Annals of Operations Research
https://doi.org/10.1007/s10479-025-06821-3

ORIGINAL RESEARCH

On the sensitivity of restless bandits solutions to uncertainty in the models of the arms

Amit Sinha[1] · **Aditya Mahajan**[1]

Received: 8 March 2023 / Accepted: 18 August 2025

Abstract
Restless multi-armed bandits (RMAB) are a popular framework for modeling resource allocation and scheduling problems arising in various applications. Such applications can be modeled as Markov decision processes (MDP), but optimal or sub-optimal solution through dynamic programming suffer from high complexity. RMAB provides a heuristic solution, where the solution complexity scales linearly with the number of alternatives. However, these heuristic solutions are derived under the assumption that the model of all arms are known perfectly. In this paper, we consider RMAB with uncertainty in the rewards and dynamics of the arms. In such a setting, using a robust MDP solution is not possible due to high computational complexity. So, we consider a certainty equivalence approach and bound the additional loss in performance due to model inaccuracy. Our bounds are directly in terms of the model uncertainty of each arm and we illustrate their use via examples.

Keywords Restless multi-armed bandits · Model mismatch · Markov decision process · Certainty equivalence · Whittle index · Gittins index

1 Introduction

Markov decision processes (MDPs) are a popular framework for solving multi-stage decision problems (Puterman, 2014). Traditional MDP models capture aleatoric uncertainty as part of the model as a probability distribution over the next state and instantaneous rewards. However, such models do not capture epistemic uncertainty. One method to capture epistemic uncertainty is via the framework of robust MDPs (White & Eldeib, 1994; Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013; Tzortzis et al., 2015; Lam, 2016).

✉ Amit Sinha
amit.sinha@mail.mcgill.ca

Aditya Mahajan
aditya.mahajan@mcgill.ca

[1] Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada

Published online: 08 September 2025

However, solving robust MDPs is NP hard and the solutions, which provide worst case performance guarantees against all models in an uncertain set, are often too pessimistic.

An alternative approach is to simply choose the optimal policy corresponding to a nominal model from the uncertain set. Such an approach is often called certainty equivalence in the systems and control literature or the plug-in estimator in the artificial intelligence literature. Unlike the robust MDP solution, the certainty equivalent solution does not provide worst case performance guarantees. However, the certainty equivalent solution is computationally simpler and, might provide a satisficing solution (Simon, 1956) in many cases, especially when the epistemic uncertainty is not large.

In order to verify if a certainty equivalent solution is satisficing, we need to characterize the sensitivity of the solution of an MDP to model uncertainty. There is a rich literature on this topic (Whitt, 1978, 1979; Müller, 1997b; Asadi et al., 2018; Gelada et al., 2019; Kara & Yüksel, 2020) which characterizes the sensitivity of the optimal solution of MDPs to model uncertainty. However, directly using these results in a real world application, which typically has multiple components, can be challenging for two reasons. First, we typically have uncertainty estimates of the dynamics and rewards of each component while the sensitivity results for MDPs require uncertainty estimates of the coupled model. Second, it is often not possible to compute an optimal solution of a large MDP due to the curse of dimensionality. So, in practice, one often uses a domain specific heuristic solution. So, one needs to generalize the sensitivity results to provide sensitivity of heuristic solutions (rather than optimal solutions) to model uncertainty. In this paper, we illustrate how to circumvent these challenges for restless multi-armed bandits (RMAB) (Whittle, 1988), which is a modeling framework used to model and solve resource allocations and scheduling problems arising in various applications, including communication networks, power systems, and machine maintenance.

In a RMAB, a decision maker controls the evolution of n alternatives or arms. Each arm is a controlled Markov process which can be *active* or *passive* at each time. The decision maker can only activate m arms, where $m < n$. The objective is to determine which arms to activate at each time to maximize the expected discounted cost over an infinite horizon. Such a problem can be modeled as an MDP but obtaining an optimal solution is PSAPCE hard in the number of arms (Papadimitriou & Tsitsiklis, 1999).

Motivated by the low-complexity index-based solution to the *rested* MAB problem (Gittins & Jones, 1979), various low-complexity heuristic solutions have been proposed for restless MAB as well. The most popular heuristic is the Whittle index policy (Whittle, 1988), which has linear complexity in the number of alternatives. The Whittle index policy is optimal in some settings (e.g., when the arms which are not selected remain frozen (Gittins & Jones, 1979), when the number of arms is asymptotically large (Weber & Weiss, 1990), and when the model satisfies some separation conditions (Lott & Teneketzis, 2000)), and performs close to optimal in a variety of applications (Niño-Mora, 2007; Ansell et al., 2003; Glazebrook et al., 2005, 2006; Ayesta et al., 2010; Akbarzadeh & Mahajan, 2019). In addition to the Whittle index policy, other heuristic solutions to RMAB have also been proposed in the literature. These include primal-dual index heuristics (Bertsimas & Niño-Mora, 2000), linear programming based methods (Verloop, 2016; Zayas-Cabán et al., 2019; Gast et al., 2022), general Lagrangian relaxations (Hu & Frazier, 2017; Brown & Smith, 2020; Hodge & Glazebrook, 2011; Killian et al., 2021).

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 are heuristic solutions such as the Whittle index policy to uncertainty in the model of the arms?* In particular, if there is some uncertainty in the model of the arms (which could be due to 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 heuristic solution 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 heuristic solution.

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 the Gittins index policy is optimal (Gittins & Jones, 1979). The question of sensitivity of the Gittins index to model mismatch has been investigated in Katehakis and Veinott (1987). However, the result and the proof technique of Katehakis and Veinott (1987) 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 (Ben-Israel & Flåm, 1990), but they are also not applicable to the restless setting.

There is some work on the robust formulation for *rested* multi-armed bandits Caro and Das Gupta (2015), Kim and Lim (2016), Cohen and Treetanthiploet (2022). These results have been generalized to a certain class of partially observed models in Kim (2016). However, as far as we are aware, there are not results on the robust formulation for *restless* bandits.

Recently, there has been a significant interest in learning Whittle index policies for RMAB (Meshram et al., 2017; Borkar & Chadha, 2018; Fu et al., 201; Avrachenkov & Borkar, 2022; Robledo et al., 2022; Akbarzadeh & Mahajan, 2022b). 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. Alternative approaches to learning in restless bandits are presented in Tekin and Liu (2012), Liu et al. (2012), which learn the arm with the largest average reward. However, 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 a heuristic solution known as Whittle index policy to model mismatch. In particular, we formalize how to define model mismatch of an arm and characterize the sensitivity of the 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 heuristic solution 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 Sect. 2, we present the model and the problem formulation and state the main results. We present some examples of our results in Sect. 3. In Sect. 4, we present the proofs of the main results and conclude in Sect. 5.

1.1 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, $\Delta(\mathrm{X})$ 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 $f : \mathrm{S} \to \mathbb{R}$, we use $\operatorname{span}(f)$ to denote the span of *f*, i.e., $\operatorname{span}(f) = \sup_{s,s' \in \mathrm{S}} |f(s) - f(s')|$ and we use $\|f\|_\infty$ to denote the supremum norm of function *f*, i.e., $\|f\|_\infty = \sup_{s \in \mathrm{S}} f(s)$.

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

$$\operatorname{Lip}(f) = \sup_{s,s' \in \mathrm{S}} \frac{|f(s) - f(s')|}{d(s,s')}. \tag{1}$$

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 $\alpha^i = \langle \mathrm{S}^i, \{0,1\}, \{p^i(a)\}_{a \in \{0,1\}}, r^i \rangle$, 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 $r^i : \mathrm{S}^i \times \{0,1\} \mapsto \mathbb{R}$ 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 $S_t^i \in \mathrm{S}^i$ and $A_t^i \in \{0,1\}$ to denote the state and action of arm *i* at time *t*. We use $\boldsymbol{S}_t = (S_t^1, \ldots, S_t^n)$ and $\boldsymbol{A}_t = (A_t^1, \ldots, A_t^n)$ 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 $\mathsf{B}^i \subset \mathrm{S}^i$, $i \in [n]$, we have

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

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

$$V^\pi(\boldsymbol{s}_0) = Q^\pi(\boldsymbol{s}_0, \pi(\boldsymbol{s}_0)), \tag{2}$$

where

$$Q^\pi(\boldsymbol{s}_0, \boldsymbol{a}_0) \quad = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t \sum_{i\in[n]} r^i(S^i_t, A^i_t) \;\middle|\; \boldsymbol{S}_0 = \boldsymbol{s}_0, \boldsymbol{A}_0 = \boldsymbol{a}_0\right], \tag{3}$$

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 Puterman (2014). However, the dynamic programming solution suffers from the curse of dimensionality because both the state space S and action space $\mathrm{A}(m)$ grow exponentially with the number of arms. To avoid the curse of dimensionality, a popular heuristic is to use the Whittle index policy (Whittle, 1988), which has a linear complexity in the number of arms. We provide an overview of the Whittle index policy below.

2.2 Indexability and the Whittle index policy

Consider an arm, say arm $\alpha^i = \langle \mathrm{S}^i, \{0,1\}, \{p^i(a)\}_{a\in\{0,1\}}, r^i\rangle$ as described before. For any $\lambda \in \mathbb{R}$, consider a modified version of the arm denoted by $\alpha^i_\lambda = \langle \mathrm{S}^i, \{0,1\}, \{p^i(a)\}_{a\in\{0,1\}}, r^i_\lambda\rangle$, where for all $s^i \in \mathrm{S}^i$, $a^i \in \{0,1\}$, we have

$$r^i_\lambda(s^i, a^i) := r^i(s^i, a^i) + \lambda a^i.$$

Now, consider the problem of individually controlling the modified arm α^i_λ. This auxiliary problem is an MDP and the optimal solution is given by the following dynamic program: find $V^{i,\star}_\lambda : \mathrm{S}^i \to \mathbb{R}$ that satisfies the Bellman optimality equation:

$$V^{i,\star}_{i\lambda}(s^i) = \max_{a^i\in\{0,1\}} \left\{ r^i_\lambda(s^i, a^i) + \gamma \int_{\mathrm{S}^i} p^i(\bar{s}^i \mid s^i, a^i) V^{i,\star}_\lambda(\bar{s}^i) d\bar{s}^i \right\}, \quad \forall s^i \in \mathrm{S}^i. \tag{4}$$

Let $\pi^{i,\star}_\lambda(s^i)$ denote the arg max of the right hand side of (4), where we choose $\pi^{i,\star}_\lambda(s^i) = 1$ when the arg max is not unique. Then, standard results from Markov decision theory (Puterman, 2014) imply that the policy $\pi^{i,\star}_\lambda$ is optimal for controlling the modified arm α^i_λ.

Define the active set Π^i_λ to be the states where the active action ($a^i = 1$) is optimal, i.e., $\Pi^i_\lambda := \{s^i \in \mathrm{S}^i : \pi^{i\star}_\lambda = 1\}$.

Definition 1 (Indexability) An arm α^i is *indexable* if the active set Π^i_λ is non-increasing in λ, i.e., for any $\lambda^1, \lambda^2 \in \mathbb{R}$ such that $\lambda^1 \leq \lambda^2$, we have $\Pi^i_{\lambda^1} \subseteq \Pi^i_{\lambda^2}$.

Definition 2 (Whittle index) The Whittle index $\omega^i : \mathrm{S}^i \to \mathbb{R}$ of an indexable arm α^i is defined as follows. for a state $s^i \in \mathrm{S}^i$, $\omega^i(s^i)$ is the smallest value of λ for which s^i is part of the active set Π^i_λ, i.e., $\omega(s^i) = \inf\{\lambda \in \mathbb{R} : s^i \in \Pi^i_\lambda\}$.

Alternatively, the Whittle index $\omega^i(s^i)$ is a value of the penalty λ for which the optimal policy is indifferent between taking the active and the passive action when the arm is in state s^i.

A restless bandit problem is said to be indexable if all arms are indexable. For an indexable restless bandit problem, the *Whittle index policy* is a heuristic policy which is defined as follows: First, we compute the Whittle indices of all arms $\{\alpha^i\}_{i\in[n]}$ offline. Then, at each time, we obtain the Whittle indices of the current state of all arms and play the arms with the m largest Whittle indices.

Various sufficient conditions for checking indexability and computing the Whittle index have been proposed in the literature. See Bertsimas and Niño-Mora (2000), Glazebrook et al. (2005), Glazebrook et al. (2006), Niño-Mora (2007), Akbarzadeh and Mahajan (2019), Gast et al. (2022) and references therein. Although the Whittle index policy is a heuristic, as mentioned in the introduction, it is optimal in some settings Gittins and Jones (1979), Weber and Weiss (1990), Lott and Teneketzis (2000) and performs close to optimal in a variety of applications (Niño-Mora, 2007; Ansell et al., 2003; Glazebrook et al., 2005, 2006; Ayesta et al., 2010; Akbarzadeh & Mahajan, 2019).

2.3 Problem formulation: model mismatch in RMAB

We start by defining a class of metrics on probability measures known as integral probability metrics (IPM) (Müller, 1997a).

Definition 3 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

$$d_{\mathfrak{F}}(\mu, \nu) := \sup_{f\in\mathfrak{F}} \left| \int_{\mathrm{X}} f d\mu - \int_{\mathrm{X}} f d\nu \right|.$$

Some examples of IPM are total variation distance, Wasserstein distance, Kolmogorov distance, Bounded-Lipschitz distance, and maximum mean discrepancy. For total variation distance, $\mathfrak{F} = \{f : \frac{1}{2}\operatorname{span}(f) \leq 1\} =: \mathfrak{F}^{\mathrm{TV}}$; for Wasserstein distance, $\mathfrak{F} = \{f : \operatorname{Lip}(f) \leq 1\} =: \mathfrak{F}^{\mathrm{W}}$. We refer the reader to Subramanian et al. (2022) for details about other examples.

Given a function class $\mathfrak{F}$ and a function f (not necessarily in $\mathfrak{F}$), the Minkowski functional (Schechter, 1996) of f with respect to $\mathfrak{F}$ is defined as:

$$\rho_{\mathfrak{F}}(f) := \inf\{\rho \in \mathbb{R}_{>0} : \rho^{-1} f \in \mathfrak{F}\}. \tag{5}$$

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

$$\left| \int_{\mathsf{X}} f d\mu - \int_{\mathsf{X}} f d\nu \right| \leq \rho_{\mathfrak{F}}(f) \cdot d_{\mathfrak{F}}(\mu, \nu), \tag{6}$$

We now formalize the notion of approximate restless bandit model.

Definition 4 Consider two arms $\alpha = \langle \mathrm{S},\ \{0,1\},\ \{p(a)\}_{a\in\{0,1\}}, r\rangle$ and $\hat{\alpha} = \langle \hat{\mathrm{S}}, \{0,1\}, \{\hat{p}(a)\}_{a\in\{0,1\}},\ \hat{r}\rangle$ defined on different state spaces S and $\hat{\mathrm{S}}$. We are also given a measurable aggregation function $\phi : \mathrm{S} \to \hat{\mathrm{S}}$. 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 $s \in \mathrm{S}$ and $a \in \{0,1\}$:

$$\left| r(s,a) - \hat{r}(\phi(s),a) \right| \leq \varepsilon, \quad d_{\mathfrak{F}}\big(\tilde{p}(\cdot|s,a), \hat{p}(\cdot|\phi(s),a)\big) \leq \delta,$$

where $\tilde{p}$ accumulates the probabilities of the true model in the aggregated state space in the following sense: for any Borel subset $\hat{B}$ of $\hat{\mathrm{S}}$

$$\tilde{p}(\hat{B} \mid s,a) := \int_{s'\in\mathrm{S}} \mathbb{1}\{\phi(s') \in \hat{B}\} p(ds'|s,a).$$

Lemma 1 Let $\hat{f} : \hat{\mathrm{S}} \to \mathbb{R}$ and $f = \hat{f} \circ \phi$ and $\tilde{p}, p$ be as described previously. Then we have for all $s \in \mathrm{S}$ and $a \in \{0,1\}$:

$$\int_{\hat{s}'\in\hat{\mathrm{S}}} \hat{f}(\hat{s}') \tilde{p}(d\hat{s}' \mid s,a) = \int_{s'\in\mathrm{S}} f(s') p(ds' \mid s,a).$$

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

Approximation setup Given a RMAB $\{\alpha^i\}_{i\in[n]}$, where $\alpha^i = \langle \mathrm{S}^i, \{0,1\},\ \{p^i(a)\}_{a\in\{0,1\}}, r^i\rangle$, consider an approximate RMAB $\{\hat{\alpha}^i\}_{i\in[n]}$, where $\hat{\alpha}^i = \langle \hat{\mathrm{S}}^i, \{0,1\}, \{\hat{p}^i(a)\}_{a\in\{0,1\}}, \hat{r}^i\rangle$ with aggregation function $\phi^i : \mathrm{S}^i \to \hat{\mathrm{S}}^i$ such that arm $\hat{\alpha}^i$ is an $(\varepsilon^i, \delta^i)$-approximation of arm α^i.

Let $\hat{\mathrm{S}} = \prod_{i\in[n]} \hat{\mathrm{S}}^i$. Define aggregation function $\phi : \mathrm{S} \to \hat{\mathrm{S}}$ given by $\phi(s^1, \ldots, s^n) = (\phi^1(s^1), \ldots, \phi^n(s^n))$. For any policy $\hat{\pi} : \hat{\mathrm{S}} \to \mathrm{A}(m)$ and initial state $\boldsymbol{s}$, let $\hat{V}^{\hat{\pi}}(\hat{\boldsymbol{s}})$ denote the performance of policy $\hat{\pi}$ in RMAB $\{\hat{\alpha}^i\}_{i\in[n]}$. Let $\pi = \hat{\pi} \circ \phi$ denote the "lifting" of the approximate policy to the original space. let $V^{\pi}(\boldsymbol{s})$ denote the performance of π in RMAB $\{\alpha^i\}_{i\in[n]}$. Let π^* denote the optimal policy for the true model $\{\alpha^i\}_{i\in[n]}$ and

let $\hat{\pi}^*$ denote the optimal policy for the approximate model $\{\hat{\alpha}^i\}_{i\in[n]}$. Note that $\pi^* \neq \hat{\pi}^* \circ \phi$ in general, because the "lifting" notation is defined for a general policy $\hat{\pi}$, whereas it need not be true that the lifted optimal approximate policy is the same as the optimal policy.

Definition 5 The *sensitivity* of any approximate policy $\hat{\pi} : \hat{\mathrm{S}} \to \mathrm{A}(m)$ is defined as $\mathrm{Gap}^{\pi} - \widehat{\mathrm{Gap}}^{\hat{\pi}}$, where $\pi = \hat{\pi} \circ \phi$, $\mathrm{Gap}^{\pi} := \|V^{\pi^*} - V^{\pi}\|_{\infty}$ is the sub-optimality gap in using lifted approximate policy π in the true model $\{\alpha^i\}_{i\in[n]}$ and $\widehat{\mathrm{Gap}}^{\hat{\pi}} := \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\pi}}\|_{\infty}$ is the sub-optimality gap in using approximate policy $\hat{\pi}$ in the approximate model $\{\hat{\alpha}^i\}_{i\in[n]}$.

Let $\hat{\mu}$ be any heuristic policy for the approximate model $\{\hat{\alpha}^i\}_{i\in[n]}$. We are interested in the following approximation characterization.

Problem 1 For the approximation setup described above, characterize the sensitivity $\mathrm{Gap}^{\hat{\mu}\circ\phi} - \widehat{\mathrm{Gap}}^{\hat{\mu}}$, i.e., the additional loss in performance when using the heuristic solution corresponding to an approximate model in the true model, in terms of the approximation errors $\{(\varepsilon^i, \delta^i)\}_{i\in[n]}$.

We present an upper bound on the sensitivity gap in Sect. 2.4. We illustrate these bounds via examples in Sect. 3, where we will use the Whittle index policy as a heuristic. For these examples, we will impose the following assumption on the model.

Assumption 1 All arms $\{\hat{\alpha}^i\}_{i\in[n]}$ are indexable.

Remark 1 A solution to Problem 1 also provides a solution to the robust RMAB problem. In particular, consider a setting where we do not know the exact model of the arms, but know that the RMAB belongs to a collection $\mathcal{A}$ of RMAB models, where all models in $\mathcal{A}$ have the same state spaces but different dynamics and rewards. Suppose $\{\hat{\alpha}^i\}_{i\in[n]}$ is a *nominal model* (not necessarily in $\mathcal{A}$). Let $\{(\varepsilon^i, \delta^i)\}_{i\in[n]}$ be such that for all $\{\alpha^i\}_{i\in[n]} \in \mathcal{A}$, and each $i \in [n]$, the arm $\hat{\alpha}^i$ is an $(\varepsilon^i, \delta^i)$-approximation of arm α^i. Then the solution to Problem 1 also provides a bound on the additional loss in performance when using a heuristic solution of the nominal model $\{\hat{\alpha}^i\}_{i\in[n]}$ instead of using the heuristic solution of the true (but unknown) model in $\mathcal{A}$.

2.4 Main result

For any Markov policy $\hat{\pi}$: $\hat{\mathrm{S}} \to \mathrm{A}(m)$, define

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

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 Sect. 2.3, we have*

$$\|Q^{\pi^*} - Q^{\hat{\mu}\circ\phi}\|_{\infty} - \|\hat{Q}^{\hat{\pi}^*} - \hat{Q}^{\hat{\mu}}\|_{\infty} \leq 3\beta_{\mathfrak{F}}^{\hat{\pi}^*} + \beta_{\mathfrak{F}}^{\hat{\mu}} \tag{7}$$

and

$$\|V^{\pi^*} - V^{\hat{\mu}\circ\phi}\|_\infty - \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty \leq 3\beta_{\mathfrak{F}}^{\hat{\pi}^*} + \beta_{\mathfrak{F}}^{\hat{\mu}}. \tag{8}$$

The proof is given in Sect. 4.

The results of Theorem 1 can be interpreted as follows. Suppose the heuristic solution is computed using a synthetic/simulation-based approximate model of the real-world. The results of Theorem 1 then characterize the sensitivity of the heuristic solution to model mismatch. In particular, if the synthetic model were accurate, we would incur a loss less than $\widehat{\mathrm{Gap}}^{\hat{\mu}}$. The results of Theorem 1 shows that when the synthetic model is not accurate, an additional loss of $\mathcal{E}_{\mathrm{model}}{:=}3\beta_{\mathfrak{F}}^{\hat{\pi}^*} + \beta_{\mathfrak{F}}^{\hat{\mu}}$ is incurred. The overall loss in using the heuristic solution of the approximate model in the real-world is less than $\mathrm{Gap}^{\hat{\mu}\circ\phi} \leq \widehat{\mathrm{Gap}}^{\hat{\mu}} + \mathcal{E}_{\mathrm{model}}$.

2.5 Discussion of the results

2.5.1 The dependence on IPM

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 Sect. 3.1 for an example on how the upper bound depends on the choice of the IPM.

2.5.2 The role of indexability

When the heuristic solution is the Whittle index policy, Theorem 1 requires only 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.

2.5.3 The special case of Gittins index

In the rested case (i.e., when only one arm can be activated at each time and the arms which are not activated remain frozen), the Whittle index policy reduces to the Gittins index policy and is optimal. Therefore, in (8), $\hat{V}^{\hat{\pi}^*} = \hat{V}^{\hat{\mu}}$ and $\mathcal{E}_{\mathrm{policy}} = 0$. Thus, Theorem 1 also provides an approximation guarantee for the rested RMAB which is different from the stopping-time based approximation guarantee in Katehakis and Veinott (1987).

2.5.4 Approximate optimality of Gittins index for "viscous when passive" restless bandits

Consider a restless bandit problem $\{\alpha^i\}_{i\in[n]}$, where $\alpha^i = \langle \mathrm{S}^i, \{0,1\}, \{p^i(a)\}_{a\in\{0,1\}}, r^i\rangle$ is such that the passive dynamics $\{p^i(0)\}_{i\in[n]}$ are close to being frozen, i.e., given a function class $\mathfrak{F}$, we have

$$d_{\mathfrak{F}}(p^i(\cdot \mid s, 0), \mathrm{id}^i(\cdot \mid s)) \leq \delta^i, \quad \forall s \in \mathrm{S}^i, i \in [n],$$

where $\mathrm{id}(\cdot \mid s)$ is a Dirac delta measure centered at s. We call such a model *viscous restless bandits*. Note that if $\delta^i = 0$ for all $i \in [n]$, then the above model reduces to the classical *rested* multi-armed bandit model, for which the Gittins index is optimal. However, when $\delta^i \neq 0$, then we are in general restless bandit settings and very little is known regarding the optimality of an index solution.

However, if $m = 1$ and $\{\delta^i\}_{i\in[n]}$ are small, then we can approximate the model $\{\alpha^i\}_{i\in[n]}$ by a model $\{\hat{\alpha}^i\}_{i\in[n]}$, where the approximate arm $\hat{\alpha}^i$ has the same state space and the same dynamics under active action as arm α^i, but the dynamics under the passive action is id (thus, frozen). Thus, we can approximate the "viscous when passive" bandits by rested badits.

Let $\hat{\pi}^*$ be the Gittins index policy for $\{\hat{\alpha}^i\}_{i\in[n]}$, which is optimal for that model. The result of Theorem 1 shows that

$$\|V^{\pi^*} - V^{\hat{\pi}^* \circ \phi}\|_\infty \leq 3\beta_{\mathfrak{F}}^{\hat{\pi}^*} + \beta_{\mathfrak{F}}^{\hat{\pi}^*}.$$

This quantifies the loss in performance incurred by the Gittins index policy in a "viscous when passive" restless bandit model.

2.6 Instance independent bounds

The bounds of Theorem 1 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

$$\begin{aligned} \|Q^{\pi^*} - Q^{\hat{\mu}\circ\phi}\|_\infty - \|\hat{Q}^{\hat{\pi}^*} - \hat{Q}^{\hat{\mu}}\|_\infty &\leq \frac{4\varepsilon}{(1-\gamma)} \\ &+ \frac{3\gamma\delta\,\mathrm{span}\,(\hat{\boldsymbol{r}})}{2(1-\gamma)^2} + \frac{\gamma\delta\,\mathrm{span}\,(\hat{V}^{\hat{\mu}})}{2(1-\gamma)} \end{aligned} \tag{9}$$

and

$$\begin{aligned} \|V^{\pi^*} - V^{\hat{\mu}\circ\phi}\|_\infty - \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty &\leq \frac{4\varepsilon}{(1-\gamma)} \\ &+ \frac{3\gamma\delta\,\mathrm{span}\,(\hat{\boldsymbol{r}})}{2(1-\gamma)^2} + \frac{\gamma\delta\,\mathrm{span}\,(\hat{V}^{\hat{\mu}})}{2(1-\gamma)}, \end{aligned} \tag{10}$$

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

See Sect. 4.4 for proof.

We now define a property of an arm.

Definition 6 Consider the function class $\mathfrak{F}^{\mathrm{W}}$, an arm $\alpha^i = \langle \mathrm{S}^i, \{0,1\}, \{p^i(a)\}_{a\in\{0,1\}}, r^i\rangle$ and a metric d^i on S^i. If

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

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

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

Similarly, consider the approximate arm $\hat{\alpha}^i = \langle \hat{\mathrm{S}}^i, \{0,1\}, \{\hat{p}^i(a)\}_{a\in\{0,1\}}, \hat{r}^i \rangle$ and a metric $\hat{d}^i$ on $\hat{\mathrm{S}}$. We define $(\mathrm{L}_{\hat{r}^i}, \mathrm{L}_{\hat{p}^i})$ analogously to Definition 6. If both of these are finite, then the arm $\hat{\alpha}^i$ is said to be $(\mathrm{L}_{\hat{r}^i}, \mathrm{L}_{\hat{p}^i})$-Lipschitz.

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

$$\begin{aligned} \|Q^{\pi^*} - Q^{\hat{\mu}\circ\phi}\|_\infty - \|\hat{Q}^{\hat{\pi}^*} - \hat{Q}^{\hat{\mu}}\|_\infty &\le \frac{4\varepsilon}{(1-\gamma)} \\ &+ \frac{3\gamma\delta \mathrm{L}_{\hat{\boldsymbol{r}}}}{(1-\gamma)(1-\gamma \mathrm{L}_{\hat{\boldsymbol{p}}})} + \frac{\gamma\delta \,\mathrm{Lip}\,(\hat{V}^{\hat{\mu}})}{(1-\gamma)} \end{aligned} \tag{11}$$

and

$$\begin{aligned} \|V^{\pi^*} - V^{\hat{\mu}\circ\phi}\|_\infty - \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty &\le \frac{4\varepsilon}{(1-\gamma)} \\ &+ \frac{3\gamma\delta \mathrm{L}_{\hat{\boldsymbol{r}}}}{(1-\gamma)(1-\gamma \mathrm{L}_{\hat{\boldsymbol{p}}})} + \frac{\gamma\delta \,\mathrm{Lip}\,(\hat{V}^{\hat{\mu}})}{(1-\gamma)}, \end{aligned} \tag{12}$$

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

See Sect. 4.5 for proof.

Remark 2 In order to compute the Lipschitz constant of $\hat{V}^{\hat{\mu}}$ in (11) and (12), we need a metric on $\hat{\mathrm{S}}$. This metric is chosen as $\hat{\boldsymbol{d}}(\hat{\boldsymbol{s}}_1, \hat{\boldsymbol{s}}_2) = \sum_{i\in[n]} \hat{d}^i(\hat{s}_1^i, \hat{s}_2^i)$.

Remark 3 The instance independent upper bounds of Propositions 1 and 2 still depend on the properties of the value function $\hat{V}^{\hat{\mu}}$. In many applications, this value function is computed numerically to characterize the performance of the proposed heuristic policy. If this value function is not available, we can upper bound its properties as follows.

1. In Proposition 1, we can show that $\mathrm{span}\,(\hat{V}^{\hat{\mu}}) \le \mathrm{span}\,(\hat{\boldsymbol{r}})/(1-\gamma)$ by following an argument similar to that used in the proof of Proposition 1.
2. In Proposition 2, using (Hinderer, 2005, Theorem 4.2), we can show that if the heuristic policy $\hat{\mu}$ is Lipschitz, then

$$\mathrm{Lip}\,(\hat{V}^{\hat{\mu}}) \leq \frac{\mathrm{L}_{\hat{\boldsymbol{r}}}(1+\mathrm{L}_{\hat{\mu}})}{1-\gamma \mathrm{L}_{\hat{\boldsymbol{p}}}(1+\mathrm{L}_{\hat{\mu}})},$$

where $\mathrm{L}_{\hat{\mu}}$ is the Lipschitz constant of the policy $\hat{\mu}$, and $\mathrm{L}_{\hat{\boldsymbol{p}}}(1+\mathrm{L}_{\hat{\mu}}) < \gamma^{-1}$.

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 $\alpha^i = \langle \mathrm{S}, \{0,1\}, \{P^i(a^i)\}_{a^i \in \{0,1\}}, r^i \rangle$, $i \in \{1,2\}$, where $\mathrm{S} = \{1,2,3\}$ shown in Fig. 1a. Suppose the arms are approximated by $\langle \mathrm{S}, \{0,1\}, \{\hat{P}^i(a^i)\}_{a^i \in \{0,1\}}, \hat{r}^i \rangle$ shown in Fig. 1b. Note that since $\hat{\mathrm{S}} = \mathrm{S}$, we take $\phi(s) = s$. Consider the heuristic solution to be the Whittle index policy. We used the open source python library provided in Gast et al. (2023) to verify that the given approximate model is indexable. Thus, Assumption 1 is satisfied.

$$P(0) = \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.3 & 0.3 \end{bmatrix}, \quad P(1) = \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.3 & 0.3 & 0.4 \\ 0.2 & 0.2 & 0.6 \end{bmatrix}, \quad r = \begin{bmatrix} 0.60 & 0.40 \\ 0.88 & 0.60 \\ 1.00 & 0.80 \end{bmatrix},$$

Arm 1

$$P(0) = \begin{bmatrix} 0.1 & 0.6 & 0.3 \\ 0.2 & 0.7 & 0.1 \\ 0.1 & 0.8 & 0.1 \end{bmatrix}, \quad P(1) = \begin{bmatrix} 0.50 & 0.40 & 0.10 \\ 0.30 & 0.60 & 0.10 \\ 0.25 & 0.55 & 0.20 \end{bmatrix}, \quad r = \begin{bmatrix} 0.52 & 0.64 \\ 0.44 & 0.96 \\ 0.76 & 0.44 \end{bmatrix}.$$

