Performance Evaluation 163 (2024) 102394

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Two families of indexable partially observable restless bandits and Whittle index computation ☆

Nima Akbarzadeh *, Aditya Mahajan

Department of Electrical and Computer Engineering, McGill University, 3480 Rue University, Montréal, QC H3A 0E9, Canada

ARTICLE INFO

Keywords:
Restless bandits
Whittle index
Indexability
Partially observable
Scheduling
Resource allocation

ABSTRACT

We consider the restless bandits with general finite state space under partial observability with two observational models: first, the state of each bandit is not observable at all, and second, the state of each bandit is observable when it is selected. Under the assumption that the models satisfy a restart property, we prove that both models are indexable. For the first model, we derive a closed-form expression for the Whittle index. For the second model, we propose an efficient algorithm to compute the Whittle index by exploiting the qualitative properties of the optimal policy. We present detailed numerical experiments for multiple instances of machine maintenance problem. The result indicates that the Whittle index policy outperforms myopic policy and can be close to optimal in different setups.

1. Introduction

Resource allocation and scheduling problems arise in various applications including telecommunication networks, sensor management, patient prioritization, and machine maintenance. Restless bandits is a widely-used solution framework for such models [1–15].

Identifying the optimal policy in restless bandits suffers from the curse of dimensionality because the state space is exponential in the number of alternatives [16]. To circumvent the curse of dimensionality, Whittle proposed an index heuristic which has a linear complexity in the number of alternatives [17]. The resulting policy, called the Whittle index policy, operates as follows: assign an index (called the Whittle index) to each state of each arm (or alternative) and then, at each time, play the arms in states with the highest indices.

The Whittle index policy is attractive for two reasons. First, it is a scalable heuristic because its complexity is linear in the number of arms. Second, although it is a heuristic, there are certain settings where it is optimal [18–21] and, in general, it performs close to optimal in many instances [10,22–26].

Nonetheless, there are two challenges in using the Whittle index policy. First, the Whittle index heuristic is applicable only when a technical condition known as indexability is satisfied. There is no general test for indexability, and the existing sufficient conditions are only applicable for specific models [10,23–25,27–30]. Second, while there are closed-form expressions to compute the Whittle index in some instances [3–6,10,24,26,29,31], in general, the Whittle index policy has to be computed numerically. For a subclass of restless bandits which satisfy an additional technical condition known as PCL (partial conservation law), the Whittle index can be computed using an algorithm called the adaptive greedy algorithm [22,32]. Recently, [31,33] presented generalizations of adaptive greedy algorithm which are applicable to all indexable restless bandits.

☆ This research was funded in part by Fonds de Recherche du Quebec-Nature et technologies (FRQNT).

* Corresponding author.
E-mail addresses: nima.akbarzadeh@mail.mcgill.ca (N. Akbarzadeh), aditya.mahajan@mcgill.ca (A. Mahajan).

https://doi.org/10.1016/j.peva.2023.102394
Received 20 November 2022; Received in revised form 2 October 2023; Accepted 27 November 2023
Available online 2 December 2023

We are interested in resource allocation and scheduling problems where the state of each arm is not fully-observed. Such *partially observable* restless bandit models are conceptually and computationally more challenging. The sufficient conditions for indexability that are derived for fully-observed bandits [10,17,19,24,29,31,34] are not directly applicable to the partially observable setting. The existing literature on partially observable restless bandits often restricts attention to models where each arm has two states [1–5,9,11,13,35–37]. In some cases, it is also assumed that the two states are positively correlated [3–5]; in others, it is assumed that the state dynamics are independent of the chosen action [6,38,39]. There are very few results for general finite state space models under partial observability [6,7,12,38,39], and, for such models, indexability is often verified numerically. In addition, there are very few algorithms to compute the Whittle index for such models.

Recently, alternative index-based heuristics for partially observable restless bandits [40] has been proposed, but we restrict to Whittle index policy in this paper.

The main contributions of our paper are as follows:

- We investigate partially observable restless bandits with general finite state spaces and consider two observation models, which we call model A and model B. We show that both models are indexable.
- For model A, we provide a closed-form expression to compute the Whittle index. For model B, we provide a refinement of the adaptive greedy algorithm of [31] to efficiently compute the Whittle index.
- We present a detailed numerical study which illustrates that the Whittle index policy performs close to optimal for small scale systems and outperforms a commonly used heuristic (the myopic policy) for large-scale systems.

The organization of the paper is as follows. In Section 2, we formulate the restless bandit problem under partial observations for two different models. Then, we define a belief state by which the partially-observable problem can be converted into a fully-observable one. In Section 3, we present a short overview of restless bandits. In Section 4, we show the restless bandit problem is indexable for both models and present a general formula to compute the index. In Section 5, we present a countable state representation of the belief state and use it to develop methods to compute the Whittle index. In Section 6, we present the proofs of the results. In Section 7, we present a detailed numerical study which compares the performance of Whittle index policy with two baseline policies. Finally, we conclude in Section 8.

1.1. Notations and definitions

We use uppercase letters to denote random variables; the corresponding lowercase letter to denote their realization and corresponding calligraphic letters to denote the set of realizations. Superscripts index arms and subscripts index time, e.g., X_t^i denotes the state of arm i at time t. The subscript $0{:}t$ denote the history of the variable from time 0 to t, e.g., $X_{0:t}^i$ denotes $(X_0, \dots, X_t)$. Bold letters denote the vector of variable for all arms, e.g., $\boldsymbol{X}_t$ denotes $(X_t^1, \dots, X_t^n)$. Given a collection of real numbers $p^1, \dots, p^n$, we use $\prod_{i=1}^n p^i$ to denote their product $p^1 \cdot p^2 \cdot \cdots \cdot p^n$. Given a collection of sets $\mathcal{X}^1, \dots, \mathcal{X}^n$, we use $\prod_{i=1}^n \mathcal{X}^i$ to denote their Cartesian product $\mathcal{X}^1 \times \mathcal{X}^2 \times \cdots \times \mathcal{X}^n$. For a finite set $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$ denote the set of probability mass functions (PMFs) on $\mathcal{X}$.

We use $\mathbb{I}$ as the indicator function, $\mathbb{E}$ as the expectation operator, $\mathbb{P}$ as the probability function, $\mathbb{R}$ as the set of real numbers, $\mathbb{Z}$ as the set of integers and $\mathbb{Z}_{\geq m}$ as the set of integers that are not lower than m.

Given ordered sets $\mathcal{X}$ and $\mathcal{Y}$, a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is called submodular if for any $x_1, x_2 \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$ such that $x_2 \geq x_1$ and $y_2 \geq y_1$, we have $f(x_1, y_2) - f(x_1, y_1) \geq f(x_2, y_2) - f(x_2, y_1)$. Furthermore, the transition probability matrix P is stochastic monotone if for any $x, y \in \mathcal{X}$ such that $x < y$, we have $\sum_{w \in \mathcal{X}_{\geq z}} P_{xw} \leq \sum_{w \in \mathcal{X}_{\geq z}} P_{yw}$ for any $z \in \mathcal{X}$. For any function $f : \mathcal{Z} \to \mathbb{R}$, $\operatorname{span}(f)$ denotes the span semi-norm of f, i.e., $\operatorname{span}(f) = \sup_{z \in \mathcal{Z}} f(z) - \inf_{z \in \mathcal{Z}} f(z)$.

2. Model and problem formulation

2.1. Restless bandit process with restart

A discrete-time restless bandit process (or arm) is a controlled Markov process $(\mathcal{X}, \{0,1\}, \{\bar{P}(a)\}_{a \in \{0,1\}}, c, \pi_0, \mathcal{Y})$ where $\mathcal{X}$ denotes the finite set of states; $\{0,1\}$ denotes the action space where the action 0 is called the *passive* action and the action 1 is the *active* action; $\bar{P}(a)$, $a \in \{0,1\}$, denotes the transition matrix when action a is chosen; $c : \mathcal{X} \times \{0,1\} \to \mathbb{R}_{\geq 0}$ denotes the cost function; π_0 denotes the initial state distribution; $\mathcal{Y}$ denotes the finite set of observations.

Assumption 1 (*Restart Property*)**.** All rows of the transition matrix $\bar{P}(1)$ are identical.

For models which satisfy Assumption 1, we denote $\bar{P}(0)$ by P and denote each (identical) row of $\bar{P}(1)$ by Q. The term *restart property* is used following the terminology of [26], where Q was a PMF on the state space (i.e., on taking active action, the state resets according to PMF Q). Note [26] considered fully observed models, while we are considering partially observable setups. Partially observable restless bandits with restart property have been considered in [11,35–37] but these papers restricted attention to models with binary state space, while we are considering general finite state spaces.

An operator has to select $m < n$ arms at each time but does not observe the state of the arms. We consider two observation models.

- **Model A**: In model A, the operator does not observe anything. We denote this by $Y_t = \mathfrak{E}$, where $\mathfrak{E}$ denotes a blank symbol.
- **Model B**: In model B, the operator observes the state of the arm after it has been reset, i.e.,

$$Y_{t+1} = \begin{cases} \mathfrak{E} & \text{if } A_t = 0 \\ X_{t+1} & \text{if } A_t = 1, \end{cases} \tag{1}$$

For model A, $\mathcal{Y} = \{\mathfrak{E}\}$ and for model B, $\mathcal{Y} = \mathcal{X} \cup \{\mathfrak{E}\}$.

2.2. Partially-observable restless multi-armed bandit problem

A partially-observable restless multi-armed bandit (PO-RMAB) problem is a collection of n independent restless bandits $(\mathcal{X}^i, \{0,1\}, \{P^i(a)\}_{a\in\{0,1\}}, c^i, \pi_0^i, \mathcal{Y}^i)$, $i \in \mathcal{N} := \{1, \dots, n\}$.

Let $\boldsymbol{\mathcal{X}} := \prod_{i\in\mathcal{N}} \mathcal{X}^i$, $\boldsymbol{\mathcal{A}}(m) := \left\{ (a^1, \dots, a^n) \in \{0,1\}^n : \sum_{i\in\mathcal{N}} a^i = m \right\}$, and $\boldsymbol{\mathcal{Y}} := \prod_{i\in\mathcal{N}} \mathcal{Y}^i$ denote the combined state, action, and observation spaces, respectively. Also, let $\boldsymbol{X}_t = (X_t^1, \dots X_t^n) \in \boldsymbol{\mathcal{X}}$, $\boldsymbol{A}_t = (A_t^1, \dots, A_t^n) \in \boldsymbol{\mathcal{A}}(m)$, and $\boldsymbol{Y}_t = (Y_t^1, \dots Y_t^n) \in \boldsymbol{\mathcal{Y}}$ denote the combined states, actions taken, and observations made by the operator at time $t \geq 0$. Due to the independent evolution of each arm, for each realization $\boldsymbol{x}_{0:t}$ of $\boldsymbol{X}_{0:t}$ and $\boldsymbol{a}_{0:t}$ of $\boldsymbol{A}_{0:t}$, we have

$$\begin{aligned} \mathbb{P}(\boldsymbol{X}_{t+1} = \boldsymbol{x}_{t+1} | \boldsymbol{X}_{0:t} = \boldsymbol{x}_{0:t}, \boldsymbol{A}_{0:t} = \boldsymbol{a}_{0:t}) &= \prod_{i\in\mathcal{N}} \mathbb{P}(X_{t+1}^i = x_{t+1}^i | X_t^i = x_t^i, A_t^i = a_t^i) \\ &= \prod_{i\in\mathcal{N}} P^i_{x_t^i, x_{t+1}^i}(a_t^i). \end{aligned}$$

Let $\boldsymbol{\pi}_0 = \prod_{i\in\mathcal{N}} \pi_0^i$ denote the initial state distribution of all arms.

