Sensitivity of Whittle index policy to model approximation

Amit Sinha, Aditya Mahajan

Department of Electrical and Computer Engineering
McGill University, Montreal, Canada.

Abstract

We consider a restless multi-armed bandit problem where the model of each arm is known approximately and provide bounds on the loss of performance in using the Whittle index policy of the approximate model. The bounds depend on the approximation errors in modeling each arm, properties of the transition dynamics and the per-step reward of each arm, and the degree of suboptimality of the Whittle index policy in the approximate model.

1. Introduction

Restless multi-armed bandits (RMAB), introduced in [36], are an important modeling framework for a variety of scheduling and resource allocation problems arising in communication networks, power systems, and machine maintenance. Such problems can be modeled as Markov decision process (MDP) but obtaining an optimal solution suffers from the curse of dimensionality. In general, the MDP solution has exponential complexity in the number of alternatives. The RMAB framework provides a scalable heuristic solution, known as the Whittle index policy, which has linear complexity in the number of alternatives (which are called *arms* in the restless bandit literature). The Whittle index policy is optimal in some settings (e.g., when the arms which are not selected remain frozen [15], when the number of arms is asymptotically large [34], and when the model satisfies some separation conditions [23]), and performs close to optimal in a variety of applications [1, 4, 6, 16, 17, 28].

However, the current literature assumes that the model of each arm is known perfectly. This is not always true, especially in applications where the models of the arms are estimated based on data. We are interested in the following question: *how sensitive is the Whittle index policy to model approximation?* In particular, if we make some approximation errors in modeling the rewards and dynamics of each arm, what is the loss in performance in taking a certainty equivalence approach and following the Whittle index policy of the approximate model? This question is also relevant for restless bandits with continuous state space, where model approximation may be required to compute the Whittle index.

For *rested* multi-armed bandits (i.e., when only one arm can be activated at each time, and the arms which are not activated remain frozen), it is known that a policy which approximates the Gittins index is approximately optimal [20, 22]. Thus, the sensitivity question reduces to the question of the senstivity of the Gittins index to model parameters. However, the result and the proof technique of [20] rely on specific features of the rested MAB settings and cannot be directly generalized to *restless* MABs. There are also other results in the literature on approximate computation of Gittins index [7], but they are also not applicable to the restless setting.

The notion of sensitivity considered in this paper is similar to the notion of certainty equivalence used in stochastic control. In our setting, the decision maker has an approximate model and wants to use the optimal policy of the approximate model in the true model. A related notion is that of robustness, where instead of choosing the optimal policy of the approximate model, the decision maker chooses a policy that optimizes the worst case performance over all realizations of the true model [19, 27]. Such robust formulations for Gittins index

Email addresses: amit.sinha@mail.mcgill.ca (Amit Sinha), aditya.mahajan@mcgill.ca (Aditya Mahajan)

have been considered in [10, 13, 22]. These results have been generalized to a certain class of partially observed models in [21]. As mentioned earlier, the notion of robustness is different from the notion of sensitivity or certainty equivalence that we consider here.

Recently, there has been a significant interest in learning Whittle index policies for RMAB [2, 5, 8, 14, 25, 31]. Most of these learn the Whittle index by using reinforcement learning to learn a Q-function of an auxiliary MDP associated with the computation of Whittle index. But, these papers do not provide an explicit answer to the sensitivity question that we are interested in.

Our main contributions are the following.

1. We formulate the question of sensitivity of the Whittle index policy to model approximation. In particular, we formalize how to define model approximation of an arm and characterize the sensitivity of Whittle index policy in terms of approximation errors in modeling individual arms and a property of the value function of the optimal policy.

2. Our results depend on the choice of metric on probability spaces. We consider a class of metrics knows as integral probability metrics (IPMs) and focus on two IPMs: total variation distance and Wasserstein distance. For these IPMs, we provide a computable upper bound on the sensitivity of the Whittle index policy which depends on the approximation errors in modeling individual arms and properties of the reward functions and transition kernels of the arms.

The rest of the paper is organized as follows. In Sec. 2, we present the model and the problem formulation and state the main results. We present some examples of our results in Sec. 3. In Sec. 4, we present the proofs of the main results and conclude in Sec. 5.

Notation Used

We use uppercase letters to denote random variables (e.g. S, A, etc.), lowercase letters to denote their realizations (e.g. s, a, etc.) and sans serif letters to denote sets (e.g. S, A, etc.). We also use superscripts (e.g. S^i, A^i, etc. for arm i) to denote quantities for a specific arm. For any set X, $\Delta(\mathsf{X})$ is used to denote the space of probability distributions on X. $\mathbb{P}$ and $\mathbb{E}$ denote the probability of an event and expectation of a random variable, respectively. For an integer n, we use $[n]$ to denote the set of integers from 1 to n.

Given a set S and a function $f \colon \mathsf{S} \to \mathbb{R}$, we use $\operatorname{span}(f)$ to denote the span of f, i.e., $\operatorname{span}(f) = \sup_{s,s' \in \mathsf{S}} |f(s) - f(s')|$ and we use $\|f\|_\infty$ to denote the supremum norm of function f, i.e., $\|f\|_\infty = \sup_{s \in \mathsf{S}} f(s)$.

When (S, d) is a metric space we use $\operatorname{Lip}(f)$ to denote the Lipschitz constant of f, i.e.,

$$\operatorname{Lip}(f) = \sup_{s,s' \in \mathsf{S}} \frac{|f(s) - f(s')|}{d(s,s')}.$$

If this constant exists and is finite, then f is said to be $\operatorname{Lip}(f)$-Lipschitz.

2. Problem formulation and main results

The results of this paper are applicable to models with discrete or continuous state spaces. For ease of exposition, we present the model and results for continuous state spaces. They can be easily translated to models with discrete state spaces.

2.1. Restless multi-armed bandits

A restless multi-armed bandit (RMAB) is a decision making problem where there are n alternatives or arms. Each arm i, $i \in [n]$, is a controlled Markov process $\alpha^i = \langle \mathsf{S}^i, \{0,1\}, \{p^i(a)\}_{a \in \{0,1\}}, r^i \rangle$, where S^i denotes the state space which is assumed to be a compact set, $\{0,1\}$ is the action space, $p^i(a)$, $a \in \{0,1\}$, denotes the transition density from S^i to S^i when action a is chosen, and $r^i : \mathsf{S}^i \times \{0,1\} \mapsto \mathbb{R}$ denotes the per-step reward which is assumed to be uniformly bounded and continuous in S^i. For some of the results, we will assume that, for each arm $i \in [n]$, the state space S^i is a metric space and use d^i to denote the metric on S^i.

The system operates in discrete time. We use $S^i_t \in \mathsf{S}^i$ and $A^i_t \in \{0,1\}$ to denote the state and action of arm i at time t. We use $\boldsymbol{S}_t = (S^1_t, \dots, S^n_t)$ and $\boldsymbol{A}_t = (A^1_t, \dots, A^n_t)$ to denote the global state and actions of all arms at time t. Each component of the global state evolves in a controlled Markov manner independently of other components. In particular, for any measurable subsets $\mathsf{B}^i \subset \mathsf{S}^i$, $i \in [n]$, we have

$$\begin{aligned}\mathbb{P}\bigg(\boldsymbol{S}_{t+1} \in \prod_{i\in[n]} \mathsf{B}^i \;\bigg|\; \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}, \boldsymbol{A}_{1:t} = \boldsymbol{a}_{1:t}\bigg) \\ = \prod_{i\in[n]}\bigg[\int_{\mathsf{B}^i} p^i(s^i_{t+1} \mid s^i_t, a^i_t) ds^i_{t+1}\bigg].\end{aligned}$$

At each time, a decision maker observes the global state $\boldsymbol{S}_t$ and can *activate* (i.e., select action $A^i_t = 1$) for at most $m < n$ arms. The decision maker chooses its actions according to a time-homogeneous Markov policy $\pi\colon \mathsf{S} \to \mathsf{A}(m)$, where $\mathsf{S} = \prod_{i\in[n]} \mathsf{S}^i$ denotes the set of all global states and $\mathsf{A}(m) \coloneqq \big\{\boldsymbol{a} \in \{0,1\}^n : \|\boldsymbol{a}\|_1 \le m\big\}$ denotes the set of feasible actions. The performance of any Markov policy π starting from an initial state $\boldsymbol{s}_0 \in \mathsf{S}$ is given by

$$V^\pi(\boldsymbol{s}_0) = Q^\pi(\boldsymbol{s}_0, \pi(\boldsymbol{s}_0)), \tag{1}$$

where

$$\begin{aligned}&Q^\pi(\boldsymbol{s}_0, \boldsymbol{a}_0) \\ &\quad = \mathbb{E}^\pi\bigg[\sum_{t=0}^\infty \gamma^t \sum_{i\in[n]} r^i(S^i_t, A^i_t) \;\bigg|\; \boldsymbol{S}_0 = \boldsymbol{s}_0, \boldsymbol{A}_0 = \boldsymbol{a}_0\bigg],\end{aligned} \tag{2}$$

where $\gamma \in (0,1)$ denotes the discount factor and $r^i \in [0,1]$. The objective is to find a Markov policy π which maximizes $V^\pi(\boldsymbol{s}_0)$.

The decision problem formulated above is a Markov decision process (MDP) and can be solved using dynamic programming. However, the dynamic programming solution suffers from the curse of dimensionality because both the state space S and action space $\mathsf{A}(m)$ grow exponentially with the number of arms. To avoid the curse of dimensionality, a popular heuristic is to use the Whittle index policy [36], which has a linear complexity in the number of arms. An overview of the Whittle index policy can be found in [38, Sec. 3.3], [24, Sec. 3.5], and [35, 36].

2.2. Problem formulation: Model approximation in RMAB

We start by defining a class of metrics on probability measures known as integral probability metrics (IPM) [26].

Definition 1. *Let $(\mathsf{X}, \mathcal{G})$ be a measurable space and $\mathfrak{F}$ denote a class of uniformly bounded measurable functions on $(\mathsf{X}, \mathcal{G})$. The integral probability metric (IPM) between two probability distributions $\mu, \nu \in \Delta(\mathsf{X})$ with respect to the function class $\mathfrak{F}$ is defined as*

$$d_{\mathfrak{F}}(\mu, \nu) \coloneqq \sup_{f\in\mathfrak{F}}\bigg|\int_{\mathsf{X}} f d\mu - \int_{\mathsf{X}} f d\nu\bigg|.$$

Some examples of IPM are total variation distance, Wasserstein distance, Kolmogorov distance, Bounded-Lipschitz distance, and maximum mean discrepancy. For total variation distance, $\mathfrak{F} = \{f : \frac{1}{2}\operatorname{span}(f) \le 1\} =: \mathfrak{F}^{\mathrm{TV}}$; for Wasserstein distance, $\mathfrak{F} = \{f : \operatorname{Lip}(f) \le 1\} =: \mathfrak{F}^{\mathrm{W}}$. We refer the reader to [33] for details about other examples.

Given a function class $\mathfrak{F}$ and a function f (not necessarily in $\mathfrak{F}$), the Minkowski functional [32] of f with respect to $\mathfrak{F}$ is defined as:

$$\rho_{\mathfrak{F}}(f) \coloneqq \inf\{\rho \in \mathbb{R}_{>0} : \rho^{-1} f \in \mathfrak{F}\}. \tag{3}$$

When $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}}$ (i.e., $d_{\mathfrak{F}}$ is the total variation distance), $\rho_{\mathfrak{F}}(f) = \frac{1}{2}\operatorname{span}(f)$; and when $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$ (i.e., $d_{\mathfrak{F}}$ is the Wasserstein distance), $\rho_{\mathfrak{F}}(f) = \operatorname{Lip}(f)$. A key implication of the definition of Minkowski functional is the following: for any function f, not necessarily in function class $\mathfrak{F}$,

$$\bigg|\int_{\mathsf{X}} f d\mu - \int_{\mathsf{X}} f d\nu\bigg| \le \rho_{\mathfrak{F}}(f) \cdot d_{\mathfrak{F}}(\mu, \nu), \tag{4}$$

We now formalize the notion of approximate restless bandit model.

Definition 2. *Consider two arms $\alpha = \langle \mathsf{S}, \{0,1\}, \{p(a)\}_{a\in\{0,1\}}, r\rangle$ and $\hat{\alpha} = \langle \mathsf{S}, \{0,1\}, \{\hat{p}(a)\}_{a\in\{0,1\}}, \hat{r}\rangle$ defined on the same state space. Given a function space $\mathfrak{F}$ and positive constants ε and δ, arm $\hat{\alpha}$ is called an (ε, δ)-*approximation *of arm α if for all $s \in \mathsf{S}$ and $a \in \{0,1\}$:*

$$\big|r(s,a) - \hat{r}(s,a)\big| \le \varepsilon, \quad d_{\mathfrak{F}}\big(p(\cdot|s,a), \hat{p}(\cdot|s,a)\big) \le \delta.$$

We fix the function space $\mathfrak{F}$ and consider the following setup.

Approximation setup. *Given a* RMAB *$\{\alpha^i\}_{i\in[n]}$, where $\alpha^i = \langle \mathsf{S}^i, \{0,1\}^i, \{p^i(a)\}_{a\in\{0,1\}}, r^i\rangle$, consider an approximate* RMAB *$\{\hat{\alpha}^i\}_{i\in[n]}$, where $\hat{\alpha}^i = \langle \mathsf{S}^i, \{0,1\}^i, \{\hat{p}^i(a)\}_{a\in\{0,1\}}, \hat{r}^i\rangle$ and for each $i \in [n]$, arm $\hat{\alpha}^i$ is an $(\varepsilon^i, \delta^i)$-approximation of arm α^i.*

For any policy $\pi\colon \mathsf{S} \to \mathsf{A}(m)$ and initial state $\boldsymbol{s}$, let $V^\pi(\boldsymbol{s})$ denote the performance of π in RMAB $\{\alpha^i\}_{i\in[n]}$ and let $\hat{V}^\pi(\boldsymbol{s})$ denote the performance of policy π in RMAB $\{\hat{\alpha}^i\}_{i\in[n]}$. Let π^* denote the optimal policy for RMAB $\{\alpha^i\}_{i\in[n]}$ and let $\hat{\pi}^*$ and $\hat{\mu}$ denote the optimal policy and the Whittle index policy[1] for RMAB $\{\hat{\alpha}^i\}_{i\in[n]}$.

For the Whittle index policy to be applicable, the model must satisfy a technical condition known as indexability [36]. So we impose the following assumption.

Assumption 1. *In the approximaion setup, all arms $\{\hat{\alpha}^i\}_{i\in[n]}$ are indexable.*

We are interested in the following approximation characterization.

Problem 1. *In the approximation setup described above, under Assumption 1, characterize the approximation error $\|V^{\pi^*} - V^{\hat{\mu}}\|_\infty$ (which is the suboptimality gap of using the Whittle index policy of the approximate model in the true model) in terms of $\|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty$ (which is the suboptimality gap of using the Whittle index policy of the approximate model in the approximate model) and the approximation errors $\{(\varepsilon^i, \delta^i)\}_{i\in[n]}$.*

2.3. Main results

We first define a property of an arm, which is needed in our main result.

Definition 3. *Consider the function class $\mathfrak{F}^{\mathrm{W}}$, an arm $\alpha^i = \langle \mathsf{S}^i, \{0,1\}^i, \{p^i(a)\}_{a\in\{0,1\}}, r^i\rangle$ and a metric d^i on S^i. If*

$$\mathrm{L}_{r^i} := \sup_{\substack{s,s'\in\mathsf{S}^i\\ a\in\{0,1\}}} \frac{|r^i(s,a) - r^i(s',a)|}{d^i(s,s')} < \infty,$$

$$\mathrm{L}_{p^i} := \sup_{\substack{s,s'\in\mathsf{S}^i\\ a\in\{0,1\}}} \frac{d_{\mathfrak{F}^{\mathrm{W}}}(p^i(\cdot\mid s,a), p^i(\cdot\mid s',a))}{d^i(s,s')} < \infty,$$

then the arm α^i is said to be $(\mathrm{L}_{r^i}, \mathrm{L}_{p^i})$-Lipschitz.

Now we present our main result. For any Markov policy π, define

$$\beta^{\pi}_{\mathfrak{F}} := \frac{\varepsilon + \gamma\delta\rho_{\mathfrak{F}}(\hat{V}^{\pi})}{1-\gamma},$$

where $(\varepsilon,\delta) = \left(\sum_{i\in[n]}\varepsilon^i, \sum_{i\in[n]}\delta^i\right)$. Then we have the following.

Theorem 1. *For the approximation setup of Sec. 2.2, under Assumption 1, we have*

$$\|Q^{\pi^*} - Q^{\hat{\mu}}\|_\infty \le 3\beta^{\hat{\pi}^*}_{\mathfrak{F}} + \beta^{\hat{\mu}}_{\mathfrak{F}} + \|\hat{Q}^{\hat{\pi}^*} - \hat{Q}^{\hat{\mu}}\|_\infty \tag{5}$$

and

$$\|V^{\pi^*} - V^{\hat{\mu}}\|_\infty \le 3\beta^{\hat{\pi}^*}_{\mathfrak{F}} + \beta^{\hat{\mu}}_{\mathfrak{F}} + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty. \tag{6}$$

The proof is given in Sec.4. The above bounds depend on the properties of the optimal value function V^{π^*}, which can be difficult to compute. We now present looser upper bounds which do not explicitly depend on V^{π^*}.

Proposition 1. *When $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}}$ (i.e. $d_{\mathfrak{F}}$ is the total variation distance) and Assumption 1 holds, then we have*

$$\|Q^{\pi^*} - Q^{\hat{\mu}}\|_\infty \le \frac{4\varepsilon}{(1-\gamma)} + \frac{3\gamma\delta\,\mathrm{span}(\hat{\boldsymbol{r}})}{2(1-\gamma)^2} + \frac{\gamma\delta\,\mathrm{span}(\hat{V}^{\hat{\mu}})}{2(1-\gamma)} + \|\hat{Q}^{\hat{\pi}^*} - \hat{Q}^{\hat{\mu}}\|_\infty \tag{7}$$

and

$$\|V^{\pi^*} - V^{\hat{\mu}}\|_\infty \le \frac{4\varepsilon}{(1-\gamma)} + \frac{3\gamma\delta\,\mathrm{span}(\hat{\boldsymbol{r}})}{2(1-\gamma)^2} + \frac{\gamma\delta\,\mathrm{span}(\hat{V}^{\hat{\mu}})}{2(1-\gamma)} + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty, \tag{8}$$

where $(\varepsilon,\delta) = \left(\sum_{i\in[n]}\varepsilon^i, \sum_{i\in[n]}\delta^i\right)$ and $\mathrm{span}(\hat{\boldsymbol{r}}) \le \sum_{i\in[n]} \mathrm{span}(\hat{r}^i)$.

See Sec. 4.4 for proof.

Proposition 2. *When $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$ (i.e. $d_{\mathfrak{F}}$ is the Wasserstein distance), suppose Assumption 1 holds, and for each $i\in[n]$, arm $\hat{\alpha}^i$ is $(\mathrm{L}_{\hat{r}^i}, \mathrm{L}_{\hat{p}^i})$-Lipschitz with $\mathrm{L}_{\hat{p}^i} < \gamma^{-1}$, we have*

$$\|Q^{\pi^*} - Q^{\hat{\mu}}\|_\infty \le \frac{4\varepsilon}{(1-\gamma)} + \frac{3\gamma\delta\mathrm{L}_{\hat{\boldsymbol{r}}}}{(1-\gamma)(1-\gamma\mathrm{L}_{\hat{\boldsymbol{p}}})} + \frac{\gamma\delta\,\mathrm{Lip}(\hat{V}^{\hat{\mu}})}{(1-\gamma)} + \|\hat{Q}^{\hat{\pi}^*} - \hat{Q}^{\hat{\mu}}\|_\infty \tag{9}$$

and

$$\|V^{\pi^*} - V^{\hat{\mu}}\|_\infty \le \frac{4\varepsilon}{(1-\gamma)} + \frac{3\gamma\delta\mathrm{L}_{\hat{\boldsymbol{r}}}}{(1-\gamma)(1-\gamma\mathrm{L}_{\hat{\boldsymbol{p}}})} + \frac{\gamma\delta\,\mathrm{Lip}(\hat{V}^{\hat{\mu}})}{(1-\gamma)} + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty, \tag{10}$$

[1] We can also consider $\hat{\mu}$ to be any arbitrary heuristic policy here (the main results stay the same). We focus on the Whittle index policy for further discussions because it is the most commonly used heuristic for RMAB problems.

$$
\begin{aligned}
P(0) &= \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.3 & 0.3 \end{bmatrix}, \\
P(1) &= \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.3 & 0.3 & 0.4 \\ 0.2 & 0.2 & 0.6 \end{bmatrix}, \\
r &= \begin{bmatrix} 0.60 & 0.40 \\ 0.88 & 0.60 \\ 1.00 & 0.80 \end{bmatrix}, \\
&\text{Arm 1}
\end{aligned}
\qquad
\begin{aligned}
P(0) &= \begin{bmatrix} 0.1 & 0.6 & 0.3 \\ 0.2 & 0.7 & 0.1 \\ 0.1 & 0.8 & 0.1 \end{bmatrix}, \\
P(1) &= \begin{bmatrix} 0.50 & 0.40 & 0.10 \\ 0.30 & 0.60 & 0.10 \\ 0.25 & 0.55 & 0.20 \end{bmatrix}, \\
r &= \begin{bmatrix} 0.52 & 0.64 \\ 0.44 & 0.96 \\ 0.76 & 0.44 \end{bmatrix}. \\
&\text{Arm 2}
\end{aligned}
$$

(a) True Model

$$
\begin{aligned}
\hat{P}(0) &= \begin{bmatrix} 0.19 & 0.29 & 0.52 \\ 0.11 & 0.51 & 0.38 \\ 0.41 & 0.29 & 0.30 \end{bmatrix}, \\
\hat{P}(1) &= \begin{bmatrix} 0.39 & 0.39 & 0.22 \\ 0.29 & 0.29 & 0.42 \\ 0.21 & 0.19 & 0.60 \end{bmatrix}, \\
\hat{r} &= \begin{bmatrix} 0.596 & 0.404 \\ 0.872 & 0.596 \\ 0.996 & 0.792 \end{bmatrix}, \\
&\text{Arm 1}
\end{aligned}
\qquad
\begin{aligned}
\hat{P}(0) &= \begin{bmatrix} 0.09 & 0.62 & 0.29 \\ 0.21 & 0.69 & 0.10 \\ 0.12 & 0.79 & 0.09 \end{bmatrix}, \\
\hat{P}(1) &= \begin{bmatrix} 0.48 & 0.42 & 0.10 \\ 0.31 & 0.59 & 0.10 \\ 0.24 & 0.55 & 0.21 \end{bmatrix}, \\
\hat{r} &= \begin{bmatrix} 0.512 & 0.636 \\ 0.432 & 0.968 \\ 0.756 & 0.448 \end{bmatrix}. \\
&\text{Arm 2}
\end{aligned}
$$

(b) Approximate Model

Figure 1: The true and approximate model for the example of Sec. 3.1

where $(\varepsilon, \delta) = \left(\sum_{i \in [n]} \varepsilon^i, \sum_{i \in [n]} \delta^i\right)$, $\mathrm{L}_{\hat{\boldsymbol{r}}} \leq \max_{i \in [n]} \mathrm{L}_{\hat{r}^i}$ *and* $\mathrm{L}_{\hat{\boldsymbol{P}}} \leq \max_{i \in [n]} \mathrm{L}_{\hat{P}^i}$.

See Sec. 4.5 for proof.

Remark 1. *In order to compute the Lipschitz constant of* $V^{\hat{\mu}}$ *in (9) and (10), we need a metric on* S. *This metric is chosen as* $\boldsymbol{d}(\boldsymbol{s}, \boldsymbol{s}') = \sum_{i \in [n]} d^i(s^i, s'^i)$.

Remark 2. *In the rested case, the Whittle index policy reduces to the Gittins index policy and is optimal. Therefore, in (6),* $\hat{V}^{\hat{\pi}^*} = \hat{V}^{\hat{\mu}}$. *Thus, Theorem 1 also provides an approximation guarantee for the rested RMAB which is different from the stopping-time based approximation guarantee in [20].*

Remark 3. *The upper bound of Theorem 1 depends on the IPM in two ways. First, the parameter* δ *(i.e. the degree of closeness of the approximate dynamics to the true dynamics) depends on the IPM. In addition, the* $\rho_{\mathfrak{F}}(\cdot)$ *term depends on the choice of IPM. See Sec. 3.1 for an example on how the upper bound depends on the choice of the IPM.*

Remark 4. *In Theorem 1, we only require the approximate model to be indexable (Assumption 1). The original model is not required to be indexable. This is a useful feature in settings where the original model is not known and only an approximate model is available.*

3. Some illustrative examples

In this section, we provide some examples to illustrate our results.

3.1. An example with finite state space

Consider an RMAB with two arms $\alpha^i = \langle \mathsf{S}, \{0,1\}, \{P^i(a^i)\}_{a^i \in \{0,1\}}, r^i \rangle$, $i \in \{1,2\}$, where $\mathsf{S} = \{1,2,3\}$ shown in Fig. 1a. Suppose these arms are approximated by $\langle \mathsf{S}, \{0,1\}, \{\hat{P}^i(a^i)\}_{a^i \in \{0,1\}}, \hat{r}^i \rangle$ show in Fig. 1b. It can be verified that the approximate model is indexable. Thus, Assumption 1 is satisfied.

Let $\hat{\omega}^i(s)$ denote the Whittle index (for the approximate model) of arm i in state s. We compute these using the modified adaptive greedy algorithm [3], and they are given by

$$
\begin{aligned}
&\hat{\omega}^1(1) = -0.308, && \hat{\omega}^1(2) = -0.309, && \hat{\omega}^1(3) = -0.140, \\
&\hat{\omega}^2(1) = 0.009, && \hat{\omega}^2(2) = 0.547, && \hat{\omega}^2(3) = -0.410.
\end{aligned}
$$

The Whittle index policy $\hat{\mu}$ is given by

$$
\hat{\mu}(s^1, s^2) = \underset{i \in \{1,2\}}{\arg\max}\, \hat{\omega}^i(s^i). \tag{11}
$$

We are interested in bounding the performance loss in using the Whittle index policy for the approximate model, in the true model. For that matter, we first compute the value function of the Whittle index policy (in the true model) using the policy evaluation equation [29]. The value function is given by[2]

$$
V^{\hat{\mu}} = \begin{bmatrix} 16.172 & 16.562 & 16.165 \\ 16.474 & 16.864 & 16.401 \\ 16.509 & 16.899 & 16.638 \end{bmatrix}.
$$

Since the model is small, we can compute the optimal value function (of the true model), which we

[2] The value function $V^{\hat{\mu}}$ is a function from $\mathsf{S}^1 \times \mathsf{S}^2 \to \mathbb{R}$. We represent it as a matrix, where the (i,j)-th element corresponds to the value $V^{\hat{\mu}}(i,j)$.

do using the value iteration algorithm [29]. The optimal value function is given by

$$V^{\pi^*} = \begin{bmatrix} 16.386 & 16.777 & 16.647 \\ 16.691 & 17.081 & 16.951 \\ 16.725 & 17.116 & 16.986 \end{bmatrix}.$$

Thus, the Whittle index policy has a suboptimality gap of $\|V^{\pi^*} - V^{\hat{\mu}}\|_\infty = 0.550$. Note that in practice we do not have access to the true model, so we cannot compute the suboptimality gap $\|V^{\pi^*} - V^{\hat{\mu}}\|_\infty$. The results of Theorem 1 provide a method to bound the suboptimality gap.

We first compute the values of approximate errors (ε, δ) for arms 1 and 2 which are shown in Table 1 (for $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$, we use $d(s, s') = |s - s'|$ as the metric on S). We also compute the value function of the Whittle index policy and the optimal value function (for the approximate model). These are given by

Parameter	**Arm 1**	**Arm 2**	**Overall**
ε	0.008	0.008	0.016
$\delta^{\mathfrak{F}^{\mathrm{TV}}}$	0.02	0.02	0.04
$\delta^{\mathfrak{F}^{\mathrm{W}}}$	0.03	0.03	0.06

Table 1: Parameters involved in Theorem 1 for Example 3.1.

$$\hat{V}^{\hat{\mu}} = \begin{bmatrix} 16.142 & 16.534 & 16.133 \\ 16.430 & 16.822 & 16.361 \\ 16.473 & 16.865 & 16.587 \end{bmatrix}$$

and

$$\hat{V}^{\hat{\pi}^*} = \begin{bmatrix} 16.349 & 16.741 & 16.597 \\ 16.641 & 17.033 & 16.889 \\ 16.683 & 17.075 & 16.931 \end{bmatrix}.$$

Thus, the Whittle index policy has a suboptimality gap of $\|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty = 0.528$ in the approximate model. Note that since we have access to the approximate model, the above value functions can be computed in practice allowing us to estimate the suboptimality gap in the approximate model. Now, we use the results of Theorem 1 to bound the suboptimality gap in the true model.

We first consider the case when $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}}$. In this case, $\rho_{\mathfrak{F}}(\cdot) = \frac{1}{2}\operatorname{span}(\cdot)$. Thus, the result (6) of Theorem 1 simplifies to

$$\begin{aligned} \|V^{\pi^*} - V^{\hat{\mu}}\|_\infty &\le \frac{4\varepsilon}{(1-\gamma)} + \frac{3\gamma\delta \operatorname{span}(\hat{V}^{\hat{\pi}^*})}{2(1-\gamma)} \\ &\quad + \frac{\gamma\delta \operatorname{span}(\hat{V}^{\hat{\mu}})}{2(1-\gamma)} + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty \\ &\le \frac{4 \times 0.016}{(1-0.9)} + \frac{3 \times 0.9 \times 0.04 \times 0.726}{2(1-0.9)} \\ &\quad + \frac{0.9 \times 0.04 \times 0.733}{2(1-0.9)} + 0.528 \\ &\le 1.163 + 0.528 = 1.691. \end{aligned}$$

Now consider the case when $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$. In this case, $\rho_{\mathfrak{F}}(\cdot) = \operatorname{Lip}(\cdot)$. Thus, the result (6) of Theorem 1 simplifies to

$$\begin{aligned} &\|V^{\pi^*} - V^{\hat{\mu}}\|_\infty \\ &\le \frac{4\varepsilon}{(1-\gamma)} + \frac{3\gamma\delta \operatorname{Lip}(\hat{V}^{\hat{\pi}^*})}{(1-\gamma)} \\ &\quad + \frac{\gamma\delta \operatorname{Lip}(\hat{V}^{\hat{\mu}})}{(1-\gamma)} + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty \\ &\le \frac{4 \times 0.016}{(1-0.9)} + \frac{3 \times 0.9 \times 0.06 \times 0.392}{(1-0.9)} \\ &\quad + \frac{0.9 \times 0.06 \times 0.461}{(1-0.9)} + 0.528 \\ &\le 1.524 + 0.528 = 2.052. \end{aligned}$$

Thus, in this example, we obtain a tighter bound by using $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}}$. The above calculations show how the result of Theorem 1 can be useful in bounding the suboptimality gap of the Whittle index policy when the true model is not known.

3.2. An example with continuous state spaces

Consider a model for machine maintenance with $n = 2$ machines and $m = 1$ repair man. $\mathsf{S} = [0, 1]$ denotes the state space of the machines where $s = 0$ denotes a machine in a pristine state and $s = 1$ denotes a completely deteriorated machine. We assume that when the passive action $a = 0$ is taken, the system incurs a per-step cost of s and deteriorates to a worse state in $[s, 1]$ uniformly at random. When the active action $a = 1$ is taken, the system incurs a per-step cost of λ and the state of the machine resets to a pristine state. Thus,

$$\begin{aligned} r^i(s^i, 0) &= -\xi^i s^i, & r^i(s^i, 1) &= -\lambda^i, \\ p^i(\cdot|s^i, 0) &= \mathcal{U}(s^i, 1), & p^i(\cdot|s^i, 1) &= \delta_D(\cdot), \end{aligned}$$

where $i \in \{1,2\}$, $s^i \in \mathsf{S}$, $\xi^1 = 1.0, \xi^2 = 0.5, \lambda^1 = 0.7, \lambda^2 = 0.3$. $\mathcal{U}(x,y)$ denotes a uniform distribution on the interval $[x,y]$, $\delta_D(\cdot)$ is the Dirac delta distribution and let $\gamma = 0.9$.

Suppose we want to compute the Whittle index by discretization. In particular, we consider a piecewise constant approximation of the model as follows. We divide the interval $[0,1]$ into H subintervals

$$[0, \tfrac{1}{H}) \cup [\tfrac{1}{H}, \tfrac{2}{H}) \cup \cdots \cup [1 - \tfrac{1}{H}, 1]$$

and consider the centers of each interval given by

$$\hat{\mathsf{S}} = \{\tfrac{1}{2H}, \tfrac{3}{2H}, \dots, \tfrac{2H-1}{2H}\}.$$

Consider a quantization function $\phi \colon \mathsf{S} \to \hat{\mathsf{S}}$, which maps any point to its closest point in $\hat{\mathsf{S}}$, i.e.,

$$\phi(s) = \begin{cases} \frac{1}{2H}, & \text{if } s \in [0, \frac{1}{H}) \\ \frac{3}{2H}, & \text{if } s \in [\frac{1}{H}, \frac{2}{H}) \\ \vdots & \vdots \\ \frac{2H-1}{2H}, & \text{if } s \in [1 - \frac{1}{H}, 1] \end{cases}$$

We then consider $H = 100$ and construct approximate arms $\hat{\alpha}^i = \langle \mathsf{S}, \{0,1\}, \{\hat{p}^i(a)\}_{a \in \{0,1\}}, \hat{r}^i \rangle$, where $i \in \{1,2\}$ and we have

$$\hat{p}^i(\cdot|s^i, 0) = \mathcal{U}(\phi(s^i) + \tfrac{1}{2H}, 1), \quad \hat{p}^i(\cdot|s^i, 1) = \delta_D(\cdot)$$

where $\mathcal{U}(x,y)$ denotes a uniform distribution on the interval $[x,y]$ and

$$\hat{r}^i(s^i, 0) = r^i(\phi(s^i), 0) = -\xi^i \phi(s^i), \quad \hat{r}^i(s^i, 1) = -\lambda^i.$$

Since the approximate model satisfies the restart property of [1, 3], it is indexable. Thus, Assumption 1 is satisfied.

Let $\hat{\omega}^i(s)$ denote the Whittle index (for the approximate model) of arm i in state s. We compute these using the modified adaptive greedy algorithm [3], and they can be visualized in Fig. 2. The Whittle index policy is given by (11).

Note that in this example, we do not solve for value functions in the true model because it has a continuous state space and so we cannot compute the suboptimality gap $\|V^{\pi^*} - V^{\hat{\mu}}\|_\infty$. So, the results of Theorem 1 provide a method to bound the suboptimality gap.

For that matter, we first compute the values of approximate errors (ε, δ) for arms 1 and 2 which are shown in Table 2 (for $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$, we use $d(s,s') = |s - s'|$ as the metric on S).

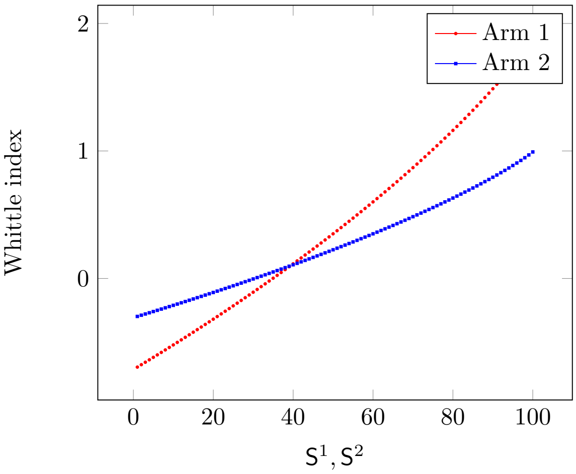


Figure 2: Whittle indices $\hat{\omega}$ plotted for all states for the example of Sec. 3.2.

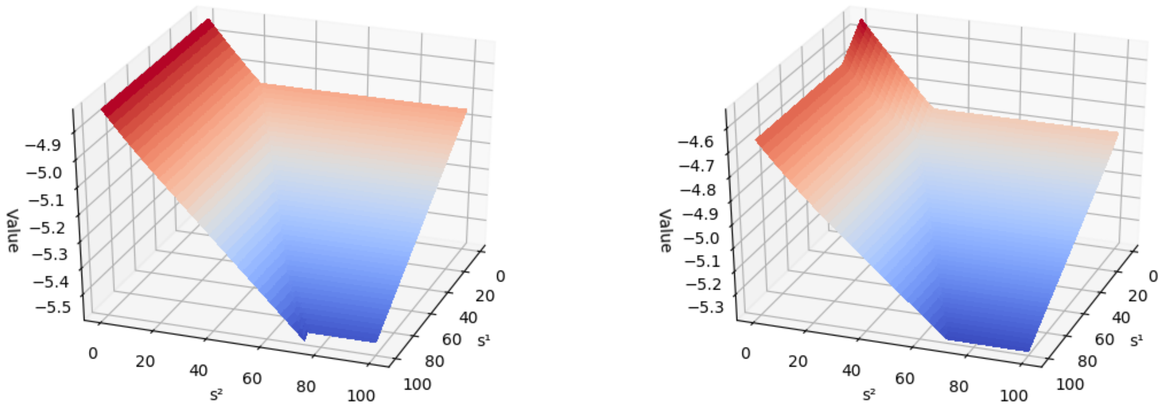
Parameter	**Arm 1**	**Arm 2**	**Overall**
ε	0.005	0.0025	0.0075
$\delta^{\mathfrak{F}^{\mathrm{W}}}$	0.005	0.005	0.01

Table 2: Parameters involved in Theorem 1 for Example 3.2.

We are interested in bounding the performance loss in using the Whittle index policy for the approximate model, in the approximate model. For that matter, we first compute the value function of the Whittle index policy $\hat{V}^{\hat{\mu}}$ using the policy evaluation equation [29]. The value function can be visualized by the 3D plot in Fig. 3a.

We can also compute the optimal value function of the approximate model $\hat{V}^{\hat{\pi}^*}$, which we do using the value iteration algorithm [29]. The value function can be visualized by the 3D plot in Fig. 3b.

Thus, the Whittle index policy has a suboptimality gap of $\|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty = 0.295$ in the approximate model. Note that since we have access to the approximate model, the above value functions can be computed in practice allowing us to estimate the suboptimality gap in the approximate model. Now, we use the results of Theorem 1 to bound the suboptimality gap in the true model.



(a) $\hat{V}^{\hat{\mu}}$

(b) $\hat{V}^{\hat{\pi}^*}$

Figure 3: Value functions $\hat{V}^{\hat{\mu}}$ and $\hat{V}^{\hat{\pi}^*}$ plotted for all states for the example of Sec. 3.2.

Thus, the result (6) of Theorem 1 simplifies to

$$\begin{aligned}\|V^{\pi^*} - V^{\hat{\mu}}\|_\infty &\le \frac{4\varepsilon}{(1-\gamma)} + \frac{3\gamma\delta\,\mathrm{Lip}(\hat{V}^{\hat{\pi}^*})}{(1-\gamma)} \\ &\quad + \frac{\gamma\delta\,\mathrm{Lip}(\hat{V}^{\hat{\mu}})}{(1-\gamma)} + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty \\ &\le \frac{4 \times 0.0075}{(1-0.9)} + \frac{3 \times 0.9 \times 0.01 \times 0.014}{(1-0.9)} \\ &\quad + \frac{0.9 \times 0.01 \times 0.040}{(1-0.9)} + 0.295 \\ &\le 0.307 + 0.295 \\ &= 0.602.\end{aligned}$$

The above calculations show how the result of Theorem 1 can be useful in bounding the suboptimality gap of the Whittle index policy when the true model is not known.

4. Proof of main result

4.1. Roadmap of the proof

The RMAB $\{\alpha^i\}_{i\in[n]}$ can be considered as an MDP $\mathcal{M} = \langle \mathbf{S}, \mathbf{A}(m), \boldsymbol{p}, \boldsymbol{r}\rangle$ where for any $\boldsymbol{s}_t, \boldsymbol{s}_{t+1} \in \mathbf{S}$ and $\boldsymbol{a}_t \in \mathbf{A}(m)$, we have

$$\boldsymbol{p}(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) = \prod_{i\in[n]} p^i(s^i_{t+1} \mid s^i_t, a^i_t), \tag{12}$$

$$\boldsymbol{r}(\boldsymbol{s}_t, \boldsymbol{a}_t) = \sum_{i\in[n]} r^i(s^i_t, a^i_t). \tag{13}$$

The approximate RMAB $\{\hat{\alpha}^i\}_{i\in[n]}$ can also be considered as an MDP $\hat{\mathcal{M}} = \langle \mathbf{S}, \mathbf{A}(m), \hat{\boldsymbol{p}}, \hat{\boldsymbol{r}}\rangle$ where $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{r}}$ are defined in a manner similar to $\boldsymbol{p}$ and $\boldsymbol{r}$.

The main intuition of our proof is that if $\hat{\alpha}^i$ is an $(\varepsilon^i, \delta^i)$-approximation of arm α^i for each $i \in [n]$, then $\hat{\mathcal{M}}$ is an (ε, δ)-approximation of $\mathcal{M}$ in some appropriate sense to be described later, where (ε, δ) can be characterized in terms of $\{(\varepsilon^i, \delta^i)\}_{i\in n}$. Then, we can use approximation results from MDPs to derive approximation bounds for RMABs. In the rest of this section, we formalize this intuition.

4.2. Preliminary results

Definition 4. *Consider the two* MDPs $\mathcal{M} = \langle \mathbf{S}, \mathbf{A}(m), \boldsymbol{p}, \boldsymbol{r}\rangle$ *and* $\hat{\mathcal{M}} = \langle \mathbf{S}, \mathbf{A}(m), \hat{\boldsymbol{p}}, \hat{\boldsymbol{r}}\rangle$ *which are defined on the same state and action spaces. Given a function space* $\mathfrak{F}$ *and positive constants* ε *and* δ, *the* MDP $\hat{\mathcal{M}}$ *is called an* (ε, δ)-*approximation of the* MDP $\mathcal{M}$ *if for all* $\boldsymbol{s} \in \mathbf{S}$ *and* $\boldsymbol{a} \in \mathbf{A}(m)$*:*

$$\big|\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a}) - \hat{\boldsymbol{r}}(\boldsymbol{s}, \boldsymbol{a})\big| \le \varepsilon, \quad d_{\mathfrak{F}}\big(\boldsymbol{p}(\cdot|\boldsymbol{s}, \boldsymbol{a}), \hat{\boldsymbol{p}}(\cdot \mid \boldsymbol{s}, \boldsymbol{a})\big) \le \delta.$$

Now we formalize the approximation bound between $\mathcal{M}$ and $\hat{\mathcal{M}}$.

Lemma 1. *When* $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}}$ *or* $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$, *then the* MDP $\hat{\mathcal{M}}$ *is an* (ε, δ)-*approximation of the* MDP $\mathcal{M}$, *where*

$$(\varepsilon, \delta) = \Big(\sum_{i\in[n]} \varepsilon^i, \sum_{i\in[n]} \delta^i\Big). \tag{14}$$

Proof. See Appendix B. □

From standard results of Markov decision theory [30], we know that for a given policy π, the performance V^π defined by (1) satisfies the following fixed point equation:

$$V^\pi(\boldsymbol{s}) = Q^\pi(\boldsymbol{s}, \pi(\boldsymbol{s})), \tag{15a}$$

$$Q^\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}[\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a})] + \gamma \int_{\mathsf{S}} V^\pi(\boldsymbol{s}')\boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}). \tag{15b}$$

Similarly, for any policy π let $\hat{V}^\pi$ denote the performance of policy π in the approximate model $\hat{\mathcal{M}}$. Then, $\hat{V}^\pi$ satisfies the following fixed point equation:

$$\hat{V}^\pi(\boldsymbol{s}) = \hat{Q}^\pi(\boldsymbol{s}, \pi(\boldsymbol{s})), \tag{16a}$$

$$\hat{Q}^\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}[\hat{\boldsymbol{r}}(\boldsymbol{s}, \boldsymbol{a})] + \gamma \int_{\mathsf{S}} \hat{V}^\pi(\boldsymbol{s}')\hat{\boldsymbol{p}}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}). \tag{16b}$$

An immediate consequence of Lemma 1 is the following.

Proposition 3. *For the approximate setup described in Sec. 2.2 and for any policy π*

$$\|V^\pi - \hat{V}^\pi\|_\infty \le \|Q^\pi - \hat{Q}^\pi\|_\infty \le \beta^\pi_{\mathfrak{F}}. \tag{17}$$

Furthermore, for any policies π^* and $\hat{\pi}^*$ which are optimal for $\mathcal{M}$ and $\hat{\mathcal{M}}$, we have

$$\|V^{\pi^*} - \hat{V}^{\hat{\pi}^*}\|_\infty \le \|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty \le \beta^{\hat{\pi}^*}_{\mathfrak{F}}. \tag{18}$$

Therefore, by the triangle inequality

$$\|Q^{\pi^*} - Q^{\hat{\pi}^*}\|_\infty \le 2\beta^{\hat{\pi}^*}_{\mathfrak{F}} \text{ and } \|V^{\pi^*} - V^{\hat{\pi}^*}\|_\infty \le 2\beta^{\hat{\pi}^*}_{\mathfrak{F}}. \tag{19}$$

Proof. For the proof of the first part of (17), observe that from (15) and (16) we have that for any $\boldsymbol{s} \in \mathsf{S}$,

$$\begin{aligned} &|V^\pi(\boldsymbol{s}) - \hat{V}^\pi(\boldsymbol{s})| \\ &= \left|Q^\pi(\boldsymbol{s}, \pi(\boldsymbol{s})) - \hat{Q}^\pi(\boldsymbol{s}, \pi(\boldsymbol{s}))\right| \\ &\overset{(a)}{\le} \|Q^\pi(\boldsymbol{s}, \cdot) - \hat{Q}^\pi(\boldsymbol{s}, \cdot)\|_\infty \\ &\overset{(b)}{\le} \|Q^\pi - \hat{Q}^\pi\|_\infty, \end{aligned}$$

where (a) and (b) follow from the definition of the sup norm. Supremizing the LHS over $\boldsymbol{s} \in \mathsf{S}$, we get

$$\|V^\pi - \hat{V}^\pi\|_\infty \le \|Q^\pi - \hat{Q}^\pi\|_\infty. \tag{20}$$

This proves the first part of (17). Now, we bound $\|Q^\pi - \hat{Q}^\pi\|_\infty$ as follows: for any fixed $\boldsymbol{s} \in \mathsf{S}$, $\boldsymbol{a} \in \mathsf{A}(m)$, from (15) and (16), we have

$$\begin{aligned} &|Q^\pi(\boldsymbol{s}, \boldsymbol{a}) - \hat{Q}^\pi(\boldsymbol{s}, \boldsymbol{a})| \\ &\overset{(c)}{\le} |\mathbb{E}[\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a})] - \mathbb{E}[\hat{\boldsymbol{r}}(\boldsymbol{s}, \boldsymbol{a})]| \\ &\quad + \gamma \int_{\mathsf{S}} |V^\pi(\boldsymbol{s}') - \hat{V}^\pi(\boldsymbol{s}')|\boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) \\ &\quad + \gamma \left| \int_{\mathsf{S}} \hat{V}^\pi(\boldsymbol{s}') \Big[\boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) - \hat{\boldsymbol{p}}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a})\Big] \right| \\ &\overset{(d)}{\le} \varepsilon + \gamma\|Q^\pi - \hat{Q}^\pi\|_\infty + \gamma\rho_{\mathfrak{F}}(\hat{V}^\pi)\delta, \end{aligned} \tag{21}$$

where (c) follows from the definition of Q^π and $\hat{Q}^\pi$, adding and subtracting the $\hat{V}^\pi$ term and the triangle inequality; (d) follows from (20) and the definition of an (ε, δ)-approximation for an MDP. Supremizing the LHS of (21) over all $\boldsymbol{s}, \boldsymbol{a} \in \mathsf{S} \times \mathsf{A}(m)$ and re-arranging terms, we get

$$\|Q^\pi - \hat{Q}^\pi\|_\infty \le \frac{\varepsilon + \gamma\rho_{\mathfrak{F}}(\hat{V}^\pi)\delta}{(1-\gamma)} = \beta^\pi_{\mathfrak{F}}. \tag{22}$$

This proves the second part of (17).