Arm 2

(a) True Model

$$\hat{P}(0) = \begin{bmatrix} 0.19 & 0.29 & 0.52 \\ 0.11 & 0.51 & 0.38 \\ 0.41 & 0.29 & 0.30 \end{bmatrix}, \quad \hat{P}(1) = \begin{bmatrix} 0.39 & 0.39 & 0.22 \\ 0.29 & 0.29 & 0.42 \\ 0.21 & 0.19 & 0.60 \end{bmatrix}, \quad \hat{r} = \begin{bmatrix} 0.596 & 0.404 \\ 0.872 & 0.596 \\ 0.996 & 0.792 \end{bmatrix},$$

Arm 1

$$\hat{P}(0) = \begin{bmatrix} 0.09 & 0.62 & 0.29 \\ 0.21 & 0.69 & 0.10 \\ 0.12 & 0.79 & 0.09 \end{bmatrix}, \quad \hat{P}(1) = \begin{bmatrix} 0.48 & 0.42 & 0.10 \\ 0.31 & 0.59 & 0.10 \\ 0.24 & 0.55 & 0.21 \end{bmatrix}, \quad \hat{r} = \begin{bmatrix} 0.512 & 0.636 \\ 0.432 & 0.968 \\ 0.756 & 0.448 \end{bmatrix}.$$

Arm 2

(b) Approximate Model

Fig. 1 The true and approximate model for the example of Sect. 3.1

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 (Akbarzadeh & Mahajan, 2022a), and they are given by

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

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

$$\hat{\mu}(s^1, s^2) = \arg\max_{i \in \{1,2\}} \hat{\omega}^i(s^i). \tag{13}$$

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 (Puterman, 2014). The value function is given by[1]

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

Since the model is small, we can compute the optimal value function (of the true model), which we do using the value iteration algorithm (Puterman, 2014). 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 $\|V^{\pi^*} - V^{\hat{\mu}\circ\phi}\|_\infty = 0.550$. Note that in practice we do not have access to the true model, so we cannot compute the suboptimality gap $\|V^{\pi^*} - V^{\hat{\mu}\circ\phi}\|_\infty$. 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 $d(\hat{s}_1^i, \hat{s}_2^i) = |\hat{s}_1^i - \hat{s}_2^i|$ as the metric on $\hat{\mathrm{S}}$).

We also compute the value function of the Whittle index policy and the optimal value function (for the approximate model) using $\boldsymbol{d}(\hat{\boldsymbol{s}}_1, \hat{\boldsymbol{s}}_2)$. These are given by

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

Table 1 Parameters involved in Theorem 1 for Example 3.1

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

$$\hat{V}^{\hat{\mu}} = \begin{bmatrix} 16.142 & 16.534 & 16.133 \\ 16.430 & 16.822 & 16.361 \\ 16.473 & 16.865 & 16.587 \end{bmatrix} \quad \text{and} \quad \hat{V}^{\hat{\pi}^*} = \begin{bmatrix} 16.349 & 16.741 & 16.597 \\ 16.641 & 17.033 & 16.889 \\ 16.683 & 17.075 & 16.931 \end{bmatrix}.$$

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, $\rho_{\mathfrak{F}}(\cdot) = \frac{1}{2}\,\mathrm{span}\,(\cdot)$. Thus, the result (8) of Theorem 1 simplifies to

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

Now consider the case when $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$. In this case, $\rho_{\mathfrak{F}}(\cdot) = \mathrm{Lip}\,(\cdot)$. Thus, the result (8) of Theorem 1 simplifies to

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

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 space

We now consider a model for machine maintenance with *n* machines and *m* repair persons. Each machine has a state $s \in \mathrm{S} := [0, 1]$, where $s = 0$ denotes a machine in a pristine state and $s = 1$ denotes a completely deteriorated machine. Active action $a = 1$ corresponds to a repair person servicing a machine in which case a per-step cost of c^i is incurred and the state of the serviced machine resets to a pristine state. Passive action $a = 0$ corresponds to the machine not being serviced in which case a per-step cost of $\xi^i s$ is incurred, where ξ^i is machine dependent coefficient and the state s deteriorates to a worse state in $[s, 1]$ uniformly at random.

Thus,

$$r^i(s^i,0) = -\xi^i s^i, \qquad r^i(s^i,1) = -c^i,$$
$$p^i(\cdot|s^i,0) = \mathcal{U}(s^i,1), \quad p^i(\cdot|s^i,1) = \delta_D(\cdot),$$

where $i \in \{1,2\}$, $s^i \in \mathrm{S}$, $\mathcal{U}(x,y)$ denotes a uniform distribution on the interval $[x,y]$, and $\delta_D(\cdot)$ is the Dirac delta distribution centered at origin.

Consider the heuristic solution to be the Whittle index policy. 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

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

and consider the centers of each interval given by

$$\hat{\mathrm{S}} = \left\{\tfrac{1}{2H}, \tfrac{3}{2H}, \ldots, \tfrac{2H-1}{2H}\right\}.$$

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

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

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

$$\hat{p}^i(\hat{s}^i_+|\hat{s}^i,0) = \begin{cases} \frac{1}{k(\hat{s}^i)}, & \text{if } \hat{s}^i_+ > \hat{s}^i \\ 0 & \text{otherwise} \end{cases}, \quad \hat{p}^i(\hat{s}^i_+|\hat{s}^i,1) = \begin{cases} 1 & \text{if } \hat{s}^i_+ = \frac{1}{2H} \\ 0 & \text{otherwise.} \end{cases}$$

where $k(\hat{s}^i)$ is a normalizing factor and

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

Since the approximate model satisfies the restart property of Akbarzadeh and Mahajan (2022a, 2019), it is indexable. Thus, Assumption 1 is satisfied. We now consider two instances of this model.

3.2.1 Case 1: An illustrative small-scale example

We consider $n = 2$ and $m = 1$ and take $\xi^1 = 1.0, \xi^2 = 0.5, c^1 = 0.7, c^2 = 0.3$ and $\gamma = 0.9$. Since the approximate arms have a finite state space, we can use the modified adaptive greedy algorithm of Akbarzadeh and Mahajan (2019) to compute the Whittle indices $\hat{\omega}^i(\hat{s})$, $i \in [n]$, of the approximate model. The computed indices are shown in Fig. 2.

To compute the sub-optimality gap, 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 and $\hat{d}(\hat{s},\hat{s}') = |\hat{s} - \hat{s}'|$ as the metric on $\hat{\mathrm{S}}$).

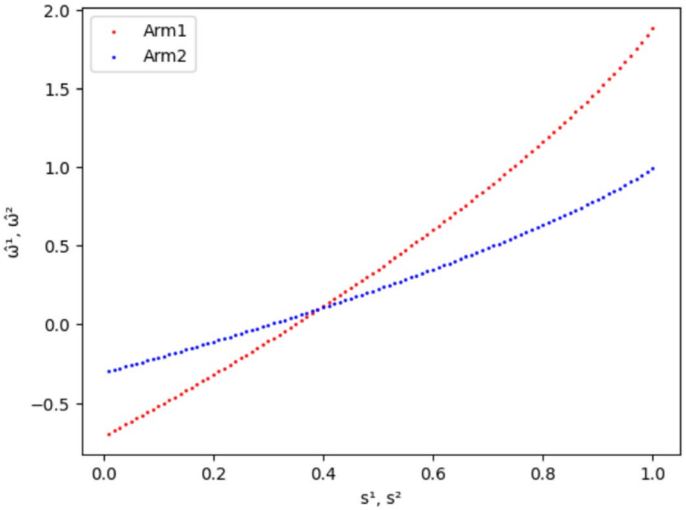
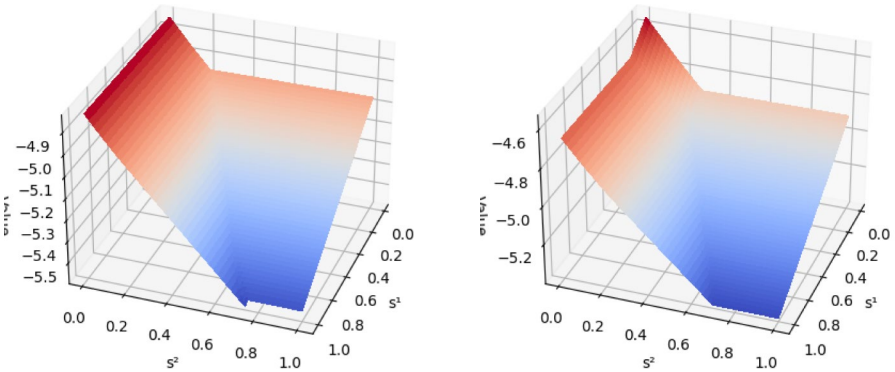


Fig. 2 Whittle indices $\hat{\omega}$ plotted for all states for the example of Sect. 3.2.1

Table 2 Approximation errors for the example of Sect. 3.2.1

Parameter	Arm 1	Arm 2	Overall
ε	0.005	0.0025	0.0075
$\delta_{\tilde{S}^W}$	0.005	0.005	0.01



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

Fig. 3 Value functions $\hat{V}^{\hat{\mu}}$ and $\hat{V}^{\hat{\pi}^*}$ plotted for all states for the example of Sect. 3.2.1

The state space of the approximate model is 100^2, which is not too large. So, it is possible to compute the value function of the Whittle index policy $\hat{V}^{\hat{\mu}}$, which we do using the policy evaluation equation (Puterman, 2014). 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 $\hat{V}^{\hat{\pi}^*}$

using the value iteration algorithm (Puterman, 2014). The value function can be visualized by the 3D plot in Fig. 3b. By comparing these two value functions, we get that the Whittle index policy has a suboptimality gap of $\|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty = 0.295$ in the approximate model.

To bound the suboptimality gap in the true model, we need the Lipschitz constants $\mathrm{Lip}\,(\hat{V}^{\hat{\pi}^*})$ and $\mathrm{Lip}\,(\hat{V}^{\hat{\mu}})$. We compute these by numerically maximizing the right hand side of (1) over all states and find that $\mathrm{Lip}\,(\hat{V}^{\hat{\pi}^*}) = 1.4$ and $\mathrm{Lip}\,(\hat{V}^{\hat{\mu}}) = 4.0$. Now, using (8) of Theorem 1, we get

$$\begin{aligned} &\|V^{\pi^*} - V^{\hat{\mu}\circ\phi}\|_\infty \\ &\quad \le \frac{4\varepsilon}{(1-\gamma)} + \frac{3\gamma\delta\,\mathrm{Lip}\,(\hat{V}^{\hat{\pi}^*})}{(1-\gamma)} + \frac{\gamma\delta\,\mathrm{Lip}\,(\hat{V}^{\hat{\mu}})}{(1-\gamma)} + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty \qquad (14) \\ &\le \frac{4\times 0.0075}{(1-0.9)} + \frac{3\times 0.9\times 0.01\times 1.4}{(1-0.9)} + \frac{0.9\times 0.01\times 4.0}{(1-0.9)} + 0.295 \qquad (15) \\ &\quad = 0.3 + 0.378 + 0.36 + 0.295 = 1.333 \end{aligned}$$

Next, note that the difference in suboptimality gaps

$$\|V^{\pi^*} - V^{\hat{\mu}\circ\phi}\|_\infty - \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty = 1.038.$$

To put this bound in perspective, observe from Fig. 3a that $\max_{\boldsymbol{s}\in\mathrm{S}} \hat{V}^{\hat{\mu}}(\phi(\boldsymbol{s})) = -4.8$ and $\min_{\boldsymbol{s}\in\mathrm{S}} \hat{V}^{\hat{\mu}}(\phi(\boldsymbol{s})) = -5.5$. Thus, the sensitivity of Whittle index policy to discretization is 18–21% of the approximate value function (depending on the start state). This large gap suggests that a finer discretization would be needed in this case.

3.2.2 Case 2: Large-scale example

We consider $n = 100$ and $m = 40$ and take a randomly generated instance where $\xi^i \sim \mathcal{U}(0,1)$, $c^i \sim \mathcal{U}(0,1)$, and $\gamma = 0.9$. We compute the Whittle index of each arm using the algorithm proposed in Gast et al. (2023).

As in Case 1, we have $\varepsilon^i = \xi^i/2\,H$ and $\delta^i = 1/2\,H$. For our randomly sampled instance, we obtain

$$\varepsilon = \sum_{i\in[n]} \varepsilon^i = 0.2203, \qquad \delta = \sum_{i\in[n]} \delta^i = 0.5$$

However, when we take $H = 100$, the state space of the approximate model is 100^{100}, so it is not possible to compute $\hat{V}^{\hat{\mu}}$ or $\hat{V}^{\hat{\pi}^*}$ as we did in the small scale model. So, we focus on the distance of the sensitivity gap of the true and the approximate model and bound or approximate $\mathrm{Lip}\,(\hat{V}^{\hat{\pi}^*})$ and $\mathrm{Lip}\,(\hat{V}^{\hat{\mu}})$ respectively.

For $\mathrm{Lip}\,(\hat{V}^{\hat{\pi}^*})$, we use the upper bound used in Prop. 2. We first compute $(\mathrm{L}_{\hat{r}^i}, \mathrm{L}_{\hat{p}^i})$ for each arm $i \in [n]$ using Definition 6. We can then compute $(\mathrm{L}_{\hat{\boldsymbol{r}}}, \mathrm{L}_{\hat{\boldsymbol{p}}})$ for the MDP corresponding to our randomly sampled instance using Lemma 4 (with $k = \infty$) as

$$\mathrm{L}_{\hat{\boldsymbol{r}}} \le \max_{i\in[n]} \mathrm{L}_{\hat{r}^i} = 0.9905, \qquad \mathrm{L}_{\hat{\boldsymbol{p}}} \le \max_{i\in[n]} \mathrm{L}_{\hat{p}^i} = 0.5.$$

Then, from (Hinderer, 2005, Theorem 4.2), we get that

$$\mathrm{Lip}\,(\hat{V}^{\hat{\pi}^*}) \leq \frac{\mathrm{L}_{\hat{\boldsymbol{r}}}^{(k)}}{(1-\gamma \mathrm{L}_{\hat{\boldsymbol{p}}}^{(k)})}.$$

We take a different approach to approximate $\mathrm{Lip}\,(\hat{V}^{\hat{\mu}})$. Instead of exact policy evaluation (which is intractable), we approximate $\hat{V}^{\hat{\mu}}$ with a neural network. To train the neural network, we randomly sample $N_e = 10^5$ initial conditions. For each initial condition $\hat{\boldsymbol{s}}$, we sample $S_e = 50$ trajectories by rolling out $T = 50$ steps of the policy $\hat{\mu}$. Note that $\gamma^T = 0.9^{50} \approx 5 \times 10^{-3}$, which is small in comparison to the total return, so we can approximate the infinite horizon return by the return of the finite length trajectory. Thus, we approximate $\hat{V}^{\hat{\mu}}(\hat{\boldsymbol{s}})$ by averaging the discounted returns of the S_e trajectories (this means we finally have a training batch of size 10^5). Then, we train a neural network using batch gradient descent to approximately learn $\hat{V}^{\hat{\mu}}$. The hyper-parameters used for training are given in Appendix D.

Given the neural network to approximate $\hat{V}^{\hat{\mu}}$, we approximate $\mathrm{Lip}\,(\hat{V}^{\hat{\mu}})$ by taking $N_L = 10^5$ randomly generated pairs of initial conditions $\hat{\boldsymbol{s}}_1$ and $\hat{\boldsymbol{s}}_2$, and maximizing $\|\hat{V}^{\hat{\mu}}(\hat{\boldsymbol{s}}_1) - \hat{V}^{\hat{\mu}}(\hat{\boldsymbol{s}}_2)\|_1 / \|\hat{\boldsymbol{s}}_1 - \hat{\boldsymbol{s}}_2\|_1$ over these initial conditions. Our calculations give $\mathrm{Lip}\,(\hat{V}^{\hat{\mu}}) \approx 0.2103$.

We can now bound the suboptimality gap in the true model using (8) of Theorem 1:

$$\begin{aligned}
&\|V^{\pi^*} - V^{\hat{\mu}\circ\phi}\|_\infty - \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty \\
&\quad \leq \frac{4\varepsilon}{(1-\gamma)} + \frac{3\gamma\delta \mathrm{L}_{\hat{\boldsymbol{r}}}}{(1-\gamma)(1-\gamma \mathrm{L}_{\hat{\boldsymbol{p}}})} + \frac{\gamma\delta\,\mathrm{Lip}\,(\hat{V}^{\hat{\mu}})}{(1-\gamma)} \\
&\quad = \frac{4 \times 0.2203}{(1-0.9)} + \frac{3 \times 0.9 \times 0.5 \times 0.9905}{(1-0.9)(1-0.9 \times 0.5)} + \frac{0.9 \times 0.5 \times 0.2103}{(1-0.9)} \\
&\quad = 8.812 + 24.312 + 0.9464 = 34.0704.
\end{aligned}$$

To put this bound in perspective, we use the random rollout data used to approximate $\mathrm{Lip}\,(\hat{V}^{\hat{\mu}})$ and obtain $\max_{\boldsymbol{s}\in\mathrm{S}} \hat{V}^{\hat{\mu}}(\phi(\boldsymbol{s})) = -156.99$ and $\min_{\boldsymbol{s}\in\mathrm{S}} \hat{V}^{\hat{\mu}}(\phi(\boldsymbol{s})) = -165.73$. Thus, the sensitivity of Whittle index policy to discretization is 20–22% of the approximate value function (depending on the start state). This is a large gap, but we conjecture that this is due to the crude bound on $\mathrm{Lip}\,(\hat{V}^{\hat{\pi}^*})$ and that the gap will be tighter if the Lipschitz constant of the optimal value function of the approximate model could be computed more accurately.

4 Proof of main result

4.1 Roadmap of the proof

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

$$\boldsymbol{p}(d\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) = \prod_{i\in[n]} p^i(ds^i_{t+1} \mid s^i_t, a^i_t), \tag{16}$$

$$\boldsymbol{r}(\boldsymbol{s}_t, \boldsymbol{a}_t) = \sum_{i\in[n]} r^i(s^i_t, a^i_t). \tag{17}$$

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

$$\hat{\boldsymbol{p}}(d\hat{\boldsymbol{s}}_{\boldsymbol{t+1}} \mid \hat{\boldsymbol{s}}_{\boldsymbol{t}}, \boldsymbol{a}_t) = \prod_{i\in[n]} \hat{p}^i(d\hat{s}^i_{t+1} \mid \hat{s}^i_t, a^i_t), \tag{18}$$

$$\hat{\boldsymbol{r}}(\hat{\boldsymbol{s}}_{\boldsymbol{t}}, \boldsymbol{a}_t) = \sum_{i\in[n]} \hat{r}^i(\hat{s}^i_t, a^i_t). \tag{19}$$

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 (Gelada et al., 2019; Subramanian et al., 2022) to derive approximation bounds for RMABs. In the rest of this section, we formalize this intuition.

4.2 Preliminary results

Definition 7 Consider the two MDPs $\mathcal{M} = \langle \mathrm{S}, \mathrm{A}(m), \boldsymbol{p}, \boldsymbol{r}\rangle$ and $\hat{\mathcal{M}} = \langle \hat{\mathrm{S}}, \mathrm{A}(m), \hat{\boldsymbol{p}}, \hat{\boldsymbol{r}}\rangle$ which are defined on the same action space. We are also given a measurable aggregation function $\phi : \mathrm{S} \to \hat{\mathrm{S}}$. 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 $\boldsymbol{s} \in \mathrm{S}$ and $\boldsymbol{a} \in \mathrm{A}(m)$:

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

where $\tilde{\boldsymbol{p}}$ accumulates the probabilities of the true model in the aggregated state space in the following sense: for any Borel subset $\hat{B}$ of $\hat{\mathrm{S}}$

$$\tilde{\boldsymbol{p}}(\hat{B} \mid \boldsymbol{s}, \boldsymbol{a}) := \int_{\boldsymbol{s}'\in\mathrm{S}} \mathbb{1}\{\phi(\boldsymbol{s}') \in \hat{B}\}\boldsymbol{p}(d\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}).$$

An immediate implication of the definition of the accumulated measure $\tilde{\boldsymbol{p}}$ is the following.

Lemma 2 Let $\hat{f} : \hat{\mathrm{S}} \to \mathbb{R}$ and $f = \hat{f} \circ \phi$ and $\tilde{\boldsymbol{p}}, \boldsymbol{p}$ be as described previously. Then we have for all $\boldsymbol{s} \in \mathrm{S}$ and $\boldsymbol{a} \in \mathrm{A}(m)$:

$$\int_{\hat{\boldsymbol{s}}'\in\hat{\mathrm{S}}} \hat{f}(\hat{\boldsymbol{s}}')\tilde{\boldsymbol{p}}(d\hat{\boldsymbol{s}}' \mid \boldsymbol{s}, \boldsymbol{a}) = \int_{\boldsymbol{s}'\in\mathrm{S}} f(\boldsymbol{s}')\boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}).$$

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

Lemma 3 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

$$(\varepsilon, \delta) = \left(\sum_{i \in [n]} \varepsilon^i, \sum_{i \in [n]} \delta^i \right). \tag{20}$$

See Appendix B for proof.