When the system is in state $\boldsymbol{x}_t$ and action $\boldsymbol{a}_t$ is taken, the system incurs a cost $\boldsymbol{c}(\boldsymbol{x}_t, \boldsymbol{a}_t) := \sum_{i\in\mathcal{N}} c^i(x_t^i, a_t^i)$. The action at time t is chosen according to

$$\boldsymbol{A}_t = \boldsymbol{g}_t(\boldsymbol{Y}_{0:t-1}, \boldsymbol{A}_{0:t-1}), \tag{2}$$

where $\boldsymbol{g}_t$ is a history-dependent policy at time t. Let $\boldsymbol{g} = (g_1, g_2, \dots)$ denote the policy for infinite time horizon and let $\boldsymbol{\mathcal{G}}$ denote the family of all such policies. Then, the performance of policy $\boldsymbol{g}$ is given by

$$J^{(\boldsymbol{g})} := (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \sum_{i\in\mathcal{N}} c^i(X_t^i, A_t^i) \Big| X_0^i \sim \pi_0^i \right], \tag{3}$$

where $\beta \in (0,1)$ denotes the discount factor.

Formally, the optimization problem of interest is as follows:

Problem 1. Given a discount factor $\beta \in (0,1)$, the total number n of arms, the number m to be selected, the system model $\{(\mathcal{X}^i, \{0,1\}, P^i(a), c^i, \mathcal{Y}^i)\}_{i\in\mathcal{N}}$ of each arm, and the observation model at the operator, choose a history-dependent policy $\boldsymbol{g} \in \boldsymbol{\mathcal{G}}$ that minimizes $J^{(\boldsymbol{g})}$ given by (3).

Some remarks

1. Problem 1 is a POMDP and the standard methodology to solve POMDPs is to convert them to a fully observable Markov decision process (MDP) by viewing the "belief state" as the information state of the system [41]. We present such a belief state representation in Section 2.4 and point out its limitations in the context of restless bandits.
2. In Problem 1, the objective $J^{(\boldsymbol{g})}$ depends on the initial state distribution $(\pi_0^i)_{i\in\mathcal{N}}$. This can give the impression that the optimal policy may depend on the initial distribution. It is well known in the MDP literature that there exist policies that are optimal for all initial distributions [42]. However, our results rely on translating the belief state representation of the POMDP into a countable state MDP formulation and such a transformation is valid only when the initial state distribution is of a specific form. See Section 5 for details. Our results do not depend on the specific choice of the initial distribution, as long as it satisfies Assumption 2 specified in Section 5.
3. In Problem 1, we consider the *normalized* expected discounted cost as an objective, where the discounted cost is multiplied by $(1-\beta)$. In the MDP literature, one typically considers unnormalized objectives. However, normalized objective is typically used constrained MDPs [43] and has also been used in some of the previous literature on restless bandits [31]. Note that multiplying the objective by a constant does not change the optimal policy. The reason that we use a normalized expected discounted cost is that it simplifies the description of the adaptive greedy algorithms to compute the Whittle index presented in Section 5.

2.3. Some examples

In this section, we present some examples corresponding to the model presented above.

Example 1. Consider a sensor network where there are n sensors, each observing an independent Markov processes. We assume that the state $\{S_t^i\}_{t\geq 1}$ of each Markov process is integer valued and evolves in an auto-regressive manner: $S_{t+1}^i = S_t^i + W_t^i$, where $\{W_t^i\}_{t\geq 1}$ are i.i.d. processes which are also independent across the sensors. An estimator can observe only m, where $m < n$, sensors at each time. If a sensor is observed, the state of the Markov process at that sensor is revealed to the estimator. If a sensor is not observed, the estimator gets no new information about its state and has to estimate the state based on previous observations. The objective is to determine a sensor scheduling policy to decide which sensors to observe at each time.

In this case, it can be shown that when the noise processes $\{W_t^i\}_{t\geq 1}$ have symmetric and unimodal distributions, the optimal estimation strategy is a *Kalman-filter* like strategy, i.e., the optimal estimate $\hat{S}_t^i$ is S_t^i when the Markov process i is transmitted, and is equal to the previous estimate $\hat{S}_{t-1}^i$ when the Markov process i is not transmitted [44]. Thus, the error process $E_t^i := S_t^i - \hat{S}_t^i$ has a restart property [45]. An instance of such a sensor network problem was considered in [46].

Example 2. Consider a maintenance company monitoring n machines which are deteriorating independently over time. Each machine has multiple deterioration states sorted from *pristine* to *ruined* levels. However, the state of the machine is not observed. There is a cost associated with running the machine and the cost is non-decreasing function of the state. If a machine is left unmonitored, then the state of the machine deteriorates and after a while, it is ruined. However, the state of the machine is not observed.

Furthermore, it is assumed the company cannot observe the state of the machines unless it sends a service-person to visit the machine. Replacing the machine is relatively inexpensive, and when service-persons visit a machine, they simply replace it with a new one. Due to manufacturing mistakes, all the machines may not be in pristine state when installed. If the service-person can observe the state of the machine when installing a new one, the observation model is same as model B. Otherwise, it is model A. There are m, where $m < n$, service-persons. The objective is to determine a scheduling policy to decide which machines should be serviced at each time. An instance of such machine maintenance problem in the context of maintaining demand response devices was considered in [9].

Example 3. Consider the problem of resource constrained health intervention delivery, where a community health center is monitoring n patients to check if they are adhering to the prescribed medication [37]. Each patient has a binary state of "Adhering" or "Not Adhering", which is hidden. There are m, where $m < n$, health workers, and if an health worker visits a patient, the state of the patient is observed. Moreover, it is assumed that after the visit by a health worker, the patient goes into the "Adhering" state. The objective is to determine a policy to schedule the health workers to maximize the number of patients in the "Adhering" state.

2.4. Belief state

Using standard results from Markov decision theory, the partially observable restless bandit problem can be converted to a fully observable restless bandit problem with belief (or posterior distribution) as states. We present the details in this section. Let define the operator's belief $\Pi_t^i \in \mathcal{P}(\mathcal{X}^i)$ on the state of arm i at time t as follows: for any, $x_t^i \in \mathcal{X}^i$, let $\Pi_t^i(x_t^i) := \mathbb{P}(X_t^i = x_t^i \mid Y_{0:t-1}^i, A_{0:t-1}^i)$. Note that Π_t^i is a distribution-valued random variable. Also, define $\boldsymbol{\Pi}_t := (\Pi_t^1, \dots, \Pi_t^n)$.

Then, for arm i, the evolution of the belief state is as follows: for model A, the belief update rule is

$$\Pi_{t+1}^i = \begin{cases} \Pi_t^i P, & \text{if } A_t^i = 0, \\ Q, & \text{if } A_t^i = 1, \end{cases} \tag{4}$$

and for model B, the belief update rule is

$$\Pi_{t+1}^i = \begin{cases} \Pi_t^i P, & \text{if } A_t^i = 0, \\ \delta_{X_{t+1}^i} \text{ where } X_{t+1}^i \sim Q, & \text{if } A_t^i = 1 \end{cases} \tag{5}$$

where δ_x is the Dirac delta distribution over the discrete state space $\mathcal{X}$ with the value of one only for state x. Fig 1 is an illustation of the dynamics for this model. The per-step cost function of the belief state Π_t^i when action A_t^i is taken is

$$\bar{c}(\Pi_t^i, A_t^i) = \mathbb{E}[c_t^i(X_t^i, A_t^i) | Y_{0:t-1}^i, A_{0:t-1}^i] = \sum_{x\in\mathcal{X}^i} \Pi_t^i(x) c^i(x, A_t^i).$$

Define the combined belief state $\Theta_t \in \mathcal{P}(\boldsymbol{\mathcal{X}})$ of the system as follows: for any $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,

$$\Theta_t(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X}_t = \boldsymbol{x} \mid \boldsymbol{Y}_{0:t-1}, \boldsymbol{A}_{0:t-1}).$$

Note that Θ_t is a random variable that takes values in $\mathcal{P}(\boldsymbol{\mathcal{X}})$. Using standard results in POMDPs [41], we have the following.

Proposition 1. *In Problem 1, Θ_t is a sufficient statistic for $(\boldsymbol{Y}_{0:t-1}, \boldsymbol{A}_{0:t-1})$. Therefore, there is no loss of optimality in restricting attention to decision policies of the form $\boldsymbol{A}_t = g_t^{belief}(\Theta_t)$. Furthermore, an optimal policy with this structure can be identified by solving an appropriate dynamic program.*

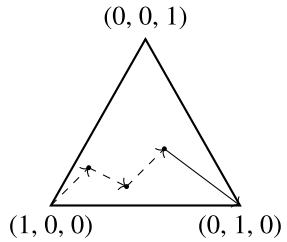

Fig. 1. Belief state dynamics for a 3-state arm i in the simplex $\mathcal{P}(\{1,2,3\})$ for model B. Dashed arrows show a sample realizations of the belief state evolution under $A_t = 0$ for three time steps and the solid arrow shows a sample realization of the belief state evolution under $A_t = 1$.

Next, we present our first simplification for the structure of optimal decision policy as follows.

Proposition 2. *For any $\boldsymbol{x} \in \mathcal{X}$, we have*

$$\Theta_t(\boldsymbol{x}) = \prod_{i \in \mathcal{N}} \Pi_t^i(x^i), \quad a.s.. \tag{6}$$

Therefore, there is no loss of optimality in restricting attention to decision policies of the form $\boldsymbol{A}_t = g_t^{simple}(\boldsymbol{\Pi}_t)$. Furthermore, an optimal policy with this structure can be identified by solving an appropriate dynamic program.

Proof. Eq. (6) follows from the conditional independence of the arms, and the nature of the observation function. The structure of the optimal policies then follow immediately from Proposition 1.

In Propositions 1 and 2, we do not present the dynamic programs because they suffer from the curse of dimensionality. In particular, obtaining the optimal policy for PO-RMAB is PSPACE-hard [16]. So, we focus on the Whittle index heuristics to solve the problem.

3. Whittle index policy solution concept

For the ease of notation, we will drop the superscript i from all relative variables for the rest of this and the next sections. Consider an arm $(\mathcal{X}, \{0,1\}, \{\bar{P}(a)\}_{a\in\{0,1\}}, c, \pi_0, \mathcal{Y})$ with a modified per-step cost function

$$\bar{c}_\lambda(\pi, a) := \bar{c}(\pi, a) + \lambda a, \quad \forall \pi \in \mathcal{P}(\mathcal{X}), \forall a \in \{0,1\}, \lambda \in \mathbb{R}. \tag{7}$$

The modified cost function implies that there is a penalty of λ for taking the active action. Given any time-homogeneous policy $g : \mathcal{P}(\mathcal{X}) \to \{0,1\}$, the modified performance of the policy is

$$J_\lambda^{(g)} := (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(\Pi_t, g(\Pi_t)) \middle| \Pi_0\right]. \tag{8}$$

Subsequently, consider the following optimization problem.