For the proof of the first part of (18), observe that from (15) and (16) we have that for any $\boldsymbol{s} \in \mathsf{S}$,

$$\begin{aligned} &|V^{\pi^*}(\boldsymbol{s}) - \hat{V}^{\hat{\pi}^*}(\boldsymbol{s})| \\ &= \left| \max_{\boldsymbol{a} \in \mathsf{A}(m)} Q^{\pi^*}(\boldsymbol{s}, \boldsymbol{a}) - \max_{\boldsymbol{a} \in \mathsf{A}(m)} \hat{Q}^{\hat{\pi}^*}(\boldsymbol{s}, \boldsymbol{a}) \right| \\ &\overset{(e)}{\le} \max_{\boldsymbol{a} \in \mathsf{A}(m)} \left| Q^{\pi^*}(\boldsymbol{s}, \boldsymbol{a}) - \hat{Q}^{\hat{\pi}^*}(\boldsymbol{s}, \boldsymbol{a}) \right| \\ &\le \|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty, \end{aligned}$$

where (e) follows from the inequality $\max f(x) - \max g(x) \le \max|f(x) - g(x)|$. Supremizing the LHS over $\boldsymbol{s} \in \mathsf{S}$, we get

$$\|V^{\pi^*} - \hat{V}^{\hat{\pi}^*}\|_\infty \le \|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty. \tag{23}$$

This proves the first part of (17). Now, we bound $\|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty$ as follows: for any fixed $\boldsymbol{s} \in \mathsf{S}$, $\boldsymbol{a} \in \mathsf{A}(m)$, from (15) and (16), we have

$$\begin{aligned} &|Q^{\pi^*}(\boldsymbol{s}, \boldsymbol{a}) - \hat{Q}^{\hat{\pi}^*}(\boldsymbol{s}, \boldsymbol{a})| \\ &\overset{(f)}{\le} |\mathbb{E}[\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a})] - \mathbb{E}[\hat{\boldsymbol{r}}(\boldsymbol{s}, \boldsymbol{a})]| \\ &\quad + \gamma \int_{\mathsf{S}} |V^{\pi^*}(\boldsymbol{s}') - \hat{V}^{\hat{\pi}^*}(\boldsymbol{s}')|\boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) \\ &\quad + \gamma \left| \int_{\mathsf{S}} \hat{V}^{\hat{\pi}^*}(\boldsymbol{s}') \Big[\boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) - \hat{\boldsymbol{p}}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a})\Big] \right| \\ &\overset{(g)}{\le} \varepsilon + \gamma\|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty + \gamma\rho_{\mathfrak{F}}(\hat{V}^{\hat{\pi}^*})\delta, \end{aligned} \tag{24}$$

where (f) is the same as (c); (g) follows from (23) and the definition of an (ε,δ)-approximation for an MDP. Supremizing the LHS of (24) over all $\boldsymbol{s},\boldsymbol{a} \in \mathsf{S} \times \mathsf{A}(m)$ and re-arranging terms, we get

$$\|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty \leq \frac{\varepsilon + \gamma \rho_{\mathfrak{F}}(\hat{V}^{\hat{\pi}^*})\delta}{(1-\gamma)} = \beta_{\mathfrak{F}}^{\hat{\pi}^*}. \tag{25}$$

This proves the second part of (17).

Finally, to show the first part of (19), consider

$$\begin{aligned}\|Q^{\pi^*} - Q^{\hat{\pi}^*}\|_\infty &\overset{(h)}{\leq} \|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty + \|Q^{\hat{\pi}^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty \\ &\overset{(i)}{\leq} \beta_{\mathfrak{F}}^{\hat{\pi}^*} + \beta_{\mathfrak{F}}^{\hat{\pi}^*} = 2\beta_{\mathfrak{F}}^{\hat{\pi}^*},\end{aligned}$$

where (h) follows from the triangle inequality; (i) follows from (22) with $\pi = \hat{\pi}^*$ and (25). To show the second part of (19), consider

$$\begin{aligned}\|V^{\pi^*} - V^{\hat{\pi}^*}\|_\infty &\overset{(j)}{\leq} \|V^{\pi^*} - \hat{V}^{\hat{\pi}^*}\|_\infty + \|V^{\hat{\pi}^*} - \hat{V}^{\hat{\pi}^*}\|_\infty \\ &\overset{(k)}{\leq} \|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty + \|Q^{\hat{\pi}^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty \\ &\overset{(l)}{\leq} \beta_{\mathfrak{F}}^{\hat{\pi}^*} + \beta_{\mathfrak{F}}^{\hat{\pi}^*} = 2\beta_{\mathfrak{F}}^{\hat{\pi}^*},\end{aligned}$$

where (j) follows from the triangle inequality; (k) follows from (20) with $\pi = \hat{\pi}^*$ and (23); (l) follows from (22) with $\pi = \hat{\pi}^*$ and (25). □

4.3. Proof of Theorem 1

For the first part of the theorem, from the triangle inequality, we have

$$\begin{aligned}\|Q^{\pi^*} - Q^{\hat{\mu}}\|_\infty &\leq \|Q^{\pi^*} - Q^{\hat{\pi}^*}\|_\infty + \|Q^{\hat{\pi}^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty \\ &\quad + \|\hat{Q}^{\hat{\pi}^*} - \hat{Q}^{\hat{\mu}}\|_\infty + \|\hat{Q}^{\hat{\mu}} - Q^{\hat{\mu}}\|_\infty \\ &\overset{(a)}{\leq} 2\beta_{\mathfrak{F}}^{\hat{\pi}^*} + \beta_{\mathfrak{F}}^{\hat{\pi}^*} \|\hat{Q}^{\hat{\pi}^*} - \hat{Q}^{\hat{\mu}}\|_\infty, + \beta_{\mathfrak{F}}^{\hat{\mu}},\end{aligned} \tag{26}$$

where each term of (a) is bound using Prop. 3. Rearranging terms proves (5).

For the second part of the theorem, from triangle inequality we have

$$\begin{aligned}\|V^{\pi^*} - V^{\hat{\mu}}\|_\infty &\leq \|V^{\pi^*} - V^{\hat{\pi}^*}\|_\infty + \|V^{\hat{\pi}^*} - \hat{V}^{\hat{\pi}^*}\|_\infty \\ &\quad + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty + \|\hat{V}^{\hat{\mu}} - V^{\hat{\mu}}\|_\infty \\ &\overset{(a)}{\leq} 2\beta_{\mathfrak{F}}^{\hat{\pi}^*} + \beta_{\mathfrak{F}}^{\hat{\pi}^*} + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty + \beta_{\mathfrak{F}}^{\hat{\mu}},\end{aligned} \tag{27}$$

where each term of (b) is bound using Prop. 3. Rearranging the terms proves (6).

4.4. Proof of Proposition 1

First, observe that for $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}}$,

$$\begin{aligned}\rho_{\mathfrak{F}}(\hat{V}^{\hat{\pi}^*}) &= \frac{1}{2}\operatorname{span}(\hat{V}^{\hat{\pi}^*}) \\ &\overset{(a)}{\leq} \frac{1}{2}\frac{\operatorname{span}(\hat{\boldsymbol{r}})}{(1-\gamma)} \\ &\overset{(b)}{\leq} \frac{1}{2}\frac{\sum_{i\in[n]}\operatorname{span}(\hat{r}^i)}{(1-\gamma)}.\end{aligned}$$

where (a) follows from [33, Lemma 39] and (b) follows because span is a semi-norm [30]. Using the above bound in (5) and (6) and using Lemma 1 to bound (ε,δ), we get (7) and (8).

4.5. Proof of Proposition 2

Suppose the state space S^i of each arm is a metric space with metric d^i. Define a metric $\boldsymbol{d}$ on S as follows: for any $q \in [1,\infty]$ and $\boldsymbol{s},\boldsymbol{s}' \in \mathsf{S}$, $\boldsymbol{d}(\boldsymbol{s},\boldsymbol{s}') = \big(\sum_{i\in[n]} d^i(s^i,s'^i)^q\big)^{1/q}$. We now define Lipshitz continuity for MDP $\mathcal{M}$.