From standard results of Markov decision theory (Puterman, 2014), we know that for a given policy π, the performance V^π defined by (2) satisfies the following fixed point equation:

$$V^\pi(\boldsymbol{s}) = Q^\pi(\boldsymbol{s}, \pi(\boldsymbol{s})), \tag{21a}$$

$$Q^\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}[\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a})] + \gamma \int_{\mathrm{S}} V^\pi(\boldsymbol{s}') \boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}). \tag{21b}$$

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

$$\hat{V}^{\hat{\pi}}(\hat{\boldsymbol{s}}) = \hat{Q}^{\hat{\pi}}(\hat{\boldsymbol{s}}, \hat{\pi}(\hat{\boldsymbol{s}})), \tag{22a}$$

$$\hat{Q}^{\hat{\pi}}(\hat{\boldsymbol{s}}, \boldsymbol{a}) = \mathbb{E}[\hat{\boldsymbol{r}}(\hat{\boldsymbol{s}}, \boldsymbol{a})] + \gamma \int_{\hat{\mathrm{S}}} \hat{V}^{\hat{\pi}}(\hat{\boldsymbol{s}}') \hat{\boldsymbol{p}}(d\hat{\boldsymbol{s}}' \mid \hat{\boldsymbol{s}}, \boldsymbol{a}). \tag{22b}$$

Then, we have the following.

Proposition 3 For the approximate setup described in Sect. 2.3 and for any policy $\hat{\pi}$

$$\|V^{\hat{\pi} \circ \phi} - \hat{V}^{\hat{\pi}} \circ \phi\|_\infty \leq \|Q^{\hat{\pi} \circ \phi} - \hat{Q}^{\hat{\pi}} \circ \phi\|_\infty \leq \beta_{\mathfrak{F}}^{\hat{\pi}}. \tag{23}$$

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

$$\|V^{\pi^*} - \hat{V}^{\hat{\pi}^*} \circ \phi\|_\infty \leq \|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*} \circ \phi\|_\infty \leq \beta_{\mathfrak{F}}^{\hat{\pi}^*}. \tag{24}$$

Therefore, by the triangle inequality

$$\|Q^{\pi^*} - Q^{\hat{\pi}^* \circ \phi}\|_\infty \leq 2\beta_{\mathfrak{F}}^{\hat{\pi}^*} \text{ and } \|V^{\pi^*} - V^{\hat{\pi}^* \circ \phi}\|_\infty \leq 2\beta_{\mathfrak{F}}^{\hat{\pi}^*}. \tag{25}$$

Proof For the proof of the first part of (23), observe that from (21) and (22) we have that for any $\boldsymbol{s} \in \mathrm{S}$,

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

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

$$
\|V^{\hat{\pi}\circ\phi} - \hat{V}^{\hat{\pi}} \circ \phi\|_\infty \leq \|Q^{\hat{\pi}\circ\phi} - \hat{Q}^{\hat{\pi}} \circ \phi\|_\infty. \tag{26}
$$

This proves the first part of (23). Now, we bound $\|Q^{\hat{\pi}\circ\phi} - \hat{Q}^{\hat{\pi}} \circ \phi\|_\infty$ as follows: for any fixed $\boldsymbol{s} \in \mathrm{S}$, $\boldsymbol{a} \in \mathrm{A}(m)$, from (21) and (22), we have

$$
\begin{aligned}
&|Q^{\hat{\pi}\circ\phi}(\boldsymbol{s}, \boldsymbol{a}) - \hat{Q}^{\hat{\pi}}(\phi(\boldsymbol{s}), \boldsymbol{a})| \\
&\overset{(c)}{\leq} |\mathbb{E}[\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a})] - \mathbb{E}[\hat{\boldsymbol{r}}(\phi(\boldsymbol{s}), \boldsymbol{a})]| + \gamma \int_{\mathrm{S}} |V^{\hat{\pi}\circ\phi}(\boldsymbol{s}') - \hat{V}^{\hat{\pi}}(\phi(\boldsymbol{s}'))| \boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) \\
&\quad + \gamma \left| \int_{\mathrm{S}} \hat{V}^{\hat{\pi}}(\phi(\boldsymbol{s}')) \boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) - \int_{\hat{\mathrm{S}}} \hat{V}^{\hat{\pi}}(\hat{\boldsymbol{s}}') \hat{\boldsymbol{p}}(d\hat{\boldsymbol{s}}' \mid \phi(\boldsymbol{s}), \boldsymbol{a}) \right| \\
&\overset{(d)}{\leq} \varepsilon + \gamma \|Q^{\hat{\pi}\circ\phi} - \hat{Q}^{\hat{\pi}}\|_\infty + \gamma \left| \int_{\hat{\mathrm{S}}} \hat{V}^{\hat{\pi}}(\hat{\boldsymbol{s}}') \tilde{\boldsymbol{p}}(d\hat{\boldsymbol{s}}' \mid \boldsymbol{s}, \boldsymbol{a}) - \int_{\hat{\mathrm{S}}} \hat{V}^{\hat{\pi}}(\hat{\boldsymbol{s}}') \hat{\boldsymbol{p}}(d\hat{\boldsymbol{s}}' \mid \phi(\boldsymbol{s}), \boldsymbol{a}) \right| \\
&\overset{(e)}{\leq} \varepsilon + \gamma \|Q^{\hat{\pi}\circ\phi} - \hat{Q}^{\hat{\pi}} \circ \phi\|_\infty + \gamma \rho_{\mathfrak{F}}(\hat{V}^{\hat{\pi}}) \delta,
\end{aligned} \tag{27}
$$

where (c) follows from the definition of $Q^{\hat{\pi}\circ\phi}$ and $\hat{Q}^{\hat{\pi}}$, adding and subtracting the $\hat{V}^{\hat{\pi}}$ term and the triangle inequality; (d) and (e) follow from (26), Lemma 2 with $\hat{f} = \hat{V}^{\hat{\pi}}$ and the definition of an (ε, δ)-approximation for an MDP. Supremizing the LHS of (27) over all $\boldsymbol{s}, \boldsymbol{a} \in \mathrm{S} \times \mathrm{A}(m)$ and re-arranging terms, we get

$$
\|Q^{\hat{\pi}\circ\phi} - \hat{Q}^{\hat{\pi}} \circ \phi\|_\infty \leq \frac{\varepsilon + \gamma \rho_{\mathfrak{F}}(\hat{V}^{\hat{\pi}}) \delta}{(1-\gamma)} = \beta_{\mathfrak{F}}^{\hat{\pi}}. \tag{28}
$$

This proves the second part of (23).

Similarly, for the proof of the first part of (24), observe that from (21) and (22) we have that for any $\boldsymbol{s} \in \mathrm{S}$,

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

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

$$\|V^{\pi^*} - \hat{V}^{\hat{\pi}^*} \circ \phi\|_\infty \leq \|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*} \circ \phi\|_\infty. \tag{29}$$

This proves the first part of (23). Now, we bound $\|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*} \circ \phi\|_\infty$ as follows: for any fixed $\boldsymbol{s} \in \mathrm{S}$, $\boldsymbol{a} \in \mathrm{A}(m)$, from (21) and (22), we have

$$\begin{aligned}
&|Q^{\pi^*}(\boldsymbol{s}, \boldsymbol{a}) - \hat{Q}^{\hat{\pi}^*}(\phi(\boldsymbol{s}), \boldsymbol{a})| \\
&\overset{(f)}{\leq} |\mathbb{E}[\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a})] - \mathbb{E}[\hat{\boldsymbol{r}}(\phi(\boldsymbol{s}), \boldsymbol{a})]| \\
&\quad + \gamma \int_{\mathrm{S}} |V^{\pi^*}(\boldsymbol{s}') - \hat{V}^{\hat{\pi}^*}(\phi(\boldsymbol{s}'))| \boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) \\
&\quad + \gamma \left| \int_{\mathrm{S}} \hat{V}^{\hat{\pi}^*}(\phi(\boldsymbol{s}')) \boldsymbol{p}(d\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) - \int_{\hat{\mathrm{S}}} \hat{V}^{\hat{\pi}^*}(\hat{\boldsymbol{s}}') \hat{\boldsymbol{p}}(d\hat{\boldsymbol{s}}' \mid \phi(\boldsymbol{s}), \boldsymbol{a}) \right| \\
&\overset{(g)}{\leq} \varepsilon + \gamma \|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*}\|_\infty \\
&\quad - \gamma \left| \int_{\hat{\mathrm{S}}} \hat{V}^{\hat{\pi}^*}(\hat{\boldsymbol{s}}') \tilde{\boldsymbol{p}}(d\hat{\boldsymbol{s}}' \mid \boldsymbol{s}, \boldsymbol{a}) - \int_{\hat{\mathrm{S}}} \hat{V}^{\hat{\pi}^*}(\hat{\boldsymbol{s}}') \hat{\boldsymbol{p}}(d\hat{\boldsymbol{s}}' \mid \phi(\boldsymbol{s}), \boldsymbol{a}) \right| \\
&\overset{(h)}{\leq} \varepsilon + \gamma \|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*} \circ \phi\|_\infty + \gamma \rho_{\mathfrak{F}}(\hat{V}^{\hat{\pi}^*}) \delta,
\end{aligned} \tag{30}$$

where (f) is the same as (c); (g), (h) follow from (29), lemma 2 with $\hat{f} = \hat{V}^{\hat{\pi}^*}$ and the definition of an (ε, δ)-approximation for an MDP. Supremizing the LHS of (30) over all $\boldsymbol{s}, \boldsymbol{a} \in \mathrm{S} \times \mathrm{A}(m)$ and re-arranging terms, we get

$$\|Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*} \circ \phi\|_\infty \leq \frac{\varepsilon + \gamma \rho_{\mathfrak{F}}(\hat{V}^{\hat{\pi}^*}) \delta}{(1 - \gamma)} = \beta_{\mathfrak{F}}^{\hat{\pi}^*}. \tag{31}$$

This proves the second part of (23).

Finally, to show the first part of (25), consider

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

where (h) follows from the triangle inequality; (i) follows from (28) with $\hat{\pi} = \hat{\pi}^*$ and (31). To show the second part of (25), consider

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

where (j) follows from the triangle inequality; (k) follows from (26) with $\hat{\pi} = \hat{\pi}^*$ and (29); (l) follows from (28) with $\hat{\pi} = \hat{\pi}^*$ and (31). □

4.3 Proof of Theorem 1

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

$$\begin{aligned}\|Q^{\pi^*} - Q^{\hat{\mu}\circ\phi}\|_\infty &\leq \|Q^{\pi^*} - Q^{\hat{\pi}^*\circ\phi}\|_\infty + \|Q^{\hat{\pi}^*\circ\phi} - \hat{Q}^{\hat{\pi}^*}\circ\phi\|_\infty \\ &\quad + \|\hat{Q}^{\hat{\pi}^*} - \hat{Q}^{\hat{\mu}}\|_\infty + \|\hat{Q}^{\hat{\mu}}\circ\phi - Q^{\hat{\mu}\circ\phi}\|_\infty \\ &\overset{(a)}{\leq} 2\beta_{\mathfrak{F}}^{\hat{\pi}^*} + \beta_{\mathfrak{F}}^{\hat{\pi}^*} + \|\hat{Q}^{\hat{\pi}^*} - \hat{Q}^{\hat{\mu}}\|_\infty, + \beta_{\mathfrak{F}}^{\hat{\mu}},\end{aligned} \tag{32}$$

where each term of (a) is bound using Prop. 3. Rearranging terms proves (7).