Problem 2. Given an arm $(\mathcal{X}, \mathcal{Y}, \{0,1\}, \{\bar{P}(a)\}_{a\in\{0,1\}}, c)$, the discount factor $\beta \in (0,1)$ and the penalty $\lambda \in \mathbb{R}$, choose a Markov policy $g : \mathcal{P}(\mathcal{X}) \to \{0,1\}$ to minimize $J_\lambda^{(g)}$ given by (8).

Problem 2 is a Markov decision process where one may use dynamic programming to obtain the optimal solution as follows.

Proposition 3. *Consider the fixed point equation*

$$V_\lambda(\pi) = \min_{a\in\{0,1\}} H_\lambda(\pi, a) \tag{9}$$

where for Model A we have

$$H_\lambda(\pi,0) = (1-\beta)\bar{c}(\pi,0) + \beta V_\lambda(\pi P), H_\lambda(\pi,1) = (1-\beta)\bar{c}(\pi,1) + (1-\beta)\lambda + \beta V_\lambda(Q)$$

and for Model B, we have

$$H_\lambda(\pi,0) = (1-\beta)\bar{c}(\pi,0) + \beta V_\lambda(\pi P), H_\lambda(\pi,1) = (1-\beta)\bar{c}(\pi,1) + (1-\beta)\lambda + \beta \sum_{x\in\mathcal{X}} Q_x V_\lambda(\delta_x).$$

Then (9) has a unique fixed point V_λ^*, and the policy

$$g_\lambda(\pi) = \begin{cases} 0, & \text{if } H_\lambda(\pi,0) < H_\lambda(\pi,1) \\ 1, & \text{otherwise} \end{cases}$$

is optimal for Problem 2.

Proof. The result follows immediately from Markov decision theory [42].

Finally, we present the following definitions.

Definition 1 (*Passive Set*)**.** Given penalty λ, define the passive set $\mathcal{W}_\lambda$ as the set of states where passive action is optimal for the modified arm, i.e.,

$$\mathcal{W}_\lambda := \left\{\pi \in \Pi : g_\lambda(\pi) = 0\right\}.$$

Definition 2 (*Indexability*)**.** An arm is indexable if $\mathcal{W}_\lambda$ is non-decreasing in λ, i.e., for any $\lambda_1, \lambda_2 \in \mathbb{R}$,

$$\lambda_1 < \lambda_2 \implies \mathcal{W}_{\lambda_1} \subseteq \mathcal{W}_{\lambda_2}.$$

A restless multi-armed bandit problem is indexable if all n arms are indexable.

Definition 3 (*Whittle Index*)**.** The Whittle index of the state π of an arm is the smallest value of λ for which state π is part of the passive set $\mathcal{W}_\lambda$, i.e.,

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

Equivalently, the Whittle index $w(\pi)$ is the smallest value of λ for which the optimal policy is indifferent between the active action and passive action when the belief state of the arm is π.

The Whittle index policy is as follows: *At each time step, select m arms which are in states with the highest indices*. The Whittle index policy is easy to implement and efficient to compute but it may not be optimal. As mentioned earlier, Whittle index is optimal in certain cases [18–21] and performs close to optimal for many other cases [10,22–26].

4. Indexability and the corresponding Whittle index for models A and B

Given an arm, let Σ denote the family of all stopping times $\tau \geq 1$, with respect to the natural filtration associated with $\{\Pi_t\}_{t\geq 0}$. For any stopping time $\tau \in \Sigma$ and an initial belief state $\pi \in \Pi$, define

$$L(\pi, \tau) := \mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t \bar{c}(\Pi_t, 0) + \beta^\tau \bar{c}(\Pi_\tau, 1) \;\middle|\; \Pi_0 = \pi\right],$$
$$B(\pi, \tau) := \mathbb{E}[\beta^\tau | \Pi_0 = \pi].$$

Theorem 1. *The PO-RMAB for model A and B is indexable. In particular, each arm is indexable and the Whittle index is given by*

$$w(\pi) = \inf\left\{\lambda \in \mathbb{R} : G(\pi) < \Omega_\lambda\right\},$$

where

$$G(\pi) := (1-\beta) \inf_{\tau \in \Sigma} \frac{L(\pi, \tau) - \bar{c}(\pi, 1)}{1 - B(\pi, \tau)}, \tag{10}$$

$$\Omega_\lambda := \lambda + \beta V_1^{\text{NEXT}}, \tag{11}$$

and $V_1^{\text{NEXT}} = V_\lambda(Q)$ for model A and $V_1^{\text{NEXT}} = \sum_{x\in\mathcal{X}} Q_x V_\lambda(\delta_x)$ for model B.

Proof. Recall that we assert that $V_\lambda(\pi)$ and Ω_λ are non-decreasing in λ for any $\pi \in \Pi$. Hence, for any policy $g : \mathcal{P}(\mathcal{X}) \to \{0, 1\}$

$$V_\lambda^{(g)}(\pi) = (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(\Pi_t, A_t) \middle| \Pi_t = \pi\right]$$

where $\bar{c}_\lambda(\pi, a) = c_\lambda(\pi, a) + \lambda a$ is non-decreasing in λ for any $\pi \in \mathcal{P}(\mathcal{X})$ and $a \in \{0, 1\}$. From Markov decision theory we know that $V_\lambda(\pi) = \inf_{g:\mathcal{P}(\mathcal{X})\to\{0,1\}} V_\lambda^{(g)}(\pi)$. Since the infimum of non-decreasing functions is non-decreasing, $V_\lambda(\pi)$ is non-decreasing in λ for any $\pi \in \mathcal{P}(\mathcal{X})$. Consequently, V_1^{NEXT} is non-decreasing which implies Ω_λ is non-decreasing in λ.

Given any stopping time $\tau \in \Sigma$, let h_τ denote a policy that takes the passive action up to and including time $\tau - 1$, takes the active action at time τ, and follows the optimal policy from time $\tau + 1$ onwards. The performance $C_\lambda(\pi, \tau)$ of policy h_τ is given by

$$C_\lambda(\pi, \tau) = (1-\beta)\mathbb{E}^{h_\tau}\left[\sum_{t=0}^{\infty} \beta^t c_\lambda(\pi_t, A_t) \;\middle|\; \Pi_0 = \pi\right] = (1-\beta)L(\pi, \tau) + \mathbb{E}[\beta^\tau \Omega_\lambda | \Pi_0 = \pi]$$
$$= (1-\beta)L(\pi, \tau) + B(\pi, \tau)\Omega_\lambda. \tag{12}$$

We use h_0 to denote a policy that takes active action at time 0 and follows the optimal policy from time 1 onwards. The performance $C_\lambda(\pi, 0)$ of policy h_0 is given by

$$C_\lambda(\pi, 0) = (1-\beta)c(\pi, 1) + \Omega_\lambda. \tag{13}$$

The next result generalizes [26, Lemma 2].

Lemma 1. *The following characterizations of the passive sets are equivalent to Definition 1.*

1. $\mathcal{W}_\lambda = \{\pi \in \Pi : H_\lambda(\pi, 0) < H_\lambda(\pi, 1)\}$.
2. $\mathcal{W}_\lambda = \{\pi \in \Pi : \exists \sigma \in \Sigma \text{ such that } C_\lambda(\pi, \sigma) < C_\lambda(\pi, 0)\}$.
3. $\mathcal{W}_\lambda = \{\pi \in \Pi : G(\pi) < \Omega_\lambda\}$.

Proof. Characterization (1) follows from the dynamic program given in Proposition 3. Characterization (2) follows from the fact that $C_\lambda(\pi, 0) = H_\lambda(\pi, 1)$ and for $\pi \in \mathcal{W}_\lambda$, $C_\lambda(\pi, \sigma) = H_\lambda(\pi, 0)$, where σ is the hitting time of the set $\mathcal{P}(\mathcal{X}) \setminus \mathcal{W}_\lambda$. Characterization (3) follows from characterization (2) and rearranging the terms using (12) and (13).

Note that $G(\pi)$ does not depend on λ while we showed that Ω_λ is non-decreasing in λ. Hence, $\mathcal{W}_\lambda = \{\pi \in \Pi : G(\pi) < \Omega_\lambda\}$ is non-decreasing in λ by the lemma. Thus, arm i is indexable. The expression for the Whittle index in the theorem follows from the definitions.

5. Whittle index computation

Computing the Whittle index using the belief state representation is intractable in general. Inspired by the approach taken in [47], we introduce a new information state which is equivalent to the belief state.

5.1. Countable information state

For models A and B, define $\mathcal{R}_A = \{QP^k : k \in \mathbb{Z}_{\geq 0}\}$, $\mathcal{R}_B = \{\delta_s P^k : s \in \mathcal{X}, k \in \mathbb{Z}_{\geq 0}\}$.

Assumption 2. For model A, $\pi_0 \in \mathcal{R}_A$ and for model B, $\pi_0 \in \mathcal{R}_B$.

For model A, define a process $\{K_t\}_{t\geq 0}$ as follows. The initial state k_0 is such that $\pi_0 = QP^{k_0}$ and for $t > 0$, K_t is given by

$$K_t = \begin{cases} 0, & \text{if } A_{t-1} = 1 \\ K_{t-1} + 1, & \text{if } A_{t-1} = 0. \end{cases} \tag{14}$$

Similarly, for model B, define a process $\{S_t, K_t\}_{t\geq 0}$ as follows. The initial state (s_0, k_0) is such that $\pi_0 = \delta_{s_0} P^{k_0}$ and for $t > 0$, K_t evolves according to (14) and S_t evolves according to

$$S_t = \begin{cases} X_{t-1} \text{ where } X_{t-1} \sim Q, & \text{if } A_{t-1} = 1 \\ S_{t-1}, & \text{if } A_{t-1} = 0. \end{cases} \tag{15}$$

Note that once the first observation has been taken in both models, K_t denotes the time elapsed since the last observation of arm i and, in addition in model B, S_t denotes the last observed states of arm i. Let $\boldsymbol{S}_t := (S_t^1, \dots S_t^n)$ and $\boldsymbol{K}_t := (K_t^1, \dots K_t^n)$. The relation between the belief state Π_t and variables S_t and K_t is characterized in the following lemma (see Fig. 1). .

Lemma 2. *The following statements hold under Assumption 2:*

- *For model A, for any t, $\Pi_t \in \mathcal{R}_A$. In particular, $\Pi_t = QP^{K_t}$.*
- *For model B, for any t, $\Pi_t \in \mathcal{R}_B$. In particular, $\Pi_t = \delta_{S_t} P^{K_t}$.*

Proof. The results immediately follow from (4)–(5) and (14)–(15).

