

TWO FAMILIES OF INDEXABLE PARTIALLY OBSERVABLE RESTLESS BANDITS AND WHITTLE INDEX COMPUTATION

NIMA AKBARZADEH,* *McGill University*

ADITYA MAHAJAN,* *McGill University*

Abstract

We consider the restless bandits with general 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 only if it is chosen. We assume both models satisfy the restart property under which we prove indexability of the models and propose the Whittle index policy as the solution. 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.

Keywords: Multi-armed bandits; Restless bandits; Whittle index; indexability; partially observable; scheduling; resource allocation.

2020 Mathematics Subject Classification: Primary 90C40

Secondary 90C39;49M20;91B32

1. Introduction

Resource allocation and scheduling problems arise in various applications including telecommunication networks, patient prioritization, machine maintenance, and sensor management. Identifying the optimal policy in such models suffers from the curse of dimensionality because the state space is exponential in the number of alternative. Restless bandits is a widely-used solution framework for such models [1, 2, 4, 14, 16–18, 20, 23, 28, 32, 33].

The key idea behind the restless bandit solution framework is as follows. For each alternative or arm, we assign an index (called the Whittle index) to each state and then, at each time, sort the arms accordingly to the Whittle index of their current state and play the arms with top- m indices. The resulting policy is called the Whittle index policy [35].

The key features of the Whittle index policy are as follows. 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 [12, 21, 22, 34] and, in general, it performs close to optimal in many instances [5, 6, 10, 14, 15, 25].

* Postal address: Department of Electrical and Computer Engineering, McGill University, 3480 Rue University, Montréal, QC H3A 0E9.

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

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 for specific models [6, 7, 9, 10, 13–15, 36]. Second, for some models, there are closed-form expressions to compute the Whittle index [1, 3, 5, 13–15, 17, 18, 20] but, 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 [24, 25]. Recently, [3] presented a generalization of adaptive greedy algorithm which is applicable to all indexable restless bandits.

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 [3, 13–15, 29, 34, 35] 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, 2, 16–18, 23, 32], and some time, it is also assumed that the two states are positively correlated [1, 17, 18]. There are very few results for general state space models under partial observability [4, 11, 20, 28], and, for such models, indexability is often verified numerically. In addition, there are very few algorithms to compute the Whittle index for such models.

The main contributions of our paper are as follows:

- We investigate partial observable restless bandits with general 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 [3] 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 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 \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 $\mathbb{Z}_{\geq 0}$ as the set of nonnegative integers. Calligraphic alphabets are used to denote sets, bold variables are

used for the vector of variables. For a finite set \mathcal{X} , $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions on \mathcal{X} . Superscript i is used to index arms and subscript t is used for time t and subscript $0:t$ shows the history of the variable from time 0 up to time t .

Given ordered sets \mathcal{X} and \mathcal{Y} , a function $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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 $z \in \mathcal{X}$.

Given a set \mathcal{Z} , $\text{span}(\mathcal{Z})$ denotes the span-norm of the set.

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 $(\mathcal{X}, \{0, 1\}, \{P(a)\}_{a \in \{0,1\}}, c, \pi_0, \mathcal{Y})$ 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; $P(a)$, $a \in \{0, 1\}$, denotes the transition matrix when action a is chosen; $c : \mathcal{X} \times \{0, 1\} \rightarrow \mathbb{R}_{\geq 0}$ denotes the cost function; π_0 denotes the initial state distribution.

In this paper, we assume that the transitions under active action satisfy the *restart property*, i.e., $P_x(1) = Q$, for all $x \in \mathcal{X}$, where Q is a known probability mass function (pmf). 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 $Y_t^i = \mathfrak{E}$, 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.,

$$Y_{t+1}^i = \begin{cases} \mathfrak{E} & \text{if } A_t^i = 0 \\ X_{t+1}^i & \text{if } A_t^i = 1 \end{cases}, \quad i \in \mathcal{N}, \quad (1)$$

We use \mathcal{Y}^i to denote the observation alphabet for arm i . For model A, $\mathcal{Y}^i = \{\mathfrak{E}\}$ and for model B, $\mathcal{Y}^i = \mathcal{X} \cup \{\mathfrak{E}\}$, for all $i \in \mathcal{N}$.

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 $(\mathcal{X}^i, \{0, 1\}, \{P^i(a)\}_{a \in \{0,1\}}, c^i, \pi_0^i)$, $i \in \mathcal{N} := \{1, \dots, n\}$.

Let $\mathcal{X} := \prod_{i \in \mathcal{N}} \mathcal{X}^i$, $\mathcal{A}(m) := \left\{ (a^1, \dots, a^n) \in \{0, 1\}^n : \sum_{i \in \mathcal{N}} a^i \leq m \right\}$, and $\mathcal{Y} := \prod_{i \in \mathcal{N}} \mathcal{Y}^i$ denote the combined state, action, and observation spaces, respectively. Also, let $\mathbf{X}_t = (X_t^1, \dots, X_t^n) \in \mathcal{X}$, $\mathbf{A}_t = (A_t^1, \dots, A_t^n) \in \mathcal{A}(m)$, and $\mathbf{Y}_t = (Y_t^1, \dots, Y_t^n) \in \mathcal{Y}$ 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 $\mathbf{x}_{0:t}$ of $\mathbf{X}_{0:t}$ and $\mathbf{a}_{0:t}$ of $\mathbf{A}_{0:t}$, we have

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

$$= \prod_{i \in \mathcal{N}} P_{x_t^i, x_{t+1}^i}^i(a_t^i).$$

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

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

where \mathbf{g}_t is the (history dependent) policy at time t . Let $\mathbf{g} = (g_1, g_2, \dots)$ denote the policy for infinite time horizon and let \mathcal{G} denote the family of all such policies. Let $\boldsymbol{\pi}_0 = \bigotimes_{i \in \mathcal{N}} \pi_0^i$ denote the initial state distribution of all arms. Then, the performance of policy \mathbf{g} is given by

$$J^{(\mathbf{g})}(\boldsymbol{\pi}_0) := (1 - \beta) \mathbb{E} \left[\sum_{t=0}^{\infty} \beta^t \sum_{i \in \mathcal{N}} c^i(X_t^i, A_t^i) \middle| \begin{array}{l} X_0^i \sim \pi_0^i, \\ \forall i \in \mathcal{N} \end{array} \right], \quad (3)$$

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 $\{(\mathcal{X}^i, \{0, 1\}, \mathcal{Y}^i, P^i(a), c^i, f^i, \pi_0^i)\}_{i \in \mathcal{N}}$ of each arm, and the observation model at the operator, choose a Markov policy $\mathbf{g} \in \mathcal{G}$ that minimizes $J^{(\mathbf{g})}(\boldsymbol{\pi}_0)$ given by (3).

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 [8].

2.3. Belief State

Let us 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 $\Pi_t^i(x_t^i) := \mathbb{P}(X_t^i = x_t^i | Y_{0:t-1}^i, A_{0:t-1}^i)$. 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

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

and for model B, the belief update rule is

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

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 $\Theta_t \in \mathcal{P}(\mathcal{X})$ of the system as follows: for any $\mathbf{x} \in \mathcal{X}$,

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

Note that Θ_t is a random variable that takes values in $\mathcal{P}(\mathcal{X})$. Using standard results in POMDPs [8], we have the following.

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

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

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

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

Therefore, there is no loss of optimality in restricting attention to decision policies of the form $\mathbf{A}_t = g_t^{\text{simple}}(\Pi_t)$. 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. \square

In Propositions 1 and 2, we do not present the DPs because they suffer from the curse of dimensionality. In particular, obtaining the optimal policy for PO-RMAB is PSPACE-hard [26]. 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 $(\mathcal{X}, \{0, 1\}, \{P(a)\}_{a \in \{0, 1\}}, c, \pi_0, \mathcal{Y})$ with a modified per-step cost function

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

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

$$J_\lambda^{(g)}(\pi_0) := (1 - \beta) \mathbb{E} \left[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(\Pi_t, g(\Pi_t)) \mid X_0 \sim \pi_0 \right]. \quad (8)$$

Subsequently, consider the following optimization problem.

Problem 2. Given an arm $(\mathcal{X}, \mathcal{Y}, \{0, 1\}, \{P(a)\}_{a \in \{0, 1\}}, c, \pi_0)$, the discount factor $\beta \in (0, 1)$ and the penalty $\lambda \in \mathbb{R}$, choose a Markov policy $g : \mathcal{P}(\mathcal{X}) \rightarrow \{0, 1\}$ to minimize $J_\lambda^{(g)}(\pi_0)$ given by (8).

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

Proposition 3. *Let $V_\lambda : \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R}$ be the unique fixed point of equation*

$$V_\lambda(\pi) = \min \left\{ (1 - \beta) \bar{c}(\pi, 0) + \beta V_\lambda(\pi P), (1 - \beta) \bar{c}(\pi, 1) + (1 - \beta) \lambda + \beta V_\lambda(Q) \right\} \quad (9)$$

for Model A and the unique fixed point of equation

$$V_\lambda(\pi) = \min \left\{ (1 - \beta)\bar{c}(\pi, 0) + \beta V_\lambda(\pi P), (1 - \beta)\bar{c}(\pi, 0) + \beta \sum_{x \in \mathcal{X}} Q_x V_\lambda(\delta_x) \right\} \quad (10)$$

for Model B. Let $g_\lambda(\pi)$ denote the arg min of the right hand side of (9) for Model A and the arg min of the right hand side of (10) for Model B. We set $g_\lambda(\pi) = 1$ if the two argument inside $\min\{\cdot, \cdot\}$ are equal. Then, the time-homogeneous policy g_λ is optimal for Problem 2.

Proof. The result follows immediately from Markov decision theory [27]. \square

Finally, we present the following definitions.

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

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

Definition 2. (*Indexability.*) an arm is indexable if \mathcal{W}_λ is weakly increasing in λ , i.e., for any $\lambda_1, \lambda_2 \in \mathbb{R}$,

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

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

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

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

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 [12, 21, 22, 34] and performs close to optimal for many other cases [5, 6, 10, 14, 15, 25].

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

Given an arm, let Σ denote the family of all stopping times 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

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

$$B(\pi, \tau) := \mathbb{E}[\beta^\tau \mid \Pi_0 = \pi].$$

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*

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

where

$$G(\pi) := (1 - \beta) \inf_{\tau \in \Sigma} \frac{L(\pi, \tau) - \bar{c}(\pi, 1)}{1 - B(\pi, \tau)}, \quad (11)$$

$$W_\lambda := \lambda + \beta V_1^{\text{NEXT}} \quad (12)$$

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

Proof. First, we assert that $V_\lambda(\pi)$ and W_λ are strictly increasing in λ for any $\pi \in \Pi$ which hold due to the fact that $\bar{c}_\lambda(\pi, a)$ is increasing in λ , $\pi \in \Pi$ and $a \in \{0, 1\}$. From [5, Lemma 2], we know that the passive set

$$\mathcal{W}_\lambda = \{ \pi \in \Pi : G(\pi) < W_\lambda \}. \quad (13)$$

Note that $G(\pi)$ does not depend on λ while we showed that W_λ is strictly increasing in λ . Hence, Π_λ is increasing in λ . Thus arm i is indexable. The expression for the Whittle index in the Theorem 1 follows immediately from (13). \square

5. Whittle index computation

Computing the Whittle index using the belief state representation is intractable in general. Inspired by the approach taken in [31], we introduce a new information state which is equivalent to the belief state.

5.1. 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 1. *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

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

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

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

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 $\mathbf{S}_t := (S_t^1, \dots, S_t^n)$ and $\mathbf{K}_t := (K_t^1, \dots, K_t^n)$. The relation between the belief state Π_t and variables S_t and K_t is characterized in the following lemma.

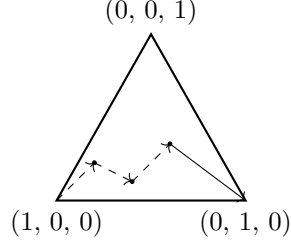


FIGURE 1: Belief state dynamics for a 3-state arm i in the simplex $\mathcal{P}(\{1, 2, 3\})$. Dashed arrows show a sample realizations of the belief state evolution under $A_t = 0$ for three time steps and the solid arrow shows a sample realization of the belief state evolution under $A_t = 1$.

Lemma 1. *The following statements hold under Assumption 1:*

- For model A, for any $i \in \mathcal{N}$ and any t , $\Pi_t \in \mathcal{R}_A$. In particular, $\Pi_t = QP^{K_t}$.
- For model B, for any $i \in \mathcal{N}$ and 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). \square

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

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

and the total expected per-step cost incurred at time t may be written as $\bar{c}(\mathbf{K}_t, \mathbf{A}_t) := \sum_{i=1}^n \bar{c}(K_t, A_t)$.

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

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

and the total expected per-step cost incurred at time t may be written as $\bar{c}(\mathbf{S}_t, \mathbