bar{\mathsf{S}}_n T \log T}$.
2) $\mathcal{R}_1(T) \le 4 \dfrac{\alpha R_{\max}}{1-\lambda^*} \sqrt{\bar{\mathsf{S}}_n T \log T}$.
3) $\mathcal{R}_2(T) \le 12 \dfrac{\alpha R_{\max}}{1-\lambda^*} \big(n + \bar{\mathsf{S}}_n \sqrt{T \log T}\big)$.

See the appendix for the proof steps.

We can obtain an alternative bound on $\mathcal{R}_2(T)$ under Assumption 4.

Lemma 3: Under Assumptions 1, 2, and 4, we have

$$\mathcal{R}_2(T) \le 16 \mathsf{D}_{\max} \boldsymbol{L}_v \sqrt{K \bar{\mathsf{S}}_n T \log(K'T)}$$

where K and K' are positive constants that do not depend on n and T.

See the appendix for the proof steps.

Remark 8: Under Assumption 6, which will be described later, we can establish a tighter bound on $\operatorname{span}(\boldsymbol{V}_\theta)$. In particular, Lipschitz continuity of $\boldsymbol{V}_\theta$ implies that $\operatorname{span}(\boldsymbol{V}_\theta) \le \boldsymbol{L}_v \operatorname{diam}(\boldsymbol{\mathcal{S}})$. This can give us a tighter bound on the term $\mathcal{R}_1(T)$, but this tighter bound does not help us in reducing the overall regret.

D. Obtaining the final bound.

Proof: [Proof of Theorem 1] From Eq. (10) and Lemma 2, we get

$$\begin{aligned}\mathcal{R}(T) &\le 2\alpha R_{\max} \sqrt{\bar{\mathsf{S}}_n T \log T} + 4 \frac{\alpha R_{\max}}{1-\lambda^*} \sqrt{\bar{\mathsf{S}}_n T \log T} \\ &\quad + 12 \frac{\alpha R_{\max}}{1-\lambda^*} \big(n + \bar{\mathsf{S}}_n \sqrt{T \log T}\big).\end{aligned}$$

By definition, $\lambda^* < 1$. Then,

$$\begin{aligned}\mathcal{R}(T) &< 6 \frac{\alpha R_{\max}}{1-\lambda^*} \sqrt{\bar{\mathsf{S}}_n T \log T} + 24 \frac{\alpha R_{\max}}{1-\lambda^*} \bar{\mathsf{S}}_n \sqrt{T \log T} \\ &< (6+24) \frac{\alpha R_{\max}}{1-\lambda^*} \bar{\mathsf{S}}_n \sqrt{T \log T} \\ &= 30 \frac{\alpha R_{\max}}{1-\lambda^*} \bar{\mathsf{S}}_n \sqrt{T \log T}.\end{aligned}$$

This completes the proof of Theorem 1. ∎

Proof: [Proof of Theorem 2] From Eq. (10), Lemma 2, and 3, we get

$$\begin{aligned}\mathcal{R}(T) &\le 2\alpha R_{\max} \sqrt{\bar{\mathsf{S}}_n T \log T} + 4 \frac{\alpha R_{\max}}{1-\lambda^*} \sqrt{\bar{\mathsf{S}}_n T \log T} \\ &\quad + 12 \bar{\mathsf{S}}_n \mathsf{D}_{\max} \boldsymbol{L}_v \sqrt{KT \log(T \max\{1, K'\})} \\ &< 12 \frac{\alpha R_{\max}}{1-\lambda^*} \sqrt{\bar{\mathsf{S}}_n T \log T} \\ &\quad + 12 \bar{\mathsf{S}}_n \mathsf{D}_{\max} \boldsymbol{L}_v \sqrt{KT \log(T \max\{1, K'\})}.\end{aligned}$$

Let $\bar{R} = \max\Big\{R_{\max}/(1-\lambda^*), \mathsf{D}_{\max} \boldsymbol{L}_v\Big\}$. Then, we have

$$\mathcal{R}(T) < 12 \max\{\alpha \sqrt{\bar{\mathsf{S}}_n}, \bar{\mathsf{S}}_n\} \bar{R} \sqrt{KT \log(T \max\{1, K'\})}.$$

This completes the proof of Theorem 2. ∎

V. Numerical examples

In this section, we demonstrate the empirical performance of RB-TSDE. In particular, we consider two environments, one for Model A (a model for machine maintenance) and one for Model B (a model for link scheduling). For both environments, we consider multiple experiments and plot the regret as a function of time and as a function of number of arms. Our results illustrate that the regret does indeed scale according to our theoretical results. We also compare the results with the empirical performance of QWI, which is a Q-learning algorithm for RBs proposed in [26], [28], [29]. Note that only the algorithm proposed in [29] is called QWI, but the algorithms of [26], [28] are conceptually similar, so we collectively call them QWI.

A. Environments

We start with a description of the two environments.

1) Environment A: We consider a machine maintenance model where a single repairman is responsible for the maintenance of a set of machines, which deteriorate over time. Each machine has multiple deterioration states sorted from *pristine* to *ruined*. There is a cost associated with running the machine and the cost is non-decreasing function of the state. If the machine is left un-monitored, then the state of the machine stochastically deteriorates over time. The repairman may visit one of the machines and replace it with a new machine at a fixed cost. The objective is to determine a scheduling policy to minimize the expected discounted cost over time.

We model the above environment as an instance of Model A. In particular, we consider n arms, where $n \in \{10, 20, \dots, 80\}$ where $m = 1$ arm can be activated at each time. The state space of each arm is of size $\mathsf{S} = 10$. Under $a = 1$, the state of the arm is reset to 1 (and this fact is known to the learner). Under $a = 0$, the transition matrix is stochastic monotone and chosen as described in [24, Appendix 1.2]. The transitions under the passive action are unknown to the learner. The per-step reward function are given by $r^i(s, 0) = (\mathsf{S}-1)^2-(s-1)^2$, $r^i(s, 1) = 0.5(\mathsf{S} - 1)^2$ for all $i \in [n]$ and $s \in [\mathsf{S}]$.

2) Environment B: We consider a link scheduling problem where there are n users who can communicate over a shared communication link. Each user has a queue, where packets arrive according to an unknown i.i.d. process. At each time, a controller may schedule one of the users and transmit all its packets over the channel. The users which are not scheduled incur a holding cost which is equal to the square of the number of packets in the queue.

We model the above environment as an instance of Model B. In particular, we consider n arms, where $n \in \{10, 20, \dots, 80\}$ where $m = n - 1$ arm can be activated at each time. The state of each arm is of size $\mathsf{S} = 10$. Under $a = 1$, the transition matrix is upper-triangular and chosen as described in [54, $\mathcal{P}_1(p)$ of Appendix A] where p is set to be different for each arm, linearly ranged from 0.05 to 0.95. The transitions under the active action are unknown to the learner. Under $a = 0$, the state of the arm is reset to 1 (and this fact is known to the learner). The per-step reward function are given by $r^i(s, 0) = 0$, $r^i(s, 1) = (\mathsf{S} - 1)^2 - (s - 1)^2$ for all $i \in [n]$ and $s \in [\mathsf{S}]$.

B. Algorithms

We compare the performance of two algorithms.

1) RB-TSDE: We consider the RB-TSDE algorithm described in Algorithm 1. We initialize the algorithm with uninformed Dirichlet prior on the unknown parameters and update the posterior according to the conjugate posterior for Dirichlet priors.

2) QWI : We also consider QWI, which is a Q-learning algorithm for RBs proposed in [26], [28], [29] as a baseline. The algorithm has two learning rates. As recommended in [28, Eq. (17)] we pick the step-size sequence which has a good performance by setting parameters C and C' of QWI as $C = 0.03$ and $C' = 0.01$, where the numerical values were obtained by running a hyper-parameter search.

C. Experimental Results

In our experiments, we pick a horizon of $T = 5,000$ and compute the Bayesian regret averaged over 250 sample paths. We repeat the experiment for $n \in \{10, 20, \dots, 80\}$. For each environment, we plot four curves: (a) plot of $\mathcal{R}(T)/\sqrt{T}$ vs T for RB-TSDE; (b) plot of $\mathcal{R}(T)$ at $T = 5,000$ vs n for RB-TSDE; (c) plot of $\mathcal{R}(T)/\sqrt{T}$ vs T for QWI; and (d) plot of $\mathcal{R}(T)$ at $T = 5,000$ vs n for QWI. For plots (b) and (d), we also fit the points with a parametric curve of the form $p_0 + p_1 n + p_2 n^{1.5}$ for Environment A and $p_0 + p_1 n$ for Environment B to obtain the scaling with number of arms.

The plots for Environment A is shown in Fig. 1. The sub-plot (a) shows that the regret essentially scales as $\sqrt{T}$ with time. The sub-plots (b) show that the regret scales as $n^{1.5}$ with the number of arms. Thus, the results are consistent with regret bounds of Theorem 2.

The plots for Environment B is shown in Fig. 2. The behavior of sub-plots (a) and (c) is the same as for Environment A. Note that for larger values of T, the plot of $\mathcal{R}(T)/\sqrt{T}$ has not yet converged to a straight line, but it is clear that these curves are upper bounded by a constant. The sub-plot (b) shows that the regret scales linearly with the number of arms. Note that in this case since $m = n - 1$, Theorem 1 suggests that the regret should scale as $n^{1.5}$ (see Remark 5) and the result is consistent with the theorem.

Note that even though no regret analysis of the QWI algorithm for RBs is presented in [26], [28], [29], the above experiment suggests that empirically the performance of QWI has similar features as RB-TSDE. However, unlike the QWI, RB-TSDE does not require any hyper-parameter tuning. Moreover, RB-TSDE has an order of magnitude lower regret than QWI.

VI. Discussion

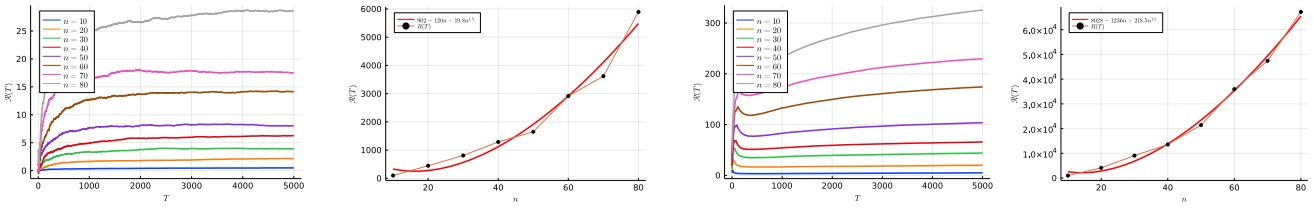
A. A relaxation of Assumption 3

Assumption 3 can be relaxed as follows.

Assumption 5: For every θ, there exists a positive integer τ^* and a real $\lambda^* \in (0, 1)$ such that $\lambda_{P_\theta^{\tau^*}} \leq \lambda^*$.

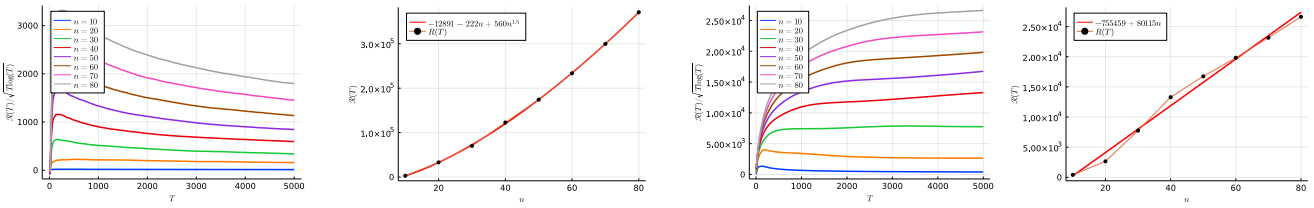
Based on this assumption, the result of Lemmas 1 changes as follows.

Lemma 4: Under Assumption 5, for any $\theta \in \Theta$, $\mathrm{span}(\boldsymbol{V}_\theta) \leq 2\tau^* \alpha R_{\max}/(1 - \lambda^*)$.



(a) $\mathcal{R}(T)/\sqrt{T}$ vs. T (b) $\mathcal{R}(T)$ vs. n (c) $\mathcal{R}(T)/\sqrt{T}$ vs. T (d) $\mathcal{R}(T)$ vs. n

Fig. 1: Regret analysis of RB-TSDE and QWI for Environment A. Note that RB-TSDE has an order of magnitude better regret than QWI.



(a) $\mathcal{R}(T)/\sqrt{T}$ vs. T (b) $\mathcal{R}(T)$ vs. n (c) $\mathcal{R}(T)/\sqrt{T}$ vs. T (d) $\mathcal{R}(T)$ vs. n

Fig. 2: Regret analysis of RB-TSDE and QWI method for Environment B. Note that RB-TSDE has an order of magnitude lower regret than QWI.

Consequently, we have the following changes.

Lemma 5: Under Assumptions 1, 2, and 5, we have

1) $\mathcal{R}_1(T) \leq 4\dfrac{\tau^* \alpha R_{\max}}{1-\lambda^*}\sqrt{\bar{\mathsf{S}}_n T \log T}$.

2) $\mathcal{R}_2(T) \leq 12\dfrac{\tau^* \alpha R_{\max}}{1-\lambda^*}\big(n + \bar{\mathsf{S}}_n\sqrt{T\log T}\big)$.

Theorem 3: Under Assumptions 1, 2, and 5, the regret of RB-TSDE is upper bounded as follows:

$$\mathcal{R}(T;\mathtt{RB-TSDE}) < 30\alpha\frac{R_{\max}}{1-\lambda^*}\bar{\mathsf{S}}_n\sqrt{T\log T}.$$

Theorem 4: Under Assumptions 1, 2, 4, and 5 the regret of RB-TSDE for both models is upper bounded as follows:

$$\mathcal{R}(T;\mathtt{RB-TSDE}) < 12\max\{\alpha\sqrt{\bar{\mathsf{S}}_n}, \bar{\mathsf{S}}_n\}$$
$$\max\Big\{\frac{\tau^* R_{\max}}{1-\lambda^*}, \mathsf{D}_{\max}\boldsymbol{L}_v\Big\}\sqrt{KT\log(T\max\{1,K'\})}.$$

The proof steps of Lemmas 4, 5 are similar to the proof steps of Lemmas 1, 2 and the proof steps of Theorems 3 and 4 are similar to the proof steps of Theorems 1 and 2. See [55] for the details.

B. A set of sufficient conditions for Assumption 4

Assumption 6: Suppose each arm $i \in [n]$ is (L_r^i, L_p^i) Lipschitz, i.e.,

$$L_r^i = \sup_{s^i_{(1)}, s^i_{(2)}, a} \frac{|r^i(s^i_{(1)}, a) - r^i(s^i_{(2)}, a)|}{d^i(s^i_{(1)}, s^i_{(2)})},$$
$$L_p^i = \sup_{s^i_{(1)}, s^i_{(2)}, a} \frac{\mathcal{K}(P^i(\cdot|s^i_{(1)}, a), P^i(\cdot|s^i_{(2)}, a))}{d^i(s^i_{(1)}, s^i_{(2)})}$$

where $L_r^i < \infty$ and $L_p^i < 1$.

Assumption 7: For all $\theta \in \Theta$, the Whittle index policy is optimal.

Assumption 7 is satisfied in some instances such as: (i) the rested multi-armed bandit setup described in Remark 3 where only one arm can be activated at a time (i.e., $m = 1$) and yield reward, and arms that are not activated remain frozen (i.e., $P(s^i_+|s^i, 0) = \delta_{s^i}(s^i_+)$, where δ_{s^i} is the Dirac delta measure centered at s^i) [17]; (ii) the number of arms are asymptotically large [18]; (iii) certain queuing models [6].

Moreover, we assume that the product measure on $\boldsymbol{S}$ is $\boldsymbol{d}(\boldsymbol{s}, \boldsymbol{s'}) = \sum_{i\in[n]} d^i(s^i, s'^i)$.

Lemma 6 ([56, Lemma 2]): Under Assumption 6, the MDP $\langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{A}}(\boldsymbol{m}), \boldsymbol{P}, \boldsymbol{R}\rangle$ is $(\max_{i\in[n]} L_r^i, \max_{i\in[n]} L_p^i)$-Lipschitz i.e.,

$$\max_{\substack{\boldsymbol{s},\boldsymbol{s'}\in\boldsymbol{S}\\ \boldsymbol{a}\in\boldsymbol{A}(m)}} \frac{|\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a}) - \boldsymbol{r}(\boldsymbol{s'},\boldsymbol{a})|}{\boldsymbol{d}(\boldsymbol{s},\boldsymbol{s'})} \leq \max_{i\in[n]} L_r^i,$$
$$\max_{\substack{\boldsymbol{s},\boldsymbol{s'}\in\boldsymbol{S}\\ \boldsymbol{a}\in\boldsymbol{A}(m)}} \frac{\mathcal{K}(\boldsymbol{P}(\cdot|\boldsymbol{s},\boldsymbol{a}) - \boldsymbol{P}(\cdot|\boldsymbol{s'},\boldsymbol{a}))}{\boldsymbol{d}(\boldsymbol{s},\boldsymbol{s'})} \leq \max_{i\in[n]} L_p^i.$$

An immediate consequence of Lemma 6 is the following.

Lemma 7: Under Assumptions 6 and 7, $\boldsymbol{V}_\theta$ is Lipschitz with the Lipschitz constant bounded by

$$\boldsymbol{L}_v \leq (\max_{i\in[n]} L_r^i)/(1 - \max_{i\in[n]} L_p^i).$$

Thus, Assumptions 6 and 7 imply Assumption 4.

Proof: The result follows from Lemma 6 and [57, Theorem 4.2]. ■

C. Regret with respect to the optimal policy

We measure regret with respect to the Whittle index policy. For models where Assumption 7 is satisfied, which will be described later, the Whittle index policy is an optimal policy. Even when the assumption is not satisfied, it is possible to generalize the results of this paper to identify the regret with respect to the optimal policy. In particular, let $\boldsymbol{\psi}^*_\theta$ denote the

optimal policy for model $\theta \in \Theta$. Then, the Bayesian regret of a learning algorithm π with respect to the optimal policy is

$$\mathcal{R}^*(T;\boldsymbol{\pi}) = \mathbb{E}^{\boldsymbol{\pi}}\left[T\boldsymbol{J}(\boldsymbol{\psi}^*_\theta) - \sum_{t=1}^{T} \boldsymbol{r}(\boldsymbol{S}_t, \boldsymbol{A}_t)\right]. \qquad (11)$$

Then, in principle, we can replace the distributed implementation presented in Algorithm 1 with a modified centralized implementation where the learner observes the state of all arms and maintains the posterior ϕ^i_t for all $i \in [n]$. At the beginning of each episode, the learner samples $\theta^i_{t_k}$ from $\phi^i_{t_k}$, computes the policy μ_{t_k}, which is optimal for the sampled model $(\theta^1_{t_k}, \dots, \theta^n_{t_k})$, and plays μ_{t_k} for the rest of the episode. The regret of this variant will be the same as the bounds in Theorems 1 and 2. However, we do not present such an analysis here because it makes the resulting algorithm impractical as computing the optimal policy is intractable when there are more than a few arms.

VII. Conclusion

In this paper, we present a Thompson-sampling based reinforcement learning algorithm for restless bandits. We show that the Bayesian regret of our algorithm with respect to an oracle that applies the Whittle index policy of the true model is either $\tilde{\mathcal{O}}(nm\sqrt{T})$, $\tilde{\mathcal{O}}(n^2\sqrt{T})$, $\tilde{\mathcal{O}}(n^{1.5}\sqrt{T})$ or $\tilde{\mathcal{O}}(n\sqrt{T})$ depending on assumptions on the model. These are in contrast to naively using any standard RL algorithm, which will have a regret that scales exponentially in n. Our results are also applicable to the rested multi-armed bandit setting, where the Whittle index policy is the same as the Gittins index and is optimal. All in all, our results illustrate that a learning algorithm which leverages the structure of the model can significantly improve regret compared to model-agnostic algorithms.