Definition 5. *Given* MDP $\mathcal{M} = \langle \mathsf{S}, \mathsf{A}(m), \boldsymbol{p}, \boldsymbol{r}\rangle$, *if*

$$\mathrm{L}_{\boldsymbol{r}} := \sup_{\substack{\boldsymbol{s},\boldsymbol{s}'\in\mathsf{S}\\ \boldsymbol{a}\in\mathsf{A}(m)}} \frac{|\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a}) - \boldsymbol{r}(\boldsymbol{s}',\boldsymbol{a})|}{d(\boldsymbol{s},\boldsymbol{s}')} < \infty,$$

$$\mathrm{L}_{\boldsymbol{p}} := \sup_{\substack{\boldsymbol{s},\boldsymbol{s}'\in\mathsf{S}\\ \boldsymbol{a}\in\mathsf{A}(m)}} \frac{d_{\mathfrak{F}^{\mathrm{W}}}(\boldsymbol{p}(\cdot \mid \boldsymbol{s},\boldsymbol{a}), \boldsymbol{p}(\cdot \mid \boldsymbol{s}',\boldsymbol{a}))}{d(\boldsymbol{s},\boldsymbol{s}')} < \infty,$$

then the MDP $\mathcal{M}$ is said to be $(\mathrm{L}_{\boldsymbol{r}},\mathrm{L}_{\boldsymbol{p}})$*-Lipschitz.*

Lemma 2. *If arms $\hat{\alpha}^i$ are $(\mathrm{L}_{\hat{r}^i},\mathrm{L}_{\hat{p}^i})$-Lipschitz, for all $i \in [n]$, and $k \in [1,\infty]$, such that $1/k + 1/q = 1$, then the* MDP $\hat{\mathcal{M}} = \langle \mathsf{S}, \mathsf{A}(m), \hat{\boldsymbol{p}}, \hat{\boldsymbol{r}}\rangle$ *is* $(\mathrm{L}^{(k)}_{\hat{\boldsymbol{r}}},\mathrm{L}^{(k)}_{\hat{\boldsymbol{p}}})$*-Lipschitz, where*

$$\mathrm{L}^{(k)}_{\hat{\boldsymbol{r}}} \leq \Bigg(\sum_{i\in[n]}(\mathrm{L}_{\hat{r}^i})^k\Bigg)^{1/k}, \quad \mathrm{L}^{(k)}_{\hat{\boldsymbol{p}}} \leq \Bigg(\sum_{i\in[n]}(\mathrm{L}_{\hat{p}^i})^k\Bigg)^{1/k}. \tag{28}$$

Proof. See Appendix C. □

Now, observe that for $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$,

$$\rho_{\mathfrak{F}}(\hat{V}^{\hat{\pi}^*}) = \operatorname{Lip}(\hat{V}^{\hat{\pi}^*}) \overset{(a)}{\leq} \frac{\mathrm{L}^{(k)}_{\hat{\boldsymbol{r}}}}{(1-\gamma\mathrm{L}^{(k)}_{\hat{\boldsymbol{p}}})}. \tag{29}$$

where (a) follows from [18, Theorem 4.2]. To prove Proposition 2, we will take $k = \infty$ because doing so gives the tighest possible bound in (29). Substititing (29) in (5) and (6) and using Lemma 1 to bound (ε,δ), we get (9) and (10).

5. Conclusions

We considered a restless multi-armed bandit problem where the model of each arm is known approximately and provided a bound on the loss of performance in using the Whittle index policy of the approximate model. The bound depends on the approximation errors in modeling each arm, properties of the transition dynamics and the per-step reward of each arm, and the degree of suboptimality of the Whittle index policy in the approximate model.

The degree of approximation of an arm depends on the choice of metric on probability spaces. We quantify our bounds for two specific choices of metrics: total variation distance and Wasserstein distance. The results are easy to generalize to other types of integral probability metrics (IPMs) as well.

6. Acknowledgements

This work was supported in part by the Innovation for Defence Excellence and Security (IDEaS) Program of the Canadian Department of National Defence through grant CFPMN2-037.

Appendix A. Preliminary Results

We first prove some preliminary results.

Lemma 3. *Consider any $f\colon \boldsymbol{\mathsf{S}} \to \mathbb{R}$. Pick an arm $i \in [n]$ and arbitrarily fix $\boldsymbol{s}^{-i} \in \boldsymbol{\mathsf{S}}^{-i}$. Define $f^i\colon \mathsf{S}^i \to \mathbb{R}$ by $f^i(s^i) = f(s^i, \boldsymbol{s}^{-i})$, for any $s^i \in \mathsf{S}^i$. Then*

(a) $\operatorname{span}(f^i) \le \operatorname{span}(f)$.

(b) $\operatorname{Lip}(f^i) \le \operatorname{Lip}(f)$.

Proof. (a) Consider for any $\boldsymbol{s}^{-i} \in \boldsymbol{\mathsf{S}}^{-i}$

$$\begin{aligned}
\operatorname{span}(f^i) &= \sup_{s^i_{(1)}, s^i_{(2)} \in \mathsf{S}^i} \left| f^i(s^i_{(1)}) - f^i(s^i_{(2)}) \right| \\
&\overset{(a)}{=} \sup_{s^i_{(1)}, s^i_{(2)} \in \mathsf{S}^i} \left| f(s^i_{(1)}, \boldsymbol{s}^{-i}) - f(s^i_{(2)}, \boldsymbol{s}^{-i}) \right| \\
&\overset{(b)}{\le} \sup_{\boldsymbol{s}_{(1)}, \boldsymbol{s}_{(2)} \in \boldsymbol{\mathsf{S}}} \left| f(\boldsymbol{s}_{(1)}) - f(\boldsymbol{s}_{(2)}) \right| \\
&= \operatorname{span}(f),
\end{aligned}$$

where (a) follows from the definition of f^i given $\boldsymbol{s}^{-i}$ and (b) follows from the fact that taking supremum over all $\boldsymbol{\mathsf{S}}^{-i}$ will given an upper bound to any specific $\boldsymbol{s}^{-i}$.

(b) Again for any $\boldsymbol{s}^{-i} \in \boldsymbol{\mathsf{S}}^{-i}$

$$\begin{aligned}
\operatorname{Lip}(f^i) &= \sup_{s^i, \tilde{s}^i \in \mathsf{S}^i} \frac{|f^i(s^i) - f^i(\tilde{s}^i)|}{d^i(s^i, \tilde{s}^i)} \\
&\overset{(c)}{=} \sup_{s^i, \tilde{s}^i \in \mathsf{S}^i} \frac{|f(s^i, \boldsymbol{s}^{-i}) - f(\tilde{s}^i, \boldsymbol{s}^{-i})|}{\boldsymbol{d}((s^i, \boldsymbol{s}^{-i}), (\tilde{s}^i, \boldsymbol{s}^{-i}))} \\
&\overset{(d)}{\le} \sup_{\boldsymbol{s}, \tilde{\boldsymbol{s}} \in \boldsymbol{\mathsf{S}}} \frac{|f(\boldsymbol{s}) - f(\tilde{\boldsymbol{s}})|}{\boldsymbol{d}(\boldsymbol{s}, \tilde{\boldsymbol{s}})} \\
&= \operatorname{Lip}(f),
\end{aligned}$$

where (c) follows from the definition of metric $\boldsymbol{d}$ and function f^i given $\boldsymbol{s}^{-i}$ and (d) follows from the fact that taking supremum over all $\boldsymbol{\mathsf{S}}^{-i}$ will given an upper bound to any specific $\boldsymbol{s}^{-i}$.

□

For the ease of notation, when $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}} = \{f\colon \boldsymbol{\mathsf{S}} \to \mathbb{R}\colon \frac{1}{2}\operatorname{span}(f) \le 1\}$, define $\mathfrak{F}^i = \{f^i\colon \mathsf{S}^i \to \mathbb{R}\colon \frac{1}{2}\operatorname{span}(f^i) \le 1\}$. Similarly when $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}} = \{f\colon \boldsymbol{\mathsf{S}} \to \mathbb{R}\colon \operatorname{Lip}(f) \le 1\}$, define $\mathfrak{F}^i = \{f^i\colon \mathsf{S}^i \to \mathbb{R}\colon \operatorname{Lip}(f^i) \le 1\}$. Lemma 3 implies that if $f \in \mathfrak{F}$, for any $\boldsymbol{s}^{-i} \in \boldsymbol{\mathsf{S}}^{-i}$, f^i (as defined in Lemma 3) belongs to $\mathfrak{F}^i$.

Lemma 4. *Let μ^i, ν^i be probability densities on S^i. Define $\boldsymbol{\mu} = \mu^1 \otimes \cdots \otimes \mu^n$ and $\boldsymbol{\nu} = \nu^1 \otimes \cdots \otimes \nu^n$. Then for $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}}$ or $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$,*

$$d_{\mathfrak{F}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \le \sum_{i \in [n]} d_{\mathfrak{F}^i}(\mu^i, \nu^i).$$

Proof. We prove the result by induction on n. The result is trivially true for $n = 1$. This forms the basis of induction. Now assume that the result is true for $n = k - 1$ and consider the case for $n = k$.

For any $f \in \mathfrak{F}$, S^{-k} and s^{-k} being the state space and the state by excluding the k^{th} component, we have

$$\begin{aligned}
&\left| \int_{\boldsymbol{\mathsf{S}}} f d\boldsymbol{\mu} - \int_{\boldsymbol{\mathsf{S}}} f d\boldsymbol{\nu} \right| \\
&= \left| \int_{\mathsf{S}^k} \int_{\mathsf{S}^{-k}} f(s^k, s^{-k}) \big[\mu^k(s^k) \mu^{-k}(s^{-k}) \right. \\
&\qquad \left. - \nu^k(s^k) \nu^{-k}(s^{-k}) \big] ds^k ds^{-k} \right|
\end{aligned}$$

$$\begin{aligned}
&\overset{(a)}{\leq} \left| \int_{\mathsf{S}^k}\int_{\mathsf{S}^{-k}} f(s^k,s^{-k})\big[\mu^k(s^k)\mu^{-k}(s^{-k}) \right.\\
&\qquad \left. - \mu^k(s^k)\nu^{-k}(s^{-k})\big] ds^k ds^{-k}\right|\\
&\quad + \left|\int_{\mathsf{S}^k}\int_{\mathsf{S}^{-k}} f(s^k,s^{-k})\big[\mu^k(s^k)\nu^{-k}(s^{-k}) \right.\\
&\qquad \left. - \nu^k(s^k)\nu^{-k}(s^{-k})\big]ds^k ds^{-k}\right|\\
&\overset{(b)}{\leq} \int_{\mathsf{S}^k}\left|\int_{\mathsf{S}^{-k}} f(s^k,s^{-k})\big[\mu^{-k}(s^{-k})-\nu^{-k}(s^{-k})\big] \right.\\
&\qquad\qquad \left. ds^{-k}\right|\mu^k(s^k)ds^k\\
&\quad + \int_{\mathsf{S}^{-k}}\left|\int_{\mathsf{S}^k} f(s^k,s^{-k})\big[\mu^k(s^k)-\nu^k(s^k)\big] \right.\\
&\qquad\qquad \left. ds^k\right|\nu^{-k}(s^{-k})ds^{-k}
\end{aligned} \tag{A.1}$$

where (a) follows from adding and subtracting the same term and using the triangle inequality and (b) also follows from the triangle inequality. Now observe that for a fixed s^k, by Lemma 3, $f(s^k,\cdot) \in \mathfrak{F}^{-k}$. Therefore,

$$\begin{aligned}
&\left|\int_{\mathsf{S}^{-k}} f(s^k,s^{-k})\big[\mu^{-k}(s^{-k})-\nu^{-k}(s^{-k})\big]ds^{-k}\right|\\
&\leq d_{\mathfrak{F}^{-k}}(\mu^{-k},\nu^{-k})
\end{aligned} \tag{A.2}$$

and similarly,

$$\begin{aligned}
&\left|\int_{\mathsf{S}^k} f(s^k,s^{-k})\big[\mu^k(s^k)-\nu^k(s^k)\big]ds^k\right|\\
&\leq d_{\mathfrak{F}^k}(\mu^k,\nu^k)
\end{aligned} \tag{A.3}$$

Substituting (A.2) and (A.3) in (A.1), we get

$$\begin{aligned}
&\left|\int_{\mathsf{S}} f d\boldsymbol{\mu} - \int_{\mathsf{S}} f d\boldsymbol{\nu}\right|\\
&\leq \int_{\mathsf{S}^k} d_{\mathfrak{F}^{-k}}(\mu^{-k},\nu^{-k})\mu^k(s^k)ds^k\\
&\quad + \int_{\mathsf{S}^{-k}} d_{\mathfrak{F}^k}(\mu^k,\nu^k)\mu^{-k}(s^{-k})ds^{-k}\\
&= d_{\mathfrak{F}^k}(\mu^k,\nu^k) + d_{\mathfrak{F}^{-k}}(\mu^{-k},\nu^{-k})\\
&\overset{(c)}{\leq} \sum_{i\in[k]} d_{\mathfrak{F}^i}(\mu^i,\nu^i),
\end{aligned}$$

where (c) follows from the induction hypothesis which is true for $k-1$. The final result follows from induction. □

Appendix B. Proof of Lemma 1

For the first part, consider

$$\begin{aligned}
|\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a}) - \hat{\boldsymbol{r}}(\boldsymbol{s},\boldsymbol{a})| &= \left|\sum_{i\in[n]} r^i(s^i,a^i) - \sum_{i\in[n]}\hat{r}^i(s^i,a^i)\right|\\
&\overset{(a)}{\leq} \sum_{i\in[n]}\left|r^i(s^i,a^i) - \hat{r}^i(s^i,a^i)\right| \overset{(b)}{\leq} \sum_{i\in[n]}\varepsilon^i.
\end{aligned}$$

where (a) follows from the triangle inequality and (b) follows from the assumption on the arms. This proves the first part of the Lemma.

The second part follows from the definition of $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$ (Eq. (12)) and Lemma 4.

Appendix C. Proof of Lemma 2

For the first part, consider

$$\begin{aligned}
&\left|\hat{\boldsymbol{r}}(\boldsymbol{s}_{(1)},\boldsymbol{a}) - \hat{\boldsymbol{r}}(\boldsymbol{s}_{(2)},\boldsymbol{a})\right|\\
&= \left|\sum_{i\in[n]}\hat{r}^i({s_{(1)}}^i,a^i) - \sum_{i\in[n]}\hat{r}^i({s_{(2)}}^i,a^i)\right|\\
&\overset{(a)}{\leq} \sum_{i\in[n]}\left|\hat{r}^i({s_{(1)}}^i,a^i) - \hat{r}^i({s_{(2)}}^i,a^i)\right|\\
&\overset{(b)}{\leq} \sum_{i\in[n]} \mathrm{L}_{\hat{r}^i} d^i({s_{(1)}}^i,{s_{(2)}}^i)\\
&\overset{(c)}{\leq} \left(\sum_{i\in[n]}(\mathrm{L}_{\hat{r}^i})^k\right)^{1/k} \boldsymbol{d}(\boldsymbol{s}_{(1)},\boldsymbol{s}_{(2)}).
\end{aligned}$$

where (a) follows from the triangle inequality, (b) follows from the assumption on the arms and (c) follows from Hölder's inequality and the definition of metric $\boldsymbol{d}$.

For the second part, consider

$$\begin{aligned}
&d_{\mathfrak{F}}(\hat{\boldsymbol{p}}(\cdot|\boldsymbol{s}_{(1)},\boldsymbol{a}),\hat{\boldsymbol{p}}(\cdot|\boldsymbol{s}_{(2)},\boldsymbol{a}))\\
&\overset{(d)}{\leq} \sum_{i\in[n]} d_{\mathfrak{F}^i}(\hat{p}^i(\cdot|{s_{(1)}}^i,a^i),\hat{p}^i(\cdot|{s_{(2)}}^i,a^i))\\
&\overset{(e)}{\leq} \sum_{i\in[n]} \mathrm{L}_{\hat{p}^i} d^i({s_{(1)}}^i,{s_{(2)}}^i)\\
&\overset{(f)}{\leq} \left(\sum_{i\in[n]}(\mathrm{L}_{\hat{p}^i})^k\right)^{1/k} \boldsymbol{d}(\boldsymbol{s}_{(1)},\boldsymbol{s}_{(2)}).
\end{aligned}$$

where (d) follows from Lemma 4, (e) follows from the assumption on the arms and (f) follows from Hölder's inequality and the definition of metric $\boldsymbol{d}$.

References

[1] N. Akbarzadeh and A. Mahajan. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *Conference on Decision and Control*, pages 7294–7300, 2019.

[2] Nima Akbarzadeh and Aditya Mahajan. On learning Whittle index policy for restless bandits with scalable regret. *arXiv preprint arXiv:2202.03463*, 2022.

[3] Nima Akbarzadeh and Aditya Mahajan. Conditions for indexability of restless bandits and an $\mathcal{O}(k^3)$ algorithm to compute Whittle index. *Journal of applied probability*, Dec 2022.

[4] P. S. Ansell, Kevin D. Glazebrook, José Niño-Mora, and M. O'Keeffe. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research*, 57(1):21–39, 2003.

[5] Konstantin E Avrachenkov and Vivek S Borkar. Whittle index based Q-learning for restless bandits with average reward. *Automatica*, 139:110186, 2022.

[6] Urtzi Ayesta, Martin Erausquin, and Peter Jacko. A modeling framework for optimizing the flow-level scheduling with time-varying channels. *Performance Evaluation*, 67(11):1014–1029, 2010.

[7] Adi Ben-Israel and Sjur Didrik Flåm. A bisection/successive approximation method for computing Gittins indices. *Mathematics for Operations Research*, 34(6):411–422, 1990.

[8] Vivek S. Borkar and Karan Chadha. A reinforcement learning algorithm for restless bandits. In *2018 Indian Control Conference (ICC)*, pages 89–94, 2018.

[9] Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

[10] Felipe Caro and Aparupa Das Gupta. Robust control of the multi-armed bandit problem. *Annals of Operations Research*, pages 1–20, 2015.

[11] Jhelum Chakravorty and Aditya Mahajan. Multi-armed bandits, Gittins index, and its calculation. *Methods and applications of statistics in clinical trials: Planning, analysis, and inferential methods*, 2(416-435):455, 2014.

[12] Xin Chen, Yutong Nie, and Na Li. Online residential demand response via contextual multi-armed bandits. *IEEE Control Systems Letters*, 5(2):433–438, 2020.

[13] Samuel N Cohen and Tanut Treetanthiploet. Gittins' theorem under uncertainty. *Electronic Journal of Probability*, 27:1–48, 2022.

[14] Jing Fu, Yoni Nazarathy, Sarat Moka, and Peter G. Taylor. Towards Q-learning the Whittle index for restless bandits. In *2019 Australian New Zealand Control Conference (ANZCC)*, pages 249–254, 2019.

[15] John C Gittins and David M Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, 66(3):561–565, 1979.

[16] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride. Some indexable families of restless bandit problems. *Advances in Applied Probability*, 38(3):643–672, 2006.

[17] 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. *European Journal of Operational Research*, 165(1):267–284, 2005.

[18] Karl Hinderer. Lipschitz continuity of value functions in markovian decision processes. *Mathematical Methods of Operations Research*, 62(1):3–22, 2005.

[19] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[20] Michael N Katehakis and Arthur F Veinott Jr. The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268, 1987.

[21] Michael Jong Kim. Robust control of partially observable failing systems. *Operations Research*, 64(4):999–1014, 2016.

[22] Michael Jong Kim and Andrew EB Lim. Robust multiarmed bandit problems. *Management Science*, 62(1):264–285, 2016.

[23] Christopher Lott and Demosthenis 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*, 14(3):259–297, 2000.

[24] Aditya Mahajan and Demosthenis Teneketzis. Multi-armed bandit problems. In *Foundations and applications of sensor management*, pages 121–151. Springer, 2008.

[25] Rahul Meshram, Aditya Gopalan, and D Manjunath. Restless bandits that hide their hand and recommendation systems. In *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*, pages 206–213. IEEE, 2017.

[26] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[27] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

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

[29] Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

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

[31] Francisco Robledo, Vivek Borkar, Urtzi Ayesta, and Konstantin Avrachenkov. QWI: Q-learning with Whittle index. *ACM SIGMETRICS Performance Evaluation Review*, 49(2):47–50, 2022.

[32] Eric Schechter. *Handbook of Analysis and its Foundations*. Academic Press, 1996.

[33] Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022.

[34] Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

[35] Peter Whittle. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):143–149, 1980.

[36] Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

[37] Shuang Wu, Jingyu Zhao, Guangjian Tian, and Jun

Wang. State-aware value function approximation with attention mechanism for restless multi-armed bandits. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 458–464, 8 2021.

[38] Qing Zhao. Multi-armed bandits: theory and applications to online learning in networks. *Synthesis Lectures on Communication Networks*, 12(1):1–165, 2019.