For model A, the expected per-step cost at time t may be written as

$$\bar{c}(K_t, A_t) := \bar{c}(QP^{K_t}, A_t) = \sum_{x\in\mathcal{X}} [QP^{K_t}]_x c(x, A_t). \tag{16}$$

Similarly, for model B, the expected per-step cost at time t may be written as

$$\bar{c}(S_t, K_t, A_t) := \bar{c}(\delta_{S_t} P^{K_t}, A_t) = \sum_{x\in\mathcal{X}} [\delta_{S_t}, P^{K_t}]_x c(x, A_t). \tag{17}$$

Proposition 4. *In Problem 1, there is no loss of optimality in restricting attention to decision policies of the form $\boldsymbol{A}_t = g_t^{info}(\boldsymbol{K}_t)$ for model A and of the form $\boldsymbol{A}_t = g_t^{info}(\boldsymbol{S}_t, \boldsymbol{K}_t)$ for model B.*

Proof. This result immediately follows from Lemma 2, (16) and (17).

5.2. Threshold policies

We assume that the model satisfies the following condition.

Assumption 3. Let $c(x,a) = (1-a)\phi(x) + a\rho(x)$ where $\phi : \mathcal{X} \to [0, \phi_{\max})$ and $\rho : \mathcal{X} \to [0, \rho_{\max})$ are non-decreasing functions in $\mathcal{X}$ and $c(x,a)$ is submodular in (x,a).

Under Assumption 3, we derive structural properties of the optimal policies for models A and B. Then, we show how $J_\lambda^{(g)}$ can be decomposed and computed.

In the following theorem, we show that the optimal policy for model A has a threshold structure and the optimal policy for model B has a threshold structure with respect to the second dimension of the information state.

Theorem 2. *Under Assumptions 2 and 3, the following statements hold:*

1. *In model A, for any $\lambda \in \mathbb{R}$, the optimal policy $g_\lambda^A(k)$ is a threshold policy, i.e., there exists a threshold $\theta_\lambda^A \in \mathbb{Z}_{\geq -1}$ such that*

$$g_\lambda^A(k) = \begin{cases} 0, & k < \theta_\lambda^A \\ 1, & \text{otherwise.} \end{cases}$$

   Moreover, the threshold θ_λ^A is non-decreasing in λ.

2. *In model B, for any $\lambda \in \mathbb{R}$, the optimal policy $g_\lambda^B(s,k)$ is a threshold policy with respect to k for every $s \in \mathcal{X}$, i.e., there exists a threshold $\theta_{s,\lambda}^B \in \mathbb{Z}_{\geq -1}$ for each $s \in \mathcal{X}$ such that*

$$g_\lambda^B(s,k) = \begin{cases} 0, & k < \theta_{s,\lambda}^B \\ 1, & \text{otherwise.} \end{cases}$$

   Moreover, for every $s \in \mathcal{X}$, the threshold $\theta_{s,\lambda}^B$ is non-decreasing in λ.

See Section 6 for the proof.

We use $\boldsymbol{\theta}^B$ to denote the vector $(\theta_s^B)_{s\in\mathcal{X}}$.

5.3. Performance of threshold based policies

We simplify the notation and denote the policy corresponding to thresholds θ^A and $\boldsymbol{\theta}^B$ by simply θ^A and $\boldsymbol{\theta}^B$ instead of $g^{(\theta^A)}$ and $g^{(\boldsymbol{\theta}^B)}$.

5.3.1. Model A

Let $J_\lambda^{(\theta^A)}(k)$ be the total discounted cost incurred under policy $g^{(\theta^A)}$ with penalty λ when the initial state is k, i.e.,

$$J_\lambda^{(\theta^A)}(k) := (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(K_t, g^{(\theta^A)}(K_t)) \;\middle|\; K_0 = k\right] := D^{(\theta^A)}(k) + \lambda N^{(\theta^A)}(k), \tag{18}$$

where

$$D^{(\theta^A)}(k) := (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t c(K_t, g^{(\theta^A)}(K_t)) \;\middle|\; K_0 = k\right],$$

$$N^{(\theta^A)}(k) := (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t g^{(\theta^A)}(K_t) \;\middle|\; K_0 = k\right].$$

$D^{(\theta^A)}(k)$ represents the expected total discounted cost while $N^{(\theta^A)}(k)$ represents the expected number of times active action is selected under policy $g^{(\theta^A)}$ starting from the initial information state k.

We will show (see Theorem 7) that the Whittle index for model A can be computed as a function of $D^{(\theta^A)}(k)$ and $N^{(\theta^A)}(k)$. First, we present a method to compute these two variables. Let

$$L^{(\theta^A)}(k) := (1-\beta)\sum_{t=k}^{\theta^A-1} \beta^{t-k}\bar{c}(t,0) + (1-\beta)\beta^{\theta^A-k}\bar{c}(\theta^A, 1)$$

$$M^{(\theta^A)}(k) := (1-\beta)\beta^{\theta^A-k}$$

where $L^{(\theta^A)}(k)$ and $M^{(\theta^A)}(k)$, respectively, denote the expected discounted cost and time starting from information state k until reaching information state θ^A for the first time.

Theorem 3. *For any $k \in \mathbb{Z}_{\geq 0}$, we have*

$$D^{(\theta^A)}(k) = L^{(\theta^A)}(k) + \beta^{\theta^A-k+1}\frac{L^{(\theta^A)}(0)}{1-\beta^{\theta^A+1}},$$

$$N^{(\theta^A)}(k) = M^{(\theta^A)}(k) + \beta^{\theta^A - k + 1} \frac{M^{(\theta^A)}(0)}{1 - \beta^{\theta^A + 1}}.$$

See Section 6 for the proof.

5.3.2. Model B

Let $J_\lambda^{(\theta^B)}(s,k)$ be the total discounted cost incurred under policy $g^{(\theta^B)}$ with penalty λ when the initial information state is (s,k), i.e.,

$$\begin{aligned} J_\lambda^{(\theta^B)}(s,k) &= (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(S_t, K_t, g^{(\theta^B)}(S_t, K_t)) \;\middle|\; (S_0, K_0) = (s,k)\right] \\ &:= D^{(\theta^B)}(s,k) + \lambda N^{(\theta^B)}(s,k), \end{aligned} \tag{19}$$

where

$$\begin{aligned} D^{(\theta^B)}(s,k) &:= (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}(S_t, K_t, g^{(\theta^B)}(S_t, K_t)) \;\middle|\; (S_0, K_0) = (s,k)\right], \\ N^{(\theta^B)}(s,k) &:= (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t g^{(\theta^B)}(S_t, K_t) \;\middle|\; (S_0, K_0) = (s,k)\right]. \end{aligned}$$

$D^{(\theta^B)}(s,k)$ and $N^{(\theta^B)}(s,k)$ have the same interpretations as the ones for model A. We will show (see Theorem 8) that Whittle index for model B can be computed as a function of $D^{(\theta^B)}(s,k)$ and $N^{(\theta^B)}(s,k)$. But first let us define vector $\boldsymbol{J}_\lambda^{(\theta^B)}(0) = (J_\lambda^{(\theta^B)}(1,0), \ldots, J_\lambda^{(\theta^B)}(|\mathcal{X}|,0))$ and vectors $\boldsymbol{D}^{(\theta^B)}(0)$ and $\boldsymbol{N}^{(\theta^B)}(0)$ in a similar manner. Then, from (19), $\boldsymbol{J}_\lambda^{(\theta^B)}(0) = \boldsymbol{D}^{(\theta^B)}(0) + \lambda \boldsymbol{N}^{(\theta^B)}(0)$. Let us also define

$$\begin{aligned} L^{(\theta^B)}(s,k) &:= (1-\beta) \sum_{t=k}^{\theta_s^B - 1} \beta^{t-k} \bar{c}(s,t,0) + (1-\beta)\beta^{\theta_s^B - k} \bar{c}(s, \theta_s^B, 1), \\ M^{(\theta^B)}(s,k) &:= (1-\beta)\beta^{\theta_s^B - k}. \end{aligned}$$

Let $\boldsymbol{L}^{(\theta^B)}(0) = (L^{(\theta^B)}(1,0), \ldots, L^{(\theta^B)}(|\mathcal{X}|,0))$ and $\boldsymbol{M}^{(\theta^B)}(0) = (M^{(\theta^B)}(1,0), \ldots, M^{(\theta^B)}(|\mathcal{X}|,0))$.

Theorem 4. *For any $(s,k) \in \mathcal{X} \times \mathbb{Z}_{\geq 0}$, we have*

$$\begin{aligned} D^{(\theta^B)}(s,k) &= L^{(\theta^B)}(s,k) + \beta^{\theta_s^B - k + 1} \sum_{r \in \mathcal{X}} Q_r D^{(\theta^B)}(r,0), \\ N^{(\theta^B)}(s,k) &= M^{(\theta^B)}(s,k) + \beta^{\theta_s^B - k + 1} \sum_{r \in \mathcal{X}} Q_r N^{(\theta^B)}(r,0). \end{aligned}$$

Let $Z^{(\theta^B)}$ be a $|\mathcal{X}| \times |\mathcal{X}|$ matrix where $Z_{sr}^{(\theta^B)} = \beta^{\theta_s^B + 1} Q_r$, for any $s, r \in \mathcal{X}$. Then,

$$\begin{aligned} \boldsymbol{D}^{(\theta^B)}(0) &= (I - Z^{(\theta^B)})^{-1} \boldsymbol{L}^{(\theta^B)}(0), \\ \boldsymbol{N}^{(\theta^B)}(0) &= (I - Z^{(\theta^B)})^{-1} \boldsymbol{M}^{(\theta^B)}(0). \end{aligned}$$

See Section 6 for the proof.

5.4. Finite state approximation

For computing Whittle index, we provide a finite state approximation of Proposition 3 for models A and B. Essentially, we truncate the countable set of possible information state K_t to a finite set and provide the approximation bound on the optimal value function for each of the models.

Theorem 5 (*Model A*). *Given $\ell \in \mathbb{N}$, let $\mathbb{N}_\ell := \{0, \ldots, \ell\}$ and $V_{\ell,\lambda} : \mathbb{N}_\ell \to \mathbb{R}$ be the unique fixed point of equation*

$$V_{\ell,\lambda}(k) = \min_{a \in \{0,1\}} H_{\ell,\lambda}(k,a), \;\; \hat{g}_{\ell,\lambda}(k) = \arg\min_{a \in \{0,1\}} H_{\ell,\lambda}(k,a)$$

where

$$\begin{aligned} H_{\ell,\lambda}(k,0) &= (1-\beta)\bar{c}(k,0) + \beta V_{\ell,\lambda}(\min\{k+1, \ell\}), \\ H_{\ell,\lambda}(k,1) &= (1-\beta)\bar{c}(k,1) + (1-\beta)\lambda + \beta V_{\ell,\lambda}(0). \end{aligned}$$

We set $\hat{g}_{\ell,\lambda}(k) = 1$ if $H_{\ell,\lambda}(k,0) = H_{\ell,\lambda}(k,1)$. Then, we have the following: (i) For any $0 \leq k \leq \ell$, we have

$$|V_\lambda(k) - V_{\ell,\lambda}(k)| \leq \frac{\beta^{\ell - k + 1} \operatorname{span}(\bar{c}_\lambda)}{1 - \beta}.$$

(ii) For all $k \in \mathbb{Z}_{\geq 0}$, $\lim_{\ell \to \infty} V_{\ell,\lambda}(k) = V_\lambda(k)$. *Moreover, let* $\hat{g}^*_\lambda(\cdot)$ *be any limit point of* $\{\hat{g}_{\ell,\lambda}(\cdot)\}_{\ell \geq 1}$. *Then, the policy* $\hat{g}^*_\lambda(\cdot)$ *is optimal for Problem 2.*

See Section 6 for the proof.