References

[1] W. Ouyang, A. Eryilmaz, and N. B. Shroff, "Downlink scheduling over Markovian fading channels," *IEEE/ACM Transactions on Networking*, vol. 24, no. 3, pp. 1801–1812, 2015.

[2] V. S. Borkar, G. S. Kasbekar, S. Pattathil, and P. Y. Shetty, "Opportunistic scheduling as restless bandits," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 4, pp. 1952–1961, 2018.

[3] K. Wang, J. Yu, L. Chen, P. Zhou, X. Ge, and M. Z. Win, "Opportunistic scheduling revisited using restless bandits: Indexability and index policy," *IEEE Transactions on Wireless Communications*, vol. 18, no. 10, pp. 4997–5010, 2019.

[4] S. Ali, A. Ferdowsi, W. Saad, and N. Rajatheva, "Sleeping multi-armed bandits for fast uplink grant allocation in machine type communications," in *IEEE Globecom*. IEEE, 2018, pp. 1–6.

[5] S. K. Singh, V. S. Borkar, and G. Kasbekar, "User association in dense mmWave networks as restless bandits," *IEEE Trans. Veh. Technol.*, 2022.

[6] C. Lott and D. Teneketzis, "On the optimality of an index rule in multichannel allocation for single-hop mobile networks with multiple service classes," *Probability in the Engineering and Informational Sciences*, vol. 14, no. 3, pp. 259–297, 2000.

[7] P. Si, F. R. Yu, H. Ji, and V. C. Leung, "Distributed multisource transmission in wireless mobile peer-to-peer networks: a restless-bandit approach," *IEEE Trans. Veh. Technol.*, vol. 59, no. 1, pp. 420–430, 2009.

[8] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[9] J. Nino-Mora, "A restless bandit marginal productivity index for opportunistic spectrum access with sensing errors," in *International Conference on Network Control and Optimization*. Springer, 2009, pp. 60–74.

[10] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: A restless bandit approach," in *2011 Proceedings IEEE INFOCOM*. IEEE, 2011, pp. 2462–2470.

[11] Q. Wang, M. Liu, and J. L. Mathieu, "Adaptive demand response: Online learning of restless and controlled bandits," in *International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 2014, pp. 752–757.

[12] C. Abad and G. Iyengar, "A near-optimal maintenance policy for automated DR devices," *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1411–1419, 2016.

[13] J. Le Ny, M. Dahleh, and E. Feron, "Multi-UAV dynamic routing with partial observations using restless bandit allocation indices," in *2008 American Control Conference*. IEEE, 2008, pp. 4220–4225.

[14] A. Dahiya, N. Akbarzadeh, A. Mahajan, and S. L. Smith, "Scalable operator allocation for multirobot assistance: A restless bandit approach," *IEEE Trans. Control Netw. Syst.*, vol. 9, no. 3, pp. 1397–1408, 2022.

[15] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control," *Math. Operat. Res.*, vol. 24, no. 2, pp. 293–305, 1999.

[16] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Prob.*, vol. 25, no. A, pp. 287–298, 1988.

[17] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society. Series B*, pp. 148–177, 1979.

[18] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *J. Appl. Prob.*, vol. 27, no. 3, pp. 637–648, 1990.

[19] K. Glazebrook and H. Mitchell, "An index policy for a stochastic scheduling model with improving/deteriorating jobs," *Naval Research Logistics*, vol. 49, no. 7, pp. 706–721, 2002.

[20] K. 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.*, vol. 165, no. 1, pp. 267–284, 2005.

[21] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride, "Some indexable families of restless bandit problems," *Adv. Appl. Prob.*, vol. 38, no. 3, pp. 643–672, 2006.

[22] J. Niño-Mora, "A $(2/3)n^3$ fast-pivoting algorithm for the gittins index and optimal stopping of a Markov Chain," *INFORMS Journal on Computing*, vol. 19, no. 4, pp. 596–606, 2007.

[23] U. Ayesta, M. Erausquin, and P. Jacko, "A modeling framework for optimizing the flow-level scheduling with time-varying channels," *Performance Evaluation*, vol. 67, no. 11, pp. 1014–1029, 2010.

[24] N. Akbarzadeh and A. Mahajan, "Conditions for indexability of restless bandits and an $\mathcal{O}(k^3)$ algorithm to compute whittle index," *Advances in Applied Probability*, vol. 54, no. 4, p. 1164–1192, 2022.

[25] R. Meshram, A. Gopalan, and D. Manjunath, "Restless bandits that hide their hand and recommendation systems," in *International Conference on Communication Systems and Networks*. IEEE, 2017, pp. 206–213.

[26] V. S. Borkar and K. Chadha, "A reinforcement learning algorithm for restless bandits," in *Indian Control Conference (ICC)*, 2018, pp. 89–94.

[27] J. Fu, Y. Nazarathy, S. Moka, and P. G. Taylor, "Towards Q-learning the Whittle index for restless bandits," in *Australian New Zealand Control Conference (ANZCC)*, 2019, pp. 249–254.

[28] K. Avrachenkov and V. S. Borkar, "Whittle index based q-learning for restless bandits with average reward," *arXiv preprint arXiv:2004.14427*, 2020.

[29] F. Robledo, V. Borkar, U. Ayesta, and K. Avrachenkov, "QWI: Q-learning with Whittle index," *ACM SIGMETRICS Performance Evaluation Review*, vol. 49, no. 2, pp. 47–50, 2022.

[30] G. Xiong, R. Singh, and J. Li, "Learning augmented index policy for optimal service placement at the network edge," *arXiv preprint arXiv:2101.03641*, 2021.

[31] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, 2012.

[32] R. Ortner, D. Ryabko, P. Auer, and R. Munos, "Regret bounds for restless Markov bandits," in *International conference on algorithmic learning theory*. Springer, 2012, pp. 214–228.

[33] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1902–1916, 2013.

[34] Y. H. Jung, M. Abeille, and A. Tewari, "Thompson sampling in non-episodic restless bandits," *arXiv preprint arXiv:1910.05654*, 2019.

[35] Y. H. Jung and A. Tewari, "Regret bounds for Thompson sampling in episodic restless bandit problems," in *Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[36] T. Gafni and K. Cohen, "Learning in restless multi-armed bandits via adaptive arm sequencing rules," *IEEE Trans. Autom. Control*, vol. 66, no. 10, pp. 5029–5036, Oct. 2020.

[37] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, no. 4, 2010.