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

$$\begin{aligned}\|V^{\pi^*} - V^{\hat{\mu}\circ\phi}\|_\infty &\leq \|V^{\pi^*} - V^{\hat{\pi}^*\circ\phi}\|_\infty + \|V^{\hat{\pi}^*\circ\phi} - \hat{V}^{\hat{\pi}^*}\circ\phi\|_\infty \\ &\quad + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty + \|\hat{V}^{\hat{\mu}} - V^{\hat{\mu}\circ\phi}\|_\infty \\ &\overset{(a)}{\leq} 2\beta_{\mathfrak{F}}^{\hat{\pi}^*} + \beta_{\mathfrak{F}}^{\hat{\pi}^*} + \|\hat{V}^{\hat{\pi}^*} - \hat{V}^{\hat{\mu}}\|_\infty + \beta_{\mathfrak{F}}^{\hat{\mu}},\end{aligned} \tag{33}$$

where each term of (b) is bound using Prop. 3. Rearranging the terms proves (8).

4.4 Proof of Proposition 1

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

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

where (a) follows from (Subramanian et al, 2022, Lemma 39) and (b) follows because span is a semi-norm (Puterman, 2014). Using the above bound in (7) and (8) and using Lemma 3 to bound (ε, δ), we get (9) and (10).

4.5 Proof of Proposition 2

Recall that the state space S^i of each arm is a metric space with metric d^i and $\{0,1\}^i$ is a metric space with metric $\hat{d}^i$. Define a metric $\boldsymbol{d}$ on S as follows: for any $q \in [1,\infty]$ and $\boldsymbol{s}, \boldsymbol{s}' \in \mathrm{S}$, $\boldsymbol{d}(\boldsymbol{s},\boldsymbol{s}') = \left(\sum_{i\in[n]} d^i(s^i, s'^i)^q\right)^{1/q}$. Define $\hat{\boldsymbol{d}}$ in an analogous manner.

We now define Lipschitz continuity for MDP $\mathcal{M}$.

Definition 8 Given MDP $\mathcal{M} = \langle \mathrm{S}, \mathrm{A}(m), \boldsymbol{p}, \boldsymbol{r} \rangle$, if

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

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

then the MDP $\mathcal{M}$ is said to be $(\mathrm{L}_{\boldsymbol{r}}, \mathrm{L}_{\boldsymbol{p}})$-Lipschitz.
Similarly, when this is true for the approximate MDP $\hat{\mathcal{M}} = \langle \hat{\mathrm{S}}, \mathrm{A}(m), \hat{\boldsymbol{p}}, \hat{\boldsymbol{r}} \rangle$, then it is said to be $(\mathrm{L}_{\hat{\boldsymbol{r}}}, \mathrm{L}_{\hat{\boldsymbol{p}}})$-Lipschitz.

Lemma 4 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 $\hat{\mathcal{M}} = \langle \hat{\mathrm{S}}, \mathrm{A}(m), \hat{\boldsymbol{p}}, \hat{\boldsymbol{r}} \rangle$ is $(\mathrm{L}^{(k)}_{\hat{\boldsymbol{r}}}, \mathrm{L}^{(k)}_{\hat{\boldsymbol{p}}})$-Lipschitz, where

$$\mathrm{L}^{(k)}_{\hat{\boldsymbol{r}}} \leq \left(\sum_{i \in [n]} (\mathrm{L}_{\hat{r}^i})^k \right)^{1/k}, \quad \mathrm{L}^{(k)}_{\hat{\boldsymbol{p}}} \leq \left(\sum_{i \in [n]} (\mathrm{L}_{\hat{p}^i})^k \right)^{1/k}. \tag{34}$$

Proof See Appendix C. □

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

$$\rho_{\mathfrak{F}}(\hat{V}^{\hat{\pi}^*}) = \mathrm{Lip}\,(\hat{V}^{\hat{\pi}^*}) \overset{(a)}{\leq} \frac{\mathrm{L}^{(k)}_{\hat{\boldsymbol{r}}}}{(1 - \gamma \mathrm{L}^{(k)}_{\hat{\boldsymbol{p}}})}. \tag{35}$$

where (a) follows from (Hinderer, 2005, Theorem 4.2). To prove Proposition 2, we will take $k = \infty$ because doing so gives the tighest possible bound in (35). Substititing (35) in (7) and (8) and using Lemma 3 to bound (ε, δ), we get (11) and (12).

5 Conclusions

In conclusion, we considered a restless multi-armed bandit problem with uncertain arm models and analyzed the sensitivity of heuristic policies such as the Whittle index policy. Our results use ideas from sensitivity analysis of MDPs, but bound the performance loss in terms of the model mismatch of each arm and the choice of metric used to compare transition matrices. Thus, our results show how to incorporate model uncertainty in existing heuristic solutions for restless bandits.

A preliminary results

We first prove some preliminary results.

Lemma 5 Consider any $f : \mathrm{S} \to \mathbb{R}$. Pick an arm $i \in [n]$ and arbitrarily fix $\boldsymbol{s}^{-i} \in \mathrm{S}^{-i}$. Define $f^i : \mathrm{S}^i \to \mathbb{R}$ by $f^i(s^i) = f(s^i, \boldsymbol{s}^{-i})$, for any $s^i \in \mathrm{S}^i$. Then

(a) $\mathrm{span}\,(f^i) \leq \mathrm{span}\,(f)$.
(b) $\mathrm{Lip}\,(f^i) \leq \mathrm{Lip}\,(f)$.

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

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

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

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

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

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 S^{-i} will given an upper bound to any specific $\boldsymbol{s}^{-i}$. □

For the ease of notation, when $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}} = \{f : \mathrm{S} \to \mathbb{R} : \frac{1}{2}\,\mathrm{span}\,(f) \leq 1\}$, define $\mathfrak{F}^i = \{f^i : \mathrm{S}^i \to \mathbb{R} : \frac{1}{2}\,\mathrm{span}\,(f^i) \leq 1\}$. Similarly when $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}} = \{f : \mathrm{S} \to \mathbb{R} :\ \mathrm{Lip}\,(f) \leq 1\}$, define $\mathfrak{F}^i = \{f^i : \mathrm{S}^i \to \mathbb{R} :\ \mathrm{Lip}\,(f^i) \leq 1\}$. Lemma 5 implies that if $f \in \mathfrak{F}$, for any $\boldsymbol{s}^{-i} \in \mathrm{S}^{-i}$, f^i (as defined in Lemma 5) belongs to $\mathfrak{F}^i$.

Lemma 6 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 $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$,

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

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

$$\begin{aligned}
&\left| \int_{\mathrm{S}} f d\boldsymbol{\mu} - \int_{\mathrm{S}} f d\boldsymbol{\nu} \right| \\
&= \left| \int_{\mathrm{S}^k} \int_{\mathrm{S}^{-k}} f(s^k, s^{-k}) \left[\mu^k(s^k) \mu^{-k}(s^{-k}) - \nu^k(s^k) \nu^{-k}(s^{-k}) \right] ds^k ds^{-k} \right| \\
&\overset{(a)}{\leq} \left| \int_{\mathrm{S}^k} \int_{\mathrm{S}^{-k}} f(s^k, s^{-k}) \left[\mu^k(s^k) \mu^{-k}(s^{-k}) - \mu^k(s^k) \nu^{-k}(s^{-k}) \right] ds^k ds^{-k} \right| \\
&\quad + \left| \int_{\mathrm{S}^k} \int_{\mathrm{S}^{-k}} f(s^k, s^{-k}) \left[\mu^k(s^k) \nu^{-k}(s^{-k}) - \nu^k(s^k) \nu^{-k}(s^{-k}) \right] ds^k ds^{-k} \right| \\
&\overset{(b)}{\leq} \int_{\mathrm{S}^k} \left| \int_{\mathrm{S}^{-k}} f(s^k, s^{-k}) \left[\mu^{-k}(s^{-k}) - \nu^{-k}(s^{-k}) \right] ds^{-k} \right| \mu^k(s^k) ds^k \\
&\quad + \int_{\mathrm{S}^{-k}} \left| \int_{\mathrm{S}^k} f(s^k, s^{-k}) \left[\mu^k(s^k) - \nu^k(s^k) \right] ds^k \right| \nu^{-k}(s^{-k}) ds^{-k}
\end{aligned} \tag{36}$$

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 5, $f(s^k, \cdot) \in \mathfrak{F}^{-k}$. Therefore,

$$\left| \int_{\mathrm{S}^{-k}} f(s^k, s^{-k}) \left[\mu^{-k}(s^{-k}) - \nu^{-k}(s^{-k}) \right] ds^{-k} \right| \leq d_{\mathfrak{F}^{-k}}(\mu^{-k}, \nu^{-k}) \tag{37}$$

and similarly,

$$\left| \int_{\mathrm{S}^k} f(s^k, s^{-k}) \left[\mu^k(s^k) - \nu^k(s^k) \right] ds^k \right| \leq d_{\mathfrak{F}^k}(\mu^k, \nu^k) \tag{38}$$

Substituting (37) and (38) in (36), we get

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

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

Lemma 7 Consider $\tilde{\boldsymbol{p}}$ as defined in Definition 7 and $\tilde{p}^i$ as defined in Definition 4 for arm α^i with state aggregation function ϕ^i. Then for all $\boldsymbol{s} \in \mathrm{S}$, $\boldsymbol{a} \in \mathrm{A}(m)$, any Borel subsets $\hat{B}^i \subset \hat{\mathrm{S}}^i$ and $\hat{\mathrm{B}} = \prod_{i\in[n]} \hat{B}^i$, we have

$$\tilde{\boldsymbol{p}}(\hat{\mathrm{B}} \mid \boldsymbol{s}, \boldsymbol{a}) = \prod_{i\in[n]} \tilde{p}^i(\hat{B}^i \mid s^i, a^i). \tag{39}$$

Proof Consider the term on the RHS, it can be re-written as follows

$$\prod_{i\in[n]} \tilde{p}^i(\hat{B}^i \mid s^i, a^i) = \prod_{i\in[n]} \int_{\bar{s}^i\in\mathrm{S}^i} \mathbb{1}\{\phi^i(\bar{s}^i) \in \hat{B}^i\} p^i(d\bar{s}^i \mid s^i, a^i)$$
$$= \int_{\bar{s}^i\in\mathrm{S}^i} \prod_{i\in[n]} \mathbb{1}\{\phi^i(\bar{s}^i) \in \hat{B}^i\} p^i(d\bar{s}^i \mid s^i, a^i)$$
$$\overset{(a)}{=} \int_{\bar{\boldsymbol{s}}\in\mathrm{S}} \mathbb{1}\{\phi(\bar{\boldsymbol{s}}) \in \hat{\mathrm{B}}\} \boldsymbol{p}(d\bar{\boldsymbol{s}} \mid \boldsymbol{s}, \boldsymbol{a})$$
$$= \tilde{\boldsymbol{p}}(\hat{\mathrm{B}} \mid \boldsymbol{s}, \boldsymbol{a}).$$

where (a) follows from the fact that the states for each arm evolve independently. □

B Proof of Lemma 3

For the first part, consider

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

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

Table 3 Hyperparameters used in computing $\text{Lip}\,(\hat{V}^{\hat{\mu}})$ approximately

Parameter	Value
Input size	$H = 100$
Number of linear layers	3
Hidden layer size	50
Activation function	ReLU
Number of episodes (N_e)	10^5
Samples per episode (S_e)	50
ADAM learning rate	0.001
Number of gradient steps	10^5
Samples for computing $\text{Lip}\,(\hat{V}^{\hat{\mu}})$, (N_L)	10^5

The second part follows from the definition of $\boldsymbol{p}$ (Eq. (16)), $\hat{\boldsymbol{p}}$ (Eq. (18)), $\tilde{\boldsymbol{p}}$ (Eq. (39)) and Lemma 6 applied with Lemma 1, Lemma 2.