Remark 1 (*Choice of* ℓ)**.** Suppose we want to ensure that $|V_\lambda(0) - V_{\ell,\lambda}(0)| \leq \alpha$, where $\alpha > 0$ is some pre-specified constant. The result of Theorem 5 implies that we can ensure the above constraint by choosing $\ell > \log_\beta(\alpha(1-\beta)/\operatorname{span}(\bar{c}_\lambda))$.

Theorem 6 (*Model B*)**.** *Given* $\ell \in \mathbb{N}$, *let* $\mathbb{N}_\ell := \{0, \dots, \ell\}$ *and* $V_{\ell,\lambda} : \mathcal{X} \times \mathbb{N}_\ell \to \mathbb{R}$ *be the unique fixed point of equation*

$$V_{\ell,\lambda}(s,k) = \min_{a \in \{0,1\}} H_{\ell,\lambda}(s,k,a), \quad \hat{g}_{\ell,\lambda}(s,k) = \arg\min_{a \in \{0,1\}} H_{\ell,\lambda}(s,k,a)$$

where

$$H_{\ell,\lambda}(s,k,0) = (1-\beta)\bar{c}(s,k,0) + \beta V_{\ell,\lambda}(s, \min\{k+1, \ell\}),$$
$$H_{\ell,\lambda}(s,k,1) = (1-\beta)\bar{c}(s,k,1) + (1-\beta)\lambda + \beta \sum_{x' \in \tilde{\mathcal{X}}} Q_{x'} V_{\ell,\lambda}(x', 0).$$

We set $\hat{g}_{\ell,\lambda}(s,k) = 1$ *if* $H_{\ell,\lambda}(s,k,0) = H_{\ell,\lambda}(s,k,1)$. *Then, we have the following: (i) For any* $0 \leq k \leq \ell$,

$$|V_\lambda(s,k) - V_{\ell,\lambda}(s,k)| \leq \frac{\beta^{\ell-k+1} \operatorname{span}(c_\lambda)}{1-\beta}, \forall s \in \mathcal{X}.$$

(ii) For all $(s,k) \in \mathcal{X} \times \mathbb{Z}_{\geq 0}$, $\lim_{\ell \to \infty} V_{\ell,\lambda}(s,k) = V_\lambda(s,k)$. *Let* $\hat{g}^*_\lambda(\cdot,\cdot)$ *be any limit point of* $\{\hat{g}_{\ell,\lambda}(\cdot,\cdot)\}_{\ell \geq 1}$. *Then, the policy* $\hat{g}^*_\lambda(\cdot,\cdot)$ *is optimal for Problem 2.*

See Section 6 for the proof. Similar to Remark 1, choosing $\ell > \log_\beta(\alpha(1-\beta)/\operatorname{span}(\bar{c}_\lambda))$, for some $\alpha > 0$, ensures that $|V_\lambda(0,k) - V_{\ell,\lambda}(0,k)| \leq \alpha$.

Due to Theorems 5 and 6, we can restrict the countable part of the information state to a finite set, $\mathbb{N}_\ell$.

5.5. Computation of Whittle index

Next, we derive a closed form expression to compute the Whittle index for model A and provide an efficient algorithm to compute the Whittle index for model B.

5.5.1. Whittle index formula for model A

For model A, we obtain the Whittle index formula based on the two variables $D^{(\theta^A)}(\cdot)$ and $N^{(\theta^A)}(\cdot)$ as follows.

Theorem 7. *Let* $\Lambda^A_k = \{k_0 \in \{0,1,\dots,(\ell+1)-1\} : N^{(k)}(k_0) \neq N^{(k+1)}(k_0)\}$. *Then, under Assumption 3,* $\Lambda^A_k \neq \emptyset$, *and the Whittle index of model A at information state* $k \in \mathbb{N}_\ell$ *is*

$$w^A(k) = \min_{k_0 \in \Lambda^A_k} \frac{D^{(k+1)}(k_0) - D^{(k)}(k_0)}{N^{(k)}(k_0) - N^{(k+1)}(k_0)}. \tag{20}$$

Proof. Since model A is a restart model, the result follows from [31, Lemma 4].

Theorem 7 gives us a closed-form expression to approximately compute the Whittle index for model A.

5.5.2. Modified adaptive greedy algorithm for model B

Let $B = |\mathcal{X}|(\ell+1)$ and $B_D (\leq B)$ denote the number of distinct Whittle indices. Let $\Lambda^* = \{\lambda_0, \lambda_1, \dots, \lambda_{B_D}\}$ where $\lambda_1 < \lambda_2 < \cdots < \lambda_{B_D}$ denote the sorted distinct Whittle indices with $\lambda_0 = -\infty$. Let $\mathcal{W}_b := \{(s,k) \in \mathcal{X} \times \mathbb{N}_\ell : w(s,k) \leq \lambda_b\}$. For any subset $S \subseteq \mathcal{X} \times \mathbb{N}_\ell$, define the policy $\bar{g}^{(S)} : \mathcal{X} \times \mathbb{N}_\ell \to \{0,1\}$ as

$$\bar{g}^{(S)}(s,k) = \begin{cases} 0, & \text{if } (s,k) \in S \\ 1, & \text{if } (s,k) \in (\mathcal{X} \times \mathbb{N}_\ell) \backslash S. \end{cases}$$

Given $\mathcal{W}_b$, define $\Phi_b = \{(s,k) \in (\mathcal{X} \times \mathbb{N}_\ell) \setminus \mathcal{W}_b : (s, \max\{0, k-1\}) \in \mathcal{W}_b\}$ and $\Gamma_{b+1} = \mathcal{W}_{b+1} \backslash \mathcal{W}_b$. Additionally, for any $b \in \{0, \dots, B_D - 1\}$, and all states $y \in \Phi_b$, define $h_b = \bar{g}^{(\mathcal{W}_b)}$, $h_{b,y} = \bar{g}^{(\mathcal{W}_b \cup \{y\})}$ and $\Lambda_{b,y} = \{(x,k) \in (\mathcal{X} \times \mathbb{N}_\ell) : N^{(h_b)}(x,k) \neq N^{(h_{b,y})}(x,k)\}$. Then, for all $(x,k) \in \Lambda_{b,y}$, define

$$\mu_{b,y}(x,k) = \frac{D^{(h_{b,y})}(x,k) - D^{(h_b)}(x,k)}{N^{(h_b)}(x,k) - N^{(h_{b,y})}(x,k)}. \tag{21}$$

Lemma 3. *For* $d \in \{0, \dots, B_D - 1\}$, *we have the following:*

Algorithm 1: Computing Whittle index of all information states of model B

input: RB $(\mathcal{X}, \{0,1\}, P, Q, c, \rho)$, discount factor β, truncation ℓ.
Initialize $b = 0$, $\mathcal{W}_b = \emptyset$.
while $\mathcal{W}_b \neq \mathcal{X} \times \mathbb{N}_\ell$ **do**
  Compute $\Lambda_{b,y}$ and $\mu_{b,y}(x)$ using (21), $\forall y \in \Phi_b$.
  Compute $\mu^*_{b,y} = \min_{x \in \Lambda_{b,y}} \mu_{b,y}(x)$, $\forall y \in \Phi_b$.
  Compute $\lambda_{b+1} = \min_{y \in \Phi_b} \mu^*_{b,y}$.
  Compute $\Gamma_{b+1} = \arg\min_{y \in \Phi_b} \mu^*_{b,y}$.
  Set $w(z) = \lambda_{b+1}$, $\forall z \in \Gamma_{b+1}$.
  Set $\mathcal{W}_{b+1} = \mathcal{W}_b \cup \Gamma_{b+1}$.
  Set $b = b + 1$.

1. *For all $y \in \Gamma_{b+1}$, we have $w(y) = \lambda_{b+1}$.*
2. *For all $y \in \Phi_b$ and $\lambda \in (\lambda_b, \lambda_{b+1}]$, we have $J_\lambda^{(h_{b,y})}(x) \geq J_\lambda^{(h_b)}(x)$ for all $x \in \mathcal{X}$ with equality if and only if $y \in \mathcal{W}_{b+1} \setminus \mathcal{W}_b$ and $\lambda = \lambda_{b+1}$.*

Proof. The result follows from [31, Lemma 3]. The only difference is that since we know from Theorem 2 that the optimal policy is a threshold policy with respect to the second dimension, we restrict to $y \in \Phi_b$.

Theorem 8. *The following properties hold:*

1. *For any $y \in \Gamma_{b+1}$, the set $\Lambda_{b,y}$ is non-empty.*
2. *For any $x \in \Lambda_{b,y}$, $\mu_{b,y}(x) \geq \lambda_{b+1}$ with equality if and only if $y \in \Gamma_{b+1}$.*

Proof. The result follows from [31, Theorem 2]. Similar to Lemma 3, we consider $y \in \Phi_b$.

By Theorem 8, we can find the Whittle indices iteratively. This approach is summarized in Algorithm 1. For a computationally-efficient implementation using the Sherman–Morrison formula, see [31, Algorithm 2].

6. Proof of main results

6.1. Proof of Theorem 2

Let μ^1 and μ^2 be two probability mass functions on totally ordered set $\tilde{\mathcal{X}}$. Then we say μ^1 *stochastically dominates* μ^2 if for all $x \in \tilde{\mathcal{X}}$, $\sum_{z \in \tilde{\mathcal{X}}_{\geq x}} \mu^1_z \geq \sum_{z \in \tilde{\mathcal{X}}_{\geq x}} \mu^2_z$. Given two $|\tilde{\mathcal{X}}| \times |\tilde{\mathcal{X}}|$ transition matrices M and N, we say M stochastically dominates N if each row of M stochastically dominates the corresponding N. A basic property of stochastic dominance is the following.

Lemma 4. *If M^1 stochastically dominates M^2 and c is an non-decreasing function defined on $\tilde{\mathcal{X}}$, then for all $x \in \tilde{\mathcal{X}}$, $\sum_{y \in \tilde{\mathcal{X}}} M^1_{xy} c(y) \geq \sum_{y \in \tilde{\mathcal{X}}} M^2_{xy} c(y)$.*

Proof. This follows from [42, Lemma 4.7.2].

Consider a fully-observable restless bandit process $\{(\tilde{\mathcal{X}}, \{0,1\}, \{\tilde{P}, \tilde{Q}\}, \tilde{c}, \tilde{\pi}_0)\}$ (note that $\mathcal{Y}$ is removed due to the observability assumption). According to [31], we say a fully-observable restless bandit process is *stochastic monotone* if it satisfies the following conditions.