[38] P. L. Bartlett and A. Tewari, "Regal: A regularization based algorithm for reinforcement learning in weakly communicating MDPs," in *Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, USA, 2009, p. 35–42.

[39] R. Fruit, M. Pirotta, A. Lazaric, and R. Ortner, "Efficient bias-span-constrained exploration-exploitation in reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1578–1586.

[40] S. Agrawal and R. Jia, "Posterior sampling for reinforcement learning: worst-case regret bounds," in *Neural Information Processing Systems*, 2017, pp. 1184–1194.

[41] Z. Zhang and X. Ji, "Regret minimization for reinforcement learning by evaluating the optimal bias function," in *Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[42] Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain, "Learning unknown Markov decision processes: A Thompson sampling approach," in *Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[43] K. Avrachenkov, U. Ayesta, J. Doncel, and P. Jacko, "Congestion control of TCP flows in internet routers by means of index policy," *Computer Networks*, vol. 57, no. 17, pp. 3463–3478, 2013.

[44] J. Wang, X. Ren, Y. Mo, and L. Shi, "Whittle index policy for dynamic multichannel allocation in remote state estimation," *IEEE Trans. Autom. Control*, vol. 65, no. 2, pp. 591–603, 2020.

[45] J. Niño-Mora, "Dynamic priority allocation via restless bandit marginal productivity indices," *TOP*, vol. 15, no. 2, pp. 161–198, 2007.

[46] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, 2010.

[47] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *International Conference on Machine Learning*. PMLR, 2013, pp. 127–135.

[48] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[49] E. Kaufmann, O. Cappé, and A. Garivier, "On the efficiency of bayesian bandit algorithms from a frequentist point of view," in *Neural Information Processing Systems*, 2011.

[50] E. Kaufmann, "Analysis of bayesian and frequentist strategies for sequential resource allocation," Ph.D. dissertation, Télécom ParisTech, 2014.

[51] A. Arapostathis, V. Borkar, E. Fernández-Gaucherand, M. Ghosh, and S. Marcus, "Discrete-time controlled Markov processes with average cost criterion: A survey," *SIAM J. Control Optim*, vol. 31, pp. 282–334, 1993.

[52] T. E. Morton and W. E. Wecker, "Discounting, ergodicity and convergence for Markov decision processes," *Management Science*, vol. 23, no. 8, pp. 890–900, 1977.

[53] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[54] N. Akbarzadeh and A. Mahajan, "Two families of indexable partially observable restless bandits and Whittle index computation," *arXiv preprint arXiv:2104.05151*, 2022.

[55] ——, "On learning Whittle index policy for restless bandits with scalable regret," *arXiv preprint arXiv:2202.03463*, 2022.

[56] A. Sinha and A. Mahajan, "Robustness of Whittle index policy to model approximation," *preprint*, 2022.

[57] K. Hinderer, "Lipschitz continuity of value functions in Markovian decision processes," *Mathematical Methods of Operations Research*, vol. 62, no. 1, pp. 3–22, 2005.

[58] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, "Inequalities for the l1 deviation of the empirical distribution," *Hewlett-Packard Labs, Tech. Rep*, 2003.

[59] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 707–738, 2015.

APPENDIX

A. Bound on $\mathcal{R}_0(T)$ (Lemma 2.1)

We first state a basic property of Thompson sampling algorithms.

Lemma 8 (TS Lemma [48]): Suppose the true parameters θ 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

$$\mathbb{E}[f(\theta)|\mathcal{H}] = \mathbb{E}[f(\theta_k)|\mathcal{H}].$$

Proof: Now we consider $\mathcal{R}_0(T)$. Let $\tilde{J}_\star = \alpha R_{\max} - J_\star$ and $\tilde{J}_k = \alpha R_{\max} - J_k$. By Lemma 1, we have that $\tilde{J}_\star, \tilde{J}_k \in [0, R_{\max}]$. Therefore,

$$\begin{aligned}\mathcal{R}_0(T) &= \mathbb{E}\left[TJ_\star - \sum_{k=1}^{K_T} T_k J_k\right] = \mathbb{E}\left[\sum_{k=1}^{K_T} T_k \tilde{J}_k - T\tilde{J}_\star\right] \\ &\overset{(a)}{\leq} \sum_{k=1}^{\infty} \mathbb{E}\left[\mathbb{1}(\{t_k \leq T\})(T_{k-1}+1)\tilde{J}_\star\right] - T\left[\tilde{J}_\star\right] \\ &\leq \mathbb{E}\left[\sum_{k=1}^{K_T}(T_{k-1}+1)\tilde{J}_\star\right] - T\mathbb{E}\left[\tilde{J}_\star\right] \overset{(b)}{\leq} \alpha R_{\max}\mathbb{E}[K_T],\end{aligned}$$

where (a) uses the TS Lemma and the fact that due to the first stopping criterion, $T_k \leq T_{k-1}+1$; (b) uses Lemma 1 and the fact that $\sum_{k=1}^{K_T} T_{k-1} \leq T$. ■

Lemma 9: The number of episodes K_T is bounded as follows:

$$K_T \leq 2\sqrt{\bar{\mathsf{S}}_n T \log(T)}.$$

Proof: Define macro episodes with start times t_{n_l}, $l = 1, 2, \ldots$ with $t_{n_1} = t_1$ and

$$t_{n_{l+1}} = \min\{t_k > t_{n_l} : N^i_{t_k}(s^i, a^i) > 2N^i_{t_{k-1}}(s^i, a^i) \text{ for some } (i, s^i, a^i)\}. \quad (12)$$

Let γ be the number of macro episodes until time T and define $n_{(\gamma+1)} = K_T + 1$. The rest of the proof is the same as [42, Eq. (8) in proof of Lemma 1] by which we get $K_T \leq \sqrt{2\gamma T}$.

For each arm-state-action tuple, define

$$\gamma^i(s^i, a^i) = |\{k \leq K_T | N^i_{t_k}(s^i, a^i) > 2N^i_{t_{k-1}}(s^i, a^i)\}|.$$

As a result $\gamma^i(s^i, a^i) \leq \log N^i_{T+1}(s^i, a^i)$. Note that for any $i \in [n]$, $N^i_{T+1}(s^i, a^i) \leq T$ and we have $2\mathsf{S}^i$ state-action pairs. Then, we have

$$\begin{aligned}\gamma &\leq 1 + \sum_{i\in[n]} \sum_{(s^i,a^i)} \gamma^i(s^i, a^i) \leq 1 + \sum_{i\in[n]} \sum_{(s^i,a^i)} \log N^i_{T+1}(s^i, a^i) \\ &= 1 + \sum_{i\in[n]} 2\mathsf{S}^i \log T \leq 2\bar{\mathsf{S}}_n \log T.\end{aligned}$$

■

B. Bound on $\mathcal{R}_1(T)$ (Lemma 2.2)

$\mathcal{R}_1(T)$ is a telescoping sum, which can be simplified as follows:

$$\begin{aligned}\mathcal{R}_1(T) &= \mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left[\boldsymbol{V}_k(\boldsymbol{S}_{t+1}) - \boldsymbol{V}_k(\boldsymbol{S}_t)\right]\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{K_T} \left[\boldsymbol{V}_k(\boldsymbol{S}_{t_{k+1}}) - \boldsymbol{V}_k(\boldsymbol{S}_{t_k})\right]\right] \leq 2\frac{\alpha R_{\max}}{1-\lambda^*}\mathbb{E}[K_T]\end{aligned}$$

where the last inequality uses Lemma 1. The result then follows by substituting the value of K_T from Lemma 9.

C. Bound on $\mathcal{R}_2(\mathsf{T})$ (Lemma 2.3)

1) Notation.: For any arm $i \in [n]$, let $N_t^i(s^i, a^i, s_+^i)$ denote the number of times $(S_\tau, A_\tau, S_{\tau+1})$ is equal to (s^i, a^i, s_+^i) until time t. Let $\hat{P}_t^i(s_+^i|s^i, a^i) = N_t^i(s^i, a^i, s_+^i)/\big(1 \vee N_t^i(s^i, a^i)\big)$ denote the empirical distribution based on observations up to time t. For a given $\delta \in (0, 1)$, we define

$$\epsilon_\delta^i(\ell) = \sqrt{\frac{2\mathsf{S}^i \log(1/\delta)}{1 \vee \ell}}. \tag{13}$$

2) Some preliminary results.: In this section, we state some preliminary properties.

Lemma 10: Let $p, q \in \Delta(\boldsymbol{\mathcal{S}})$. Then, for any function $f : \boldsymbol{\mathcal{S}} \to \mathbb{R}$, we have $\big|\langle f, p\rangle - \langle f, q\rangle\big| \le 0.5 \operatorname{span}(f) \big\|p - q\big\|_1$.

Proof: Let $\bar{f} = (\max f + \min f)/2$. Then

$$\begin{aligned}\big|\langle f, p\rangle - \langle f, q\rangle\big| &= \big|\langle f - \bar{f}, p - q\rangle\big| \le \|f - \bar{f}\|_\infty \big|\langle \mathbf{1}, p - q\rangle\big| \\ &\le \tfrac{1}{2} \operatorname{span}(f) \big\|p - q\big\|_1.\end{aligned}$$

■

Lemma 11: Consider any arm i, episode k, $\delta \in (0, 1)$, $\ell > 1$, and state-action pair (s^i, a^i). Define events $\mathcal{E}_\ell^i = \{N_{t_k}^i(s^i, a^i) = \ell\}$, $\mathcal{F}^i = \{\|P^i(\cdot\,|s^i, a^i) - \hat{P}_{t_k}^i(\cdot\,|s^i, a^i)\|_1 \le \epsilon_\delta(N_{t_k}^i(s^i, a^i))\}$, and $\mathcal{F}_k^i = \{\|P_k^i(\cdot\,|s^i, a^i) - \hat{P}_{t_k}^i(\cdot\,|s^i, a^i)\|_1 \le \epsilon_\delta(N_{t_k}^i(s^i, a^i))\}$. Then, we have

$$\mathbb{P}\Big(\big\|P^i - \hat{P}_{t_k}^i(\cdot\,|s^i, a^i)\big\|_1 > \epsilon_\delta^i(\ell) \;\Big|\; \mathcal{E}_\ell^i\Big) \le \delta.$$

The above inequality implies that

$$\mathbb{E}\Big[\big\|P^i - \hat{P}_{t_k}^i(\cdot\,|s^i, a^i)\big\|_1 \;\Big|\; \mathcal{F}^i\Big] \le \mathbb{E}[\epsilon_\delta^i(\ell)|\mathcal{F}^i] + 2\delta.$$

A similar bound holds if $(P^i, \mathcal{F}^i)$ is replaced by $(P_k^i, \mathcal{F}_k^i)$.

Proof: Given arm i, state s^i of the arm and action a^i chosen for the arm, we know from [58] that for any $\varepsilon > 0$, the $L1$-deviation of the true and the empirical distributions over $\mathcal{S}^i$ with $N_{t_k}^i(s^i, a^i) = \ell$ samples is bounded by

$$\begin{aligned}&\mathbb{P}\Big(\|P^i(\cdot\,|s^i, a^i) - \hat{P}_{t_k}^i(\cdot\,|s^i, a^i)\|_1 \ge \varepsilon \;\Big|\; \mathcal{E}_\ell^i\Big) \\ &\qquad \le 2^{\mathsf{S}^i} \exp\Big(-\frac{\ell\varepsilon^2}{2}\Big) < \exp\Big(\mathsf{S}^i - \frac{\ell\varepsilon^2}{2}\Big).\end{aligned}$$

Let $\delta = \exp(\mathsf{S}^i - \ell\varepsilon^2/2)$. Note that, $\mathsf{S}^i \ge 2$, therefore $\mathsf{S}^i + \log(1/\delta) \le \mathsf{S}^i \log(1/\delta)$. Hence,

$$\mathbb{P}\left(\big\|P^i(\cdot\,|s^i, a^i) - \hat{P}_{t_k}^i(\cdot\,|s^i, a^i)\big\|_1 > \sqrt{\frac{2\mathsf{S}^i \log(1/\delta)}{1 \vee \ell}} \;\middle|\; \mathcal{E}_\ell^i\right) < \delta.$$

The next result is driven by showing that $P((\mathcal{F}^i)^c) \le \delta$ and

$$\begin{aligned}&\mathbb{E}\Big[\big\|P^i(\cdot\,|s^i, a^i) - \hat{P}_{t_k}^i(\cdot\,|s^i, a^i)\big\|_1 \;\Big|\; \mathcal{F}^i\Big] \\ &\qquad \le 2\delta + \mathbb{E}[\epsilon_\delta^i(N_{t_k}^i(s^i, a^i))|\mathcal{F}^i],\end{aligned}$$

See [55] for details. ■

Lemma 12: For any episode k, and $\delta \in (0, 1)$, we have

$$\begin{aligned}&\mathbb{E}\Big[\big\|\boldsymbol{P}_\star(\cdot|\boldsymbol{s}, \boldsymbol{a}) - \boldsymbol{P}_k(\cdot|\boldsymbol{s}, \boldsymbol{a})\big\|_1\Big] \\ &\quad \le 4n\delta + \sum_{i\in[n]} \sum_{f\in\{\mathcal{F}^i, \mathcal{F}_k^i\}} \mathbb{E}\big[\epsilon_\delta^i(N_{t_k}^i(s^i, a^i))|f\big].\end{aligned}$$