C Proof of Lemma 4

For the first part, consider for any $\hat{\boldsymbol{s}}_{(1)}, \hat{\boldsymbol{s}}_{(2)} \in \hat{\mathrm{S}}, \boldsymbol{a} \in \mathrm{A}$

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

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 for any $\hat{\boldsymbol{s}}_{(1)}, \hat{\boldsymbol{s}}_{(2)} \in \hat{\mathrm{S}}, \boldsymbol{a} \in \mathrm{A}$

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

where (*d*) follows from Lemma 6, (*e*) follows from the assumption on the arms and (*f*) follows from Hölder's inequality and the definition of metric $\boldsymbol{d}$.

D Hyperparameters used in the example in Sect. 3.2.2

The parameters used in our experiment are described in Table 3.

We use a neural network with 3 hidden layers of size 50 each followed by ReLU activations. The input to the neural network is a vector of length $H = 100$ which contains the randomly initialized state $\hat{\boldsymbol{s}}$, where each component of the vector is a component $\hat{s}$ of $\hat{\boldsymbol{s}}$. The random initialization for these state components of each of the arms is done on the basis of the model dynamics given by ξ^i and δ^i which are fixed throughout the experiment.

A total of $N_e = 10^5$ initializations $\hat{\boldsymbol{s}}$ are used to construct the batch. For each of these initializations, we have S_e rollouts from the starting state to get a proper estimate of the average. Finally, we use this entire batch to do gradient descent using ADAM for 10^5 gradient steps.

Funding 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.

Declarations

Conflict of interest Amit Sinha declares that he has no Conflict of interest. Aditya Mahajan declares that he has no Conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Akbarzadeh, N., & Mahajan, A. (2019). Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *Conference on Decision and Control (CDC)*, IEEE (pp. 7294–7300).

Akbarzadeh, N., & Mahajan, A. (2022a). Conditions for indexability of restless bandits and an $\mathcal{O}(\|^{\ni})$ algorithm to compute Whittle index. *Journal of Applied Probability, 54*(4), 1164–1192.

Akbarzadeh, N., & Mahajan, A. (2022b). On learning whittle index policy for restless bandits with scalable regret. arXiv preprint arXiv:2202.03463.

Ansell, P. S., Glazebrook, K. D., Niño-Mora, J., & O'Keeffe, M. (2003). Whittle's index policy for a multiclass queueing system with convex holding costs. *Mathematical Methods of Operations Research, 57*(1), 21–39.

Asadi, K., Misra, D., & Littman, M. (2018). Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning* (vol. 80, pp. 264–273). PMLR https://proceedings.mlr.press/v80/asadi18a.html.

Avrachenkov, K. E., & Borkar, V. S. (2022). Whittle index based Q-learning for restless bandits with average reward. *Automatica, 139*(110), 186.

Ayesta, U., Erausquin, M., & Jacko, P. (2010). A modeling framework for optimizing the flow-level scheduling with time-varying channels. *Performance Evaluation, 67*(11), 1014–1029.

Ben-Israel, A., & Flåm, S. D. (1990). A bisection/successive approximation method for computing Gittins indices. *Mathematics for Operations Research, 34*(6), 411–422.

Bertsimas, D., & Niño-Mora, J. (2000). Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research, 48*(1), 80–90.

Borkar, V. S., & Chadha, K. (2018). A reinforcement learning algorithm for restless bandits. In *2018 Indian Control Conference (ICC)* (pp. 89–94).
Brown, D. B., & Smith, J. E. (2020). Index policies and performance bounds for dynamic selection problems. *Management Science, 66*(7), 3029–3050. https://doi.org/10.1287/mnsc.2019.3342
Caro, F., & Das Gupta, A. (2015). Robust control of the multi-armed bandit problem. *Annals of Operations Research,* 1–20.
Cohen, S. N., & Treetanthiploet, T. (2022). Gittins' theorem under uncertainty. *Electronic Journal of Probability, 27*, 1–48.
Fu, J., Nazarathy, Y., Moka, S., & Taylor, P. G. (2019). Towards Q-learning the Whittle index for restless bandits. In *2019 Australian New Zealand Control Conference (ANZCC)* (pp. 249–254).
Gast, N., Gaujal, B., & Yan, C. (2022). LP-based policies for restless bandits: Necessary and sufficient conditions for (exponentially fast) asymptotic optimality. arXiv:2106.10067.
Gast, N., Gaujal, B., & Khun, K. (2023). Testing indexability and computing whittle and gittins index in subcubic time. *Mathematical Methods of Operations Research, 97*(3), 391–436.
Gelada, C., Kumar, S., Buckman, J., Nachum, O., & Bellemare, M. G. (2019). DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*. PMLR (pp. 2170–2179). https://proceedings.mlr.press/v97/gelada19a.html.
Gittins, J. C., & Jones, D. M. (1979). A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika, 66*(3), 561–565.
Glazebrook, K. D., Mitchell, H. M., & Ansell, P. S. (2005). Index policies for the maintenance of a collection of machines by a set of repairmen. *European Journal of Operational Research, 165*(1), 267–284.
Glazebrook, K. D., Ruiz-Hernandez, D., & Kirkbride, C. (2006). Some indexable families of restless bandit problems. *Advances in Applied Probability, 38*(3), 643–672.
Hinderer, K. (2005). Lipschitz continuity of value functions in Markovian decision processes. *Mathematical Methods of Operations Research, 62*(1), 3–22.
Hodge, D. J., & Glazebrook, K. D. (2011). Dynamic resource allocation in a multi-product make-to-stock production system. *Queueing Systems, 67*(4), 333–364. https://doi.org/10.1007/s11134-011-9217-2
Hu, W., & Frazier, P. (2017). An asymptotically optimal index policy for finite-horizon restless bandits. https://doi.org/10.48550/arXiv.1707.00205, arXiv:1707.00205 [math].
Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research, 30*(2), 257–280.
Kara, A. D., & Yüksel, S. (2020). Robustness to incorrect system models in stochastic control. *SIAM Journal on Control and Optimization, 58*(2), 1144–1182. https://doi.org/10.1137/18m1208058
Katehakis, M. N., & Veinott, A. F., Jr. (1987). The multi-armed bandit problem: Decomposition and computation. *Mathematics of Operations Research, 12*(2), 262–268.
Killian, J. A., Perrault, A., & Tambe, M. (2021). Beyond "To act or not to act": Fast Lagrangian approaches to general multi-action restless bandits. In *International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 710–718).
Kim, M. J. (2016). Robust control of partially observable failing systems. *Operations Research, 64*(4), 999–1014.
Kim, M. J., & Lim, A. E. (2016). Robust multiarmed bandit problems. *Management Science, 62*(1), 264–285.
Lam, H. (2016). Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research, 41*(4), 1248–1275. https://doi.org/10.1287/moor.2015.0776
Liu, H., Liu, K., & Zhao, Q. (2012). Learning in a changing world: Restless multiarmed bandit with unknown dynamics. *IEEE Transactions on Information Theory, 59*(3), 1902–1916.
Lott, C., & Teneketzis, D. (2000). 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.
Meshram, R., Gopalan, A., & Manjunath, D. (2017). Restless bandits that hide their hand and recommendation systems. In *International Conference on Communication Systems and Networks (COMSNETS)*, IEEE (pp. 206–213).
Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability, 29*(2), 429–443.
Müller, A. (1997). How does the value function of a markov decision process depend on the transition probabilities? *Mathematics of Operations Research, 22*(4), 872–885. https://doi.org/10.1287/moor.22.4.872
Nilim, A., & El Ghaoui, L. (2005). Robust control of markov decision processes with uncertain transition matrices. *Operations Research, 53*(5), 780–798.
Niño-Mora, J. (2007). Dynamic priority allocation via restless bandit marginal productivity indices. *TOP, 15*(2), 161–198.
Papadimitriou, C. H., & Tsitsiklis, J. N. (1999). The complexity of optimal queuing network control. *Mathematics of Operations Research, 24*(2), 293–305. https://doi.org/10.1287/moor.24.2.293

Puterman, M. L. (2014). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons.

Robledo, F., Borkar, V., Ayesta, U., & Avrachenkov, K. (2022). QWI: Q-learning with Whittle index. *ACM SIGMETRICS Performance Evaluation Review, 49*(2), 47–50.

Schechter, E. (1996). *Handbook of Analysis and its Foundations*. Academic Press.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129–138. https://doi.org/10.1037/h0042769

Subramanian, J., Sinha, A., Seraj, R., & Mahajan, A. (2022). Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research, 23*(12), 1–83.

Tekin, C., & Liu, M. (2012). Online learning of rested and restless bandits. *IEEE Transactions on Information Theory, 58*(8), 5588–5611.

Tzortzis, I., Charalambous, C. D., & Charalambous, T. (2015). Dynamic programming subject to total variation distance ambiguity. *SIAM Journal on Control and Optimization, 53*(4), 2040–2075. https://doi.org/10.1137/140955707

Verloop, I. M. (2016). Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *The Annals of Applied Probability, 26*(4), 1947–1995. https://doi.org/10.1214/15-AAP1137

Weber, R. R., & Weiss, G. (1990). On an index policy for restless bandits. *Journal of Applied Probability, 27*(3), 637–648.

White, C. C., & Eldeib, H. K. (1994). Markov decision processes with imprecise transition probabilities. *Operations Research, 42*(4), 739–749. https://doi.org/10.1287/opre.42.4.739

Whitt, W. (1978). Approximations of dynamic programs, I. *Mathematics of Operations Research, 3*(3), 231–243. https://doi.org/10.1287/moor.3.3.231

Whitt, W. (1979). Approximations of dynamic programs, II. *Mathematics of Operations Research, 4*(2), 179–185. https://doi.org/10.1287/moor.4.2.179

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

Wiesemann, W., Kuhn, D., & Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research, 38*(1), 153–183. https://doi.org/10.1287/moor.1120.0566

Zayas-Cabán, G., Jasin, S., & Wang, G. (2019). An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits. *Advances in Applied Probability, 51*(3), 745–772. https://doi.org/10.1017/apr.2019.29