(D1) $\tilde{P}$ and $\tilde{Q}$ are stochastic monotone transition matrices.
(D2) For any $z \in \tilde{\mathcal{X}}$, $\sum_{w \in \tilde{\mathcal{X}}_{\geq z}} [\tilde{P} - \tilde{Q}]_{xw}$ is non-decreasing in $x \in \tilde{\mathcal{X}}$.
(D3) For any $a \in \{0,1\}$, $\tilde{c}(x,a)$ is non-decreasing in x.
(D4) $\tilde{c}(x,a)$ is submodular in (x,a).

The following is established in [31, Lemma 5].

Proposition 5. *The optimal policy of a stochastic monotone fully-observable restless bandit process is a threshold policy denoted by $\tilde{g}$, which is a policy which takes passive action for states below a threshold denoted by $\tilde{\theta}$ and active action for the rest of the states, i.e.,*

$$\tilde{g} = \begin{cases} 0, & x < \tilde{\theta} \\ 1, & \textit{otherwise}. \end{cases}$$

6.1.1. Proof of Theorem 2, part 1

We show that each machine in model A is a stochastic monotone fully-observable restless bandit process. Each condition of stochastic monotone fully-observable restless bandit process is presented and proven for model A below.

(D1') The transition probability matrix under passive action for model A based on the information states is $P^A_{xy} = \mathbb{I}_{\{y=x+1\}}$ and the transition probability matrix under active action for model A is $Q^A_{xy} = \mathbb{I}_{\{y=0\}}$. Thus, P^A and Q^A are stochastic monotone matrices.

(D2') Since P^A is a stochastic monotone matrix and Q^A has constant rows, $\sum_{r\geq z}[P^A - Q^A]_{sr}$ is non-decreasing in s for any integer $z \geq 0$.

(D3') As P stochastically dominates the identity matrix, we infer from [48, Theorem 1.1-b and Theorem 1.2-c], that QP^ℓ stochastically dominates QP^k for any $\ell > k \geq 0$. Additionally, $c_\lambda(x,a)$ is non-decreasing in x for any $a \in \{0,1\}$. By (16) we have $\bar{c}_\lambda(k,a) = \sum_{x\in\mathcal{X}}[(QP)^k]_x c_\lambda(x,a)$. Therefore, by Lemma 4, $\bar{c}_\lambda(k,a)$ is non-decreasing in k.

(D4') As $c(x,a)$ is submodular in (x,a) and as shown in (D3'), QP^ℓ stochastically dominates QP^k for any $\ell > k \geq 0$. Therefore, by Lemma 4, $\bar{c}_\lambda(k,0) - \bar{c}_\lambda(k,1) = \sum_{x\in\mathcal{X}}[(QP)^k]_x(c_\lambda(x,0) - c_\lambda(x,1))$ is non-decreasing in (k,a).

Therefore, according to Proposition 5, the optimal policy of a fully-observable restless bandit process under model A is a threshold based policy.

Finally, since the optimal policy is threshold based, the passive set $\mathcal{W}_\lambda$ is given by $\{k \in \mathbb{Z}_{\geq -1} : k < \theta^A_\lambda\}$. As shown in Theorem 1, model A is indexable. Therefore, the passive set must be non-decreasing in λ, which implies that the threshold θ^A_λ is non-decreasing in λ.

6.1.2. Proof of Theorem 2, part 2

We first characterize the behavior of value function and state–action value function for Model B.

Lemma 5. *We have*

a. $\bar{c}_\lambda(s,k,a)$ is non-decreasing in k for any $s \in \mathcal{X}$ and $a \in \{0,1\}$.
b. Given a fixed λ, $V_\lambda(s,k)$ is non-decreasing in k for any $s \in \mathcal{X}$.
c. $\bar{c}_\lambda(s,k,a)$ is submodular in (k,a), for any $s \in \mathcal{X}$.
d. $H_\lambda(s,k,a)$ is submodular in (k,a), for any $s \in \mathcal{X}$.

Proof. The proof of each part is as follows.

a. By definition, we have

$$\bar{c}_\lambda(s,k,a) = \sum_{x\in\mathcal{X}}[\delta_s P^k](x)c(x,a) + \lambda a.$$

Similar to the proof of (D3') in Proposition 5, for a given $s \in \mathcal{X}$ and $a \in \{0,1\}$, $[\delta_s P^k](x)$ is non-decreasing in k and x and as $c(x,a)$ is non-decreasing in x, $\bar{c}_\lambda(s,k,a)$ is non-decreasing in k.

b. Let

$$\begin{aligned}
H^j_\lambda(s,k,0) &:= (1-\beta)\bar{c}(s,k,0) + \beta V^j_\lambda(s,k+1),\\
H^j_\lambda(s,k,1) &:= (1-\beta)\bar{c}(s,k,1) + (1-\beta)\lambda + \beta\sum_r Q_r V^j_\lambda(r,0),\\
V^{j+1}_\lambda(s,k) &:= \min_{a\in\{0,1\}}\{H^j_\lambda(s,k,a)\},
\end{aligned}$$

where $V^0_\lambda(\cdot,\cdot) = 0$ for all $(s,k) \in \mathcal{X}\times\mathbb{Z}_{\geq 0}$.
Claim: $V^j_\lambda(s,k)$ is non-decreasing in k for any $s \in \mathcal{X}$ and $j \geq 0$.
We prove the claim by induction. By construction, $V^0_\lambda(s,k)$ is non-decreasing in k for any $s \in \mathcal{X}$. This forms the basis of induction. Now assume that $V^j_\lambda(s,k)$ is non-decreasing in k for any $s \in \mathcal{X}$ and some $j \geq 0$. Consider $\ell > k \geq 0$. Then, by induction hypothesis we have

$$\begin{aligned}
H^j_\lambda(s,\ell,0) &= (1-\beta)\bar{c}(s,\ell,0) + \beta V^j_\lambda(s,\ell+1)\\
&\geq (1-\beta)\bar{c}(s,k,0) + \beta V^j_\lambda(s,k+1) = H^j_\lambda(s,k,0),\\
H^j_\lambda(s,\ell,1) &= (1-\beta)\bar{c}(s,\ell,1) + (1-\beta)\lambda + \beta\sum_r Q_r V^j_\lambda(r,0)\\
&\geq (1-\beta)\bar{c}(s,k,1) + (1-\beta)\lambda + \beta\sum_r Q_r V^j_\lambda(r,0) = H^j_\lambda(s,k,1).
\end{aligned}$$

Therefore,

$$V^{j+1}_\lambda(s,\ell) = \min_a\{H^j_\lambda(s,\ell,a)\} \geq \min_a\{H^j_\lambda(s,k,a)\} = V^{j+1}_\lambda(s,k).$$

Thus, $V^{j+1}_\lambda(s,k)$ is non-decreasing in k for any $s \in \mathcal{X}$. This completes the induction step. $V_\lambda(s,k) = \lim_{j\to\infty} V^j_\lambda(s,k)$ and monotonicity is preserved under limits, the induction proof is complete.

c. $c(x,a)$ is submodular in (x,a). Also, note that $\delta_s P^k$ is the sth row of P^k. Thus, $\delta_s P^{k+1}$ stochastically dominates $\delta_s P^k$ and by Lemma 4 we have

$$\sum_{x\in\mathcal{X}} [\delta_s(P^{k+1}-P^k)]_x (c(x,0)-c(x,1)) \geq 0.$$

Therefore,

$$\sum_{x\in\mathcal{X}} [\delta_s(P^{k}-P^{k+1})]_x c(x,1) \geq \sum_{x\in\mathcal{X}} [\delta_s(P^{k}-P^{k+1})]_x c(x,0).$$

Consequently,

$$\sum_{x\in\mathcal{X}} [\delta_s P^k]_x c(x,1) - \sum_{x\in\mathcal{X}} [\delta_s P^k]_x c(x,0) \geq \sum_{x\in\mathcal{X}} [\delta_s P^{k+1}]_x c(x,1) - \sum_{x\in\mathcal{X}} [\delta_s P^{k+1}]_x c(x,0).$$

Hence,

$$\bar{c}(s,k,1)-\bar{c}(s,k,0) \geq \bar{c}(s,k+1,1)-\bar{c}(s,k+1,0).$$

d. As for any $s\in\mathcal{X}$, $V_\lambda(s,k)$ is non-decreasing in k, and $\bar{c}_\lambda(s,k,a)$ is submodular in (k,a), for any $k\in\mathbb{N}_\ell$ and $a\in\{0,1\}$, we have

$$\begin{aligned}
H_\lambda(s,k,1)-H_\lambda(s,k,0) &= (1-\beta)\bar{c}(s,k,1)+(1-\beta)\lambda+\beta\sum_r Q_r V_\lambda(r,0)\\
&\quad-(1-\beta)\bar{c}(s,k,0)-\beta V_\lambda(s,k+1)\\
&\geq (1-\beta)\bar{c}(s,k+1,1)+(1-\beta)\lambda+\beta\sum_r Q_r V_\lambda(r,0)\\
&\quad-(1-\beta)\bar{c}(s,k+1,0)-\beta V_\lambda(s,k+2)\\
&= H_\lambda(s,k+1,1)-H_\lambda(s,k+1,0).
\end{aligned}$$

Lemma 6. *Suppose $f:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is a submodular function and for each $x\in\mathcal{X}$, $\min_{y\in\mathcal{Y}} f(x,y)$ exists. Then,* $\max\{\arg\min_{y\in\mathcal{Y}} f(x,y)\}$ *is monotone non-decreasing in x.*

Proof. This result follows from [42, Lemma 4.7.1].

Now, we conclude that as $H_\lambda(s,k,a)$ is submodular in (k,a) for any $s\in\mathcal{X}$, then, based on Lemma 6 and as only two actions is available, the optimal policy is a threshold policy specified in the theorem statement.

Finally, since the optimal policy is threshold based, the passive set $\mathcal{W}_\lambda$ is given by $\{(s,k)\in\mathcal{X}\times\mathbb{Z}_{\geq -1} : k<\theta^B_{s,\lambda}\}$. As shown in Theorem 1, model B is indexable. Therefore, the passive set must be non-decreasing in λ, which implies that, for every $s\in\mathcal{X}$, the threshold $\theta^B_{s,\lambda}$ is non-decreasing in λ.