Proof: The proof is as follows:

$$\begin{aligned}&\mathbb{E}\Big[\big\|\boldsymbol{P}_\star(\cdot|\boldsymbol{s}, \boldsymbol{a}) - \boldsymbol{P}_k(\cdot|\boldsymbol{s}, \boldsymbol{a})\big\|_1\Big] \\ &\overset{(a)}{=} \sum_{i\in[n]} \mathbb{E}\Big[\big\|P^i(\cdot\,|s^i, a^i) - P_k^i(\cdot\,|s^i, a^i)\big\|_1\Big] \\ &\overset{(b)}{\le} \sum_{i\in[n]} \mathbb{E}\Big[\big\|P^i(\cdot\,|s^i, a^i) - \hat{P}_{t_k}^i(\cdot\,|s^i, a^i)\big\|_1 \\ &\qquad + \big\|P_k^i(\cdot\,|s^i, a^i) - \hat{P}_{t_k}^i(\cdot\,|s^i, a^i)\big\|_1\Big]\end{aligned}$$

where (a) follows from [34, Lemma 13] and (b) follows from triangle inequality. The result then follows from Lemma 11. ■

3) Bounding $\mathcal{R}_2(\mathsf{T})$.: Now, consider the inner summation in the expression for $\mathcal{R}_2(T)$:

$$\begin{aligned}&\mathbb{E}\Big[\big\langle \boldsymbol{P}_k(\cdot\,|\boldsymbol{S}_t, \boldsymbol{A}_t), \boldsymbol{V}_k\big\rangle - \boldsymbol{V}_k(\boldsymbol{S}_{t+1})\Big] \\ &\overset{(a)}{\le} \mathbb{E}\Big[\tfrac{1}{2} \operatorname{span}(\boldsymbol{V}_k) \big\|\boldsymbol{P}_k(\cdot|\boldsymbol{S}_t, \boldsymbol{A}_t) - \boldsymbol{P}_\star(\cdot|\boldsymbol{S}_t, \boldsymbol{A}_t)\big\|_1\Big] \\ &\overset{(b)}{\le} \frac{\alpha R_{\max}}{1 - \lambda^*} \mathbb{E}\Big[\big\|\boldsymbol{P}_k(\cdot|\boldsymbol{S}_t, \boldsymbol{A}_t) - \boldsymbol{P}_\star(\cdot|\boldsymbol{S}_t, \boldsymbol{A}_t)\big\|_1\Big] \qquad (14)\end{aligned}$$

where (a) follows from Lemma 10 and (b) follows from Lemma 1. Then, by Lemma 12, we have

$$\begin{aligned}&\mathcal{R}_2(T) = \mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \big[\boldsymbol{P}_k \boldsymbol{V}_k\big](\boldsymbol{S}_t) - \boldsymbol{V}_k(\boldsymbol{S}_{t+1})\Big] \\ &\le \frac{\alpha R_{\max}}{1 - \lambda^*} \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \Big(4n\delta + \sum_{(i,f)\in\mathcal{D}_k} \mathbb{E}\big[\epsilon_\delta^i(N_{t_k}^i(S_t^i, A_t^i))|f\big]\Big) \qquad (15)\end{aligned}$$

where $\mathcal{D}_k = \{(i, f) : i \in [n], f \in \{\mathcal{F}, \mathcal{F}_k^i\}\}$.

For the first inner term of (15), we have

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} 4n\delta = \sum_{t=1}^{T} 4n\delta = 4n\delta T. \tag{16}$$

For the second inner term of (15), fix $(i, f) \in \mathcal{D}_k$ and let $\bar{\delta} = \sqrt{2\mathsf{S}^i \log(1/\delta)}$. Note $\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} = \sum_{t=1}^{T}$. Then,

$$\begin{aligned}&\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \mathbb{E}\left[\frac{\bar{\delta}}{\sqrt{1 \vee N_{t_k}^i(S_t^i, A_t^i)}} \;\middle|\; f\right] \\ &= \bar{\delta} \sum_{(s^i,a^i)} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(S_t^i = s^i, A_t^i = a^i) \sqrt{\frac{1}{1 \vee N_t^i(s^i, a^i)}} \;\middle|\; f\right] \\ &= \bar{\delta} \sum_{(s^i,a^i)} \mathbb{E}\left[\mathbb{1}(N_{T+1}^i(s^i, a^i) > 0) + \sum_{j=1}^{N_{T+1}^i(s^i,a^i)-1} \frac{1}{\sqrt{j}} \;\middle|\; f\right] \\ &\le \bar{\delta} \sum_{(s^i,a^i)} \mathbb{E}\left[\mathbb{1}(N_{T+1}^i(s^i, a^i) > 0) + 2\sqrt{N_{T+1}^i(s^i, a^i)} \;\middle|\; f\right] \\ &\le \bar{\delta} \sum_{(s^i,a^i)} 3\mathbb{E}\left[\sqrt{N_{T+1}^i(s^i, a^i)} \;\middle|\; f\right] \\ &\overset{(a)}{\le} 3\bar{\delta}\mathbb{E}\left[\sqrt{2\mathsf{S}^i \sum_{(s^i,a^i)} N_{T+1}^i(s^i, a^i)} \;\middle|\; f\right]\end{aligned}$$

$$\overset{(b)}{=} 6\mathsf{S}^i\sqrt{T\log(1/\delta)} \tag{17}$$

where (a) uses Cauchy-Schwartz inequality and (b) uses the fact that $\sum_{(x^i,a^i)} N^i_{T+1}(x^i,a^i) = T$. Adding (17) over $(i,f)\in\mathcal{D}_k$, we have

$$\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\sum_{(i,f)\in\mathcal{D}_k}\mathbb{E}\big[\epsilon^i_\delta(N^i_{t_k}(S^i_t,A^i_t))|f\big] \le 12\bar{\mathsf{S}}_n\sqrt{T\log(1/\delta)}. \tag{18}$$

Finally, by setting $\delta = 1/T$, and substituting (17) and (16) in (15), we get the final result.

D. Bound on $\mathcal{R}_2(\mathsf{T})$ (Lemma 3)

1) Some preliminary results.: For any arm $i\in[n]$, let $N^i_t(s^i,a^i,s^i_+)$ and $\hat{P}^i_t(s^i_+|s^i,a^i)$ denote the same variables as defined in Sec. C.1.

Lemma 13: For any Lipschitz function $f:\mathcal{X}\to\mathbb{R}$ with Lipschitz coefficient L_f, and any probability measures ζ_1 and ζ_2 on $(\mathcal{X},d_X)$ we have

$$\left|\sum_{x\in\mathcal{X}} f(x)\zeta_1(x) - \sum_{x\in\mathcal{X}} f(x)\zeta_2(x)\right| \le L_f\mathcal{K}(\zeta_1,\zeta_2).$$

Proof: The result is immediately derived from the definition of Kantorovich distance. ∎

Lemma 14 (From Theorem 2 of [59]): Let ν denote a probability measure on $(\mathbb{R},|\cdot|)$ and let $\hat{\nu}_n$ denote the estimated probability measure by n samples from ν. Then, for all $n\ge 1$ and all $\epsilon>0$, there exist constants C and c which depend on ν such that $\mathbb{P}\left(\mathcal{K}(\nu,\hat{\nu}_n)\ge\epsilon\right) \le C\exp(-cn\epsilon)\mathbb{1}(\epsilon\le 1) + C\exp(-cn\epsilon^2)\mathbb{1}(\epsilon>1)$.

Proof: The lemma follows directly by applying [59, Theorem 2] and setting $d=1$, $p=1$ and $\alpha=2$ which satisfies condition (C1) of [59]. ∎

Let $\epsilon_\delta(\ell) = \sqrt{\log(C/\delta)/(c(1\vee\ell))}$.

Lemma 15: Consider any arm i, any episode k, $\delta\in(0,1)$, $\ell>1$, and state-action pair (s^i,a^i). Define events $\mathcal{F}^i = \{\mathcal{K}(P^i,\hat{P}^i_{t_k}(\cdot\,|s^i,a^i)) \le \epsilon_\delta(N^i_{t_k}(s^i,a^i))\}$, and $\mathcal{F}^i_k = \{\mathcal{K}(\tilde{P}^i_k,\hat{P}^i_{t_k}(\cdot\,|s^i,a^i))\|_1 \le \epsilon_\delta(N^i_{t_k}(s^i,a^i))\}$. we have

$$\mathbb{P}\Big(\mathcal{K}(P^i,\hat{P}^i_{t_k}(\cdot\,|s^i,a^i)) > \epsilon^i_\delta(\ell)\Big) \le \delta,$$

Furthermore, the above inequality implies

$$\mathbb{E}\Big[\mathcal{K}(P^i,\hat{P}^i_{t_k}(\cdot\,|s^i,a^i)) \;\Big|\; \mathcal{E}^i\Big] \le \mathbb{E}[\epsilon^i_\delta(\ell) \;\Big|\; \mathcal{F}^i] + 2\,\mathrm{diam}(\mathcal{S}^i)\delta.$$

A similar bound holds if $(P^i,\mathcal{F}^i)$ is replaced by $(P^i_k,\mathcal{F}^i_k)$.

Proof: The result follows by using Lemmas 13 and 14 and a similar approach as the proof of Lemma 11. ∎

Lemma 16: For any episode k, and $\delta\in(0,1)$, we have

$$\mathbb{E}\Big[\mathcal{K}(\boldsymbol{P}_\star(\cdot|\boldsymbol{s},\boldsymbol{a}),\boldsymbol{P}_k(\cdot|\boldsymbol{s},\boldsymbol{a}))\Big] \le 4n\mathsf{D}_{\max}\delta + \sum_{f\in\{\mathcal{F}^i,\mathcal{F}^i_k\}}\sum_{i\in[n]}\mathbb{E}\Big[\epsilon^i_\delta(N^i_{t_k}(s^i,a^i)) \;\Big|\; f\Big].$$

Proof: The proof is similar to that of Lemma 12, where we use Kantorovich distance instead of total variation distance. The equivalent of equality (a) (in the proof of Lemma 12) follows from [56, Lemma 4], and the rest of the argument follows from Lemma 15. ∎

2) Bounding $\mathcal{R}_2(\mathsf{T})$.: First, consider the inner summation in the expression for $\mathcal{R}_2(T)$:

$$\mathbb{E}\Big[\big\langle \boldsymbol{P}_k(\cdot\,|\boldsymbol{S}_t,\boldsymbol{A}_t),\boldsymbol{V}_k\big\rangle - \boldsymbol{V}_k(\boldsymbol{S}_{t+1})\Big] \overset{(a)}{\le} \boldsymbol{L}_v\mathbb{E}\Big[\mathcal{K}\left(\boldsymbol{P}_\star(\cdot|\boldsymbol{S}_t,\boldsymbol{A}_t) - \boldsymbol{P}_k(\cdot|\boldsymbol{S}_t,\boldsymbol{A}_t)\right)\Big] \tag{19}$$

where (a) follows from Lemma 13. Then, by Lemma 16, we have

$$\mathcal{R}_2(T) = \mathbb{E}\Big[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\big[\boldsymbol{P}_k\boldsymbol{V}_k\big](\boldsymbol{S}_t) - \boldsymbol{V}_k(\boldsymbol{S}_{t+1})\Big] \le \boldsymbol{L}_v\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(4n\mathsf{D}_{\max}\delta + \sum_{(i,f)\in\mathcal{D}_k}\mathbb{E}\Big[\epsilon^i_\delta(N^i_{t_k}(S^i_t,A^i_t)) \;\Big|\; f\Big]\right). \tag{20}$$

where $\mathcal{D}_k = \{(i,f): i\in[n], f\in\{\mathcal{F},\mathcal{F}^i_k\}\}$.

For the first inner term of (20), we have

$$\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} 4n\mathsf{D}_{\max}\delta = \sum_{t=1}^{T} 4n\mathsf{D}_{\max}\delta = 2n\mathsf{D}_{\max}\delta T. \tag{21}$$

For the second inner term of (20), we can follow an argument similar to (18) to show that

$$\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\sum_{(i,f)\in\mathcal{D}_k}\mathbb{E}\Big[\epsilon^i_\delta(N^i_{t_k}(S^i_t,A^i_t)) \;\Big|\; f\Big] \le 12\sqrt{\frac{\bar{\mathsf{S}}_n\log(C/\delta)T}{c}}. \tag{22}$$

See [55] for details. Finally, by setting $\delta = 1/T$, and substituting (22) and (21) in (20), we get the result.

Nima Akbarzadeh (S'17) is a PhD student in the Electrical and Computer Engineering, McGill University, Canada. He received the B.Sc. degree in Electrical and Computer Engineering from Shiraz University, Iran, in 2014, the M.Sc. in Electrical and Electronics Engineering from Bilkent University, Turkey, in 2017. He is a recipient of 2020 FRQNT PhD Scholarship. His research interests include stochastic control, reinforcement learning and multi-armed bandits.

Aditya Mahajan (S'06-M'09-SM'14) is an Associate Professor in the department of Electrical and Computer Engineering, McGill University, Montreal, Canada. He received the B.Tech degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, India and the MS and PhD degrees in Electrical Engineering and Computer Science from the University of Michigan, Ann Arbor, USA. His principal research interests are learning and control of centralized and decentralized stochastic systems.