6.2. Proof of Theorem 3

By the strong Markov property, we have

$$\begin{aligned}
D^{(\theta^A)}(k) &= (1-\beta)\sum_{j=k}^{\theta^A}\beta^t\bar{c}(t,g(t))+\beta^{\theta^A-k+1}D^{(\theta^A)}(0) = L^{(\theta^A)}(k)+\beta^{\theta^A-k+1}D^{(\theta^A)}(0),\\
N^{(\theta^A)}(k) &= (1-\beta)\beta^{\theta^A-k}+\beta^{\theta^A-k+1}N^{(\theta^A)}(0) = M^{(\theta^A)}(k)+\beta^{\theta^A-k+1}N^{(\theta^A)}(0).
\end{aligned}$$

If we set $k=0$ in the above,

$$D^{(\theta^A)}(0)=\frac{L^{(\theta^A)}(0)}{1-\beta^{\theta^A+1}} \text{ and } N^{(\theta^A)}(0)=\frac{M^{(\theta^A)}(0)}{1-\beta^{\theta^A+1}}.$$

6.3. Proof of Theorem 4

By the strong Markov property, we have

$$\begin{aligned}
D^{(\theta^B)}(s,k) &= (1-\beta)\sum_{j=k}^{\theta^B_s}\beta^t\bar{c}(s,t,g(s,t))+\beta^{\theta^B_s-k+1}\sum_{r\in\mathcal{X}}Q_r D^{(\theta^B)}(r,0)\\
&= L^{(\theta^B)}(s,k)+\beta^{\theta^B_s-k+1}\sum_{r\in\mathcal{X}}Q_r D^{(\theta^B)}(r,0),\\
N^{(\theta^B)}(s,0) &= (1-\beta)\beta^{\theta^B_s-k}+\beta^{\theta^B_s-k+1}\sum_{r\in\mathcal{X}}Q_r N^{(\theta^B)}(r,0)\\
&= M^{(\theta^B)}(s,k)+\beta^{\theta^B_s-k+1}\sum_{r\in\mathcal{X}}Q_r N^{(\theta^B)}(r,0).
\end{aligned}$$

If we set $k = 0$ in the above,

$$D^{(\theta^B)}(s,0) = L^{(\theta^B)}(s,0) + \beta^{\theta^B_s+1} \sum_{r\in\mathcal{X}} Q_r D^{(\theta^B)}(r,0),$$
$$N^{(\theta^B)}(s,0) = M^{(\theta^B)}(s,0) + \beta^{\theta^B_s+1} \sum_{r\in\mathcal{X}} Q_r N^{(\theta^B)}(r,0).$$

which results in

$$\boldsymbol{D}^{(\theta^B)}(0) = \boldsymbol{L}^{(\theta^B)}(0) + Z^{(\theta^B)} \boldsymbol{D}^{(\theta^B)}(0),$$
$$\boldsymbol{N}^{(\theta^B)}(0) = \boldsymbol{M}^{(\theta^B)}(0) + Z^{(\theta^B)} \boldsymbol{N}^{(\theta^B)}(0)$$

and hence, the statement is obtained by reformation of the terms inside the equations.

6.4. Proof of Theorem 5

(i): Starting from information state $k \in \mathbb{N}_\ell$, the cost incurred by $\hat{g}_{\ell,\lambda}(\cdot)$ is the same as $g^A_\lambda(\cdot)$ for information states $\{k, \dots, \ell\}$. The per-step cost incurred by $\hat{g}_{\ell,\lambda}(\cdot)$ differs from $g^A_\lambda(\cdot)$ for information states $\{\ell+1, \dots\}$ by at most $\mathrm{span}(c_\lambda)$.

(ii): The sequence of finite-state models described above is an *augmentation type approximation sequence*. As a result, a limit point of $\hat{g}^*_\lambda$ exists and the final result holds by [49, Proposition B.5, Theorem 4.6.3].

6.5. Proof of Theorem 6

(i): Starting from information state (s,k), given any $s \in \mathcal{X}$ and $k \in \mathbb{N}_\ell$, the cost incurred by $\hat{g}_{\ell,\lambda}(\cdot,\cdot)$ is the same as $g^B_\lambda(\cdot,\cdot)$ for information states $\{(s,l)\}_{l=k}^{\ell}$. The per-step cost incurred by $\hat{g}_{\ell,\lambda}(\cdot,\cdot)$ differs from $g^B_\lambda(\cdot,\cdot)$ for later realized information states by at most Δc_λ. Thus, the bound holds.

(ii): The sequence of finite-state models described above is an *augmentation type approximation sequence*. As a result, a limit point of $\hat{g}^*_\lambda$ exists and the final result holds [49, Proposition B.5, Theorem 4.6.3].

7. Numerical analysis

In this section, we consider Example 2 and compare the performance of the following policies:

- OPT: the optimal policy obtained using dynamic programming. As discussed earlier, the dynamic programming computation to obtain the optimal policy suffers from the curse of dimensionality. Therefore, the optimal policy can be computed only for small-scale models.
- MYP: myopic policy, which is a heuristic which sequentially selects m machines as follows. Suppose $\zeta < m$ machines have been selected. Then select machine $\zeta+1$ to be the machine which provides the smallest increase in the total per-step cost. The detailed description for model B is shown in Alg. 2.
- WIP: whittle index heuristic, as described in this paper.

Algorithm 2: Myopic Heuristic (Model B)

input: RB $(\mathcal{X}, \{0,1\}, P, Q, c, \rho)$, discount factor β, m.
Initialize $t = 0$.
while $t \geq 0$ **do**
  Set $\zeta = 0$.
  while $\zeta \leq m$ **do**
    Compute $i^*_\zeta \in \arg\min_{i\in\mathcal{Z}} \sum_{j\in\mathcal{Z}\setminus\{i\}} \bar{c}^j(S^j_t, K^j_t, 0) + \bar{c}^i(S^i_t, K^i_t, 1)$.
    Let $\mathcal{M} = \mathcal{M} \cup \{i^*_\zeta\}$, $\mathcal{Z} = \mathcal{Z} \setminus \{i^*_\zeta\}$.
    Set $\zeta = \zeta + 1$.
  Service the machines with indices collected in $\mathcal{M}$.
  Update K^i_t according to (14) and S^i_t according to (15) for all $i \in \mathcal{N}$.
  Set $t = t + 1$.

7.1. Experiments and results

We conduct numerical experiments for both models A and B, and vary the number n of machines, the number m of service-persons and the parameters associated with each machine. There are three parameters associated with each machine: the deterioration probability matrix P^i, the reset PMF Q^i and the per-step cost $c^i(x,a)$. We assume the matrix P^i is chosen from a family of four types of structured transition matrices $\mathcal{P}_\gamma(p)$, $\gamma \in \{1,2,3,4\}$ where p is a parameter of the model. The details of all these models are presented in Appendix. We assume each element of Q^i is sampled from Exp(1), i.e., exponential distribution with the rate

Table 1
ε_{MYP} for different choice of parameters of Model A in Experiment 2.

(a) Model A, $m = 1$

ε_{MYP}		γ			
		1	2	3	4
n	20	1.99	2.54	2.24	7.44
	40	3.41	6.90	4.71	8.14
	60	2.97	6.19	2.80	6.70

(b) Model A, $m = 5$

ε_{MYP}		γ			
		1	2	3	4
n	20	0.21	0.26	0.19	0.97
	40	0.68	1.73	1.28	4.54
	60	1.36	2.35	2.32	6.41

Table 2
ε_{MYP} for different choice of parameters of Model B in Experiment 2.

(a) Model B, $m = 1$

ε_{MYP}		γ			
		1	2	3	4
n	20	7.67	11.17	12.12	9.39
	40	14.96	13.85	14.55	9.17
	60	15.02	12.12	13.39	6.63

(b) Model B, $m = 5$

ε_{MYP}		γ			
		1	2	3	4
n	20	0.63	1.62	1.01	2.92
	40	2.92	3.14	3.21	6.57
	60	4.86	7.22	6.99	9.96

parameter of 1, and then normalized such that the sum of all elements becomes 1. Finally, we assume that the per-step cost is given by $c^i(x, 0) = (x-1)^2$ and $c^i(x, 1) = 0.5|\mathcal{X}^i|^2$.

In all experiments, the discount factor is $\beta = 0.99$. The performance of every policy is evaluated using Monte-Carlo simulation of length 1000 averaged over 5000 sample paths.

In Experiment 1, we consider a small scale problem where we can compute OPT and we compare the performance of WIP with it. However, in Experiment 2, we consider a large scale problem where we compare the performance of WIP with MYP as computing the optimal policy is highly time-consuming.

The code for both experiments is available at [50].

Experiment (1) comparison of whittle index with the optimal policy.

In this experiment, we compare the performance of WIP with OPT. We assume $|\mathcal{X}| = 4$, $(\ell + 1) = 6$ and $n = 3$, $m = 1$ for both models A and B. In order to model heterogeneous machines, we consider the following. Let $(p_1, \dots, p_n)$ denote n equispaced points in the interval $[0.05, 0.95]$. Then we choose $\mathcal{P}_{\gamma}(p_i)$ as the transition matrix of machine i. We denote the accumulated discounted cost of WIP and OPT by $J(\text{WIP})$ and $J(\text{OPT})$, respectively. In order to have a better perspective of the performances, we compute the relative performance of WIP with respect to OPT by computing

$$\alpha_{\text{OPT}} = 100 \times \frac{J(\text{OPT})}{J(\text{WIP})}. \tag{22}$$

The closer α is to 100, the closer WIP is to OPT. The results of α_{OPT} for all different combinations of parameter were 100 which means the Whittle policy is as good as the optimal policy.

Experiment (2) comparison of whittle index with the myopic policy for structured models.

In this experiment, we increase the state space size to $|\mathcal{X}| = 20$ and we set $(\ell + 1) = 40$, we select n from the set $\{20, 40, 60\}$ and m from the set $\{1, 5\}$. We denote the accumulated discounted cost of MYP by $J(\text{MYP})$. In order to have a better perspective of the performances, we compute the relative improvement of WIP with respect to MYP by computing

$$\varepsilon_{\text{MYP}} = 100 \times \frac{J(\text{MYP}) - J(\text{WIP})}{J(\text{MYP})}. \tag{23}$$

Note that $\varepsilon_{\text{MYP}} > 0$ means that WIP performs better than MYP. We generate structured transition matrices, similar to Experiment 1, and apply the same procedure to build heterogeneous machines. The results of ε_{MYP} for different choice of the parameters for models A and B are shown in Tables 1 and 2, respectively.

7.2. Discussion

In Experiment 1 where WIP is compared with OPT, we observe α_{OPT} is very close to 100 for almost all experiments, implying that WIP performs as well as OPT for these experiments. α_{OPT} in model B is less than model A as model B is more complex than model A for a given set of parameters and hence, the difference between the performance of the two polices is more than model A.

In Experiment 2 where WIP is compared with MYP, we observe ε_{MYP} ranges from 0.2% to 15%. In a similar interpretation as Experiment 1, as model B is more complex than model A, ε_{MYP} for model B is higher than the ones model A given the same set of parameters.

Furthermore, we observe that as n increases, ε_{MYP} also increases overall. Also, as m increases, ε_{MYP} decreases in general. This suggests that as m increases, there is an overlap between the set of machines chosen according to WIP and MYP, and hence, the performance of WIP and MYP become close to each other.

8. Conclusion

We investigated partially observable restless bandits. Unlike most of the existing literature which restricts attention to models with binary state space, we consider general state space models. We presented two observation models, which we call model A and model B, and showed that the partially observable restless bandits are indexable for both models.

To compute the Whittle index, we work with a countable space representation rather than the belief state representation. We established certain qualitative properties of the auxiliary problem to compute the Whittle index. In particular, for both models we showed that the optimal policies of the auxiliary problem satisfy threshold properties. For model A, we used the threshold property to obtain a closed form expression to compute the Whittle index. For model B, we used the threshold policy to present a refinement of the adaptive greedy algorithm of [31] to compute the Whittle index.

Finally, we presented a detailed numerical study of a machine maintenance model. We observed that for small-scale models, the Whittle index policy is close-to-optimal and for large-scale models, the Whittle index policy outperforms the myopic policy baseline.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nima Akbarzadeh reports financial support was provided by Quebec Research Fund Nature and Technology.

Data availability

No data was used for the research described in the article.

Appendix. Structured Markov chains

Consider a Markov chain with $|\mathcal{X}|$ states. Then a family of structured stochastic monotone matrices which dominates the identity matrix is illustrated below.

1. **Matrix** $\mathcal{P}_1(p)$: Let $q_1 = 1-p$ and $q_2 = 0$. Then,

$$\mathcal{P}_1(p) = \begin{bmatrix} p & q_1 & q_2 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & p & q_1 & q_2 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & p & q_1 & q_2 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & p & q_1 & q_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & p & q_1 & q_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & q_1+q_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

2. **Matrix** $\mathcal{P}_2(p)$: Similar to $\mathcal{P}_1(p)$ with $q_1 = (1-p)/2$ and $q_2 = (1-p)/2$.
3. **Matrix** $\mathcal{P}_3(p)$: Similar to $\mathcal{P}_1(p)$ with $q_1 = 2(1-p)/3$ and $q_2 = (1-p)/3$.
4. **Matrix** $\mathcal{P}_4(p)$: Let $q_i = (1-p)/(\mathcal{X}-i)$. Then,

$$\mathcal{P}_4(p) = \begin{bmatrix} p & q_1 & q_1 & \cdots & q_1 & q_1 \\ 0 & p & q_2 & \cdots & q_2 & q_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & p & q_{n-1} \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

References

[1] R. Meshram, D. Manjunath, A. Gopalan, On the whittle index for restless multiarmed hidden Markov bandits, IEEE Trans. Automat. Control 63 (9) (2018) 3046–3053.

[2] S. Guha, K. Munagala, P. Shi, Approximation algorithms for restless bandit problems, J. ACM 58 (1) (2010) 3.

[3] K. Kaza, R. Meshram, V. Mehta, S.N. Merchant, Sequential decision making with limited observation capability: Application to wireless networks, IEEE Trans. Cogn. Commun. Netw. 5 (2) (2019) 237–251.

[4] K. Kaza, V. Mehta, R. Meshram, S. Merchant, Restless bandits with cumulative feedback: Applications in wireless networks, in: Wireless Communications and Networking Conference, IEEE, 2018, pp. 1–6.

[5] S. Aalto, P. Lassila, P. Osti, Whittle index approach to size-aware scheduling for time-varying channels with multiple states, Queueing Syst. 83 (3–4) (2016) 195–225.

[6] M. Larrañaga, M. Assaad, A. Destounis, G.S. Paschos, Dynamic pilot allocation over Markovian fading channels: A restless bandit approach, in: Information Theory Workshop, IEEE, 2016, pp. 290–294.

[7] N. Akbarzadeh, A. Mahajan, Dynamic spectrum access under partial observations: A restless bandit approach, in: Canadian Workshop on Information Theory, IEEE, 2019, pp. 1–6.

[8] S.S. Villar, J. Bowden, J. Wason, Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges, Statist. Sci. Rev. J. Inst. Math. Statist. 30 (2) (2015) 199.

[9] C. Abad, G. Iyengar, A near-optimal maintenance policy for automated DR devices, IEEE Trans. Smart Grid 7 (3) (2016) 1411–1419.

[10] K.D. Glazebrook, H.M. Mitchell, P.S. Ansell, Index policies for the maintenance of a collection of machines by a set of repairmen, European J. Oper. Res. 165 (1) (2005) 267–284.
[11] S.S. Villar, Indexability and optimal index policies for a class of reinitialising restless bandits, Probab. Engrg. Inform. Sci. 30 (1) (2016) 1–23.
[12] Y. Qian, C. Zhang, B. Krishnamachari, M. Tambe, Restless poachers: Handling exploration-exploitation tradeoffs in security domains, in: Int. Conf. on Autonomous Agents & Multiagent Systems, 2016, pp. 123–131.
[13] V.S. Borkar, Whittle index for partially observed binary Markov decision processes, IEEE Trans. Automat. Control 62 (12) (2017) 6614–6618.
[14] V.S. Borkar, G.S. Kasbekar, S. Pattathil, P. Shetty, Opportunistic scheduling as restless bandits, IEEE Trans. Control Netw. Syst. (2017).
[15] K.E. Avrachenkov, V.S. Borkar, Whittle index policy for crawling ephemeral content, IEEE Trans. Control Netw. Syst. 5 (1) (2018) 446–455.
[16] C.H. Papadimitriou, J.N. Tsitsiklis, The complexity of optimal queuing network control, Math. Oper. Res. 24 (2) (1999) 293–305.
[17] P. Whittle, Restless bandits: Activity allocation in a changing world, J. Appl. Probab. 25 (A) (1988) 287–298.
[18] J.C. Gittins, Bandit processes and dynamic allocation indices, J. R. Stat. Soc. Ser. B Stat. Methodol. (1979) 148–177.
[19] R.R. Weber, G. Weiss, On an index policy for restless bandits, J. Appl. Probab. 27 (3) (1990) 637–648.
[20] C. Lott, D. Teneketzis, On the optimality of an index rule in multichannel allocation for single-hop mobile networks with multiple service classes, Probab. Engrg. Inform. Sci. 14 (3) (2000) 259–297.
[21] K. Liu, Q. Zhao, Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access, IEEE Trans. Inform. Theory 56 (11) (2010) 5547–5567.
[22] J. Niño-Mora, Dynamic priority allocation via restless bandit marginal productivity indices, TOP 15 (2) (2007) 161–198.
[23] P.S. Ansell, K.D. Glazebrook, J. Niño-Mora, M. O'Keeffe, Whittle's index policy for a multi-class queueing system with convex holding costs, Math. Methods Oper. Res. 57 (1) (2003) 21–39.
[24] K.D. Glazebrook, D. Ruiz-Hernandez, C. Kirkbride, Some indexable families of restless bandit problems, Adv. Appl. Probab. 38 (3) (2006) 643–672.
[25] U. Ayesta, M. Erausquin, P. Jacko, A modeling framework for optimizing the flow-level scheduling with time-varying channels, Perform. Eval. 67 (11) (2010) 1014–1029.
[26] N. Akbarzadeh, A. Mahajan, Restless bandits with controlled restarts: Indexability and computation of Whittle index, in: Conference on Decision and Control, 2019, pp. 7294–7300.
[27] T.W. Archibald, D.P. Black, K.D. Glazebrook, Indexability and index heuristics for a simple class of inventory routing problems, Oper. Res. 57 (2) (2009) 314–326.
[28] K. Avrachenkov, U. Ayesta, J. Doncel, P. Jacko, Congestion control of TCP flows in internet routers by means of index policy, Comput. Netw. 57 (17) (2013) 3463–3478.
[29] K. Glazebrook, D. Hodge, C. Kirkbride, Monotone policies and indexability for bidirectional restless bandits, Adv. Appl. Probab. 45 (1) (2013) 51–85.
[30] Z. Yu, Y. Xu, L. Tong, Deadline scheduling as restless bandits, 63(8), 2018, pp. 2343–2358.
[31] N. Akbarzadeh, A. Mahajan, Conditions for indexability of restless bandits and an $\mathcal{O}\left(K^3\right)$ algorithm to compute Whittle index, Adv. Appl. Probab. 54 (4) (2022) 1164–1192.
[32] J. Niño-Mora, Restless bandits, partial conservation laws and indexability, Adv. Appl. Probab. 33 (1) (2001) 76–98.
[33] N. Gast, B. Gaujal, K. Khun, Computing Whittle (and Gittins) index in subcubic time, 2022, arXiv preprint arXiv:2203.05207.
[34] D. Ruiz-Hernández, J.M. Pinar-Pérez, D. Delgado-Gómez, Multi-machine preventive maintenance scheduling with imperfect interventions: A restless bandit approach, Comput. Oper. Res. 119 (2020) 104927.
[35] K. Liu, R. Weber, Q. Zhao, Indexability and Whittle index for restless bandit problems involving reset processes, in: 2011 50th IEEE Conference on Decision and Control and European Control Conference, 2011, pp. 7690–7696.
[36] K. Liu, Q. Zhao, Dynamic intrusion detection in resource-constrained cyber networks, in: 2012 IEEE International Symposium on Information Theory Proceedings, IEEE, 2012, pp. 970–974.
[37] A. Mate, J. Killian, H. Xu, A. Perrault, M. Tambe, Collapsing bandits and their application to public health intervention, Adv. Neural Inf. Process. Syst. 33 (2020) 15639–15650.
[38] C.R. Dance, T. Silander, Optimal policies for observing time series and related restless bandit problems, Journal of Machine Learning Research 20 (35) (2019) 1–93.
[39] K. Liu, Index policy for a class of partially observable Markov decision processes, 2021, arXiv preprint arXiv:2107.11939.
[40] D.B. Brown, J.E. Smith, Index policies and performance bounds for dynamic selection problems, Manage. Sci. 66 (7) (2020) 3029–3050.
[41] K.J. Astrom, Optimal control of Markov processes with incomplete state information, J. Math. Anal. Appl. 10 (1) (1965) 174–205.
[42] M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, 2014.
[43] E. Altman, Constrained Markov Decision Processes, Vol. 7, CRC Press, 1999.
[44] G.M. Lipsa, N.C. Martins, Remote state estimation with communication costs for first-order LTI systems, IEEE Trans. Automat. Control 56 (9) (2011) 2013–2025.
[45] J. Chakravorty, A. Mahajan, Fundamental limits of remote estimation of autoregressive Markov processes under communication constraints, IEEE Trans. Automat. Control 62 (3) (2017) 1109–1124.
[46] A. Mahajan, Remote estimation over control area networks, in: 2017 IEEE 86th Vehicular Technology Conference, VTC-Fall, 2017, pp. 1–5.
[47] D.I. Shuman, A. Nayyar, A. Mahajan, Y. Goykhman, K. Li, M. Liu, D. Teneketzis, M. Moghaddam, D. Entekhabi, Measurement scheduling for soil moisture sensing: From physical models to optimal control, Proc. IEEE 98 (11) (2010) 1918–1933.
[48] J. Keilson, A. Kester, Monotone matrices and monotone Markov processes, Stochastic Process. Appl. 5 (3) (1977) 231–241.
[49] L.I. Sennott, Stochastic Dynamic Programming and the Control of Queueing Systems, Vol. 504, John Wiley & Sons, 2009.
[50] N. Akbarzadeh, A. Mahajan, Partially Observable Restless Bandits with Restarts, https://codeocean.com/capsule/6654724/tree/v1.

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

Aditya Mahajan is Associate Professor in the department of Electrical and Computer Engineering, McGill University, Montreal, Canada. He is Associate Editor for IEEE Transactions on Automatic Control and Springer Mathematics of Control, Signal, and Systems. He was an Associate Editor of the IEEE Control Systems Society Conference Editorial Board from 2014 to 2017. He is the recipient of the 2015 George Axelby Outstanding Paper Award, 2014 CDC Best Student Paper Award (as supervisor), and the 2016 NecSys Best Student Paper Award (as supervisor). His principal research interests include learning and control of centralized and decentralized stochastic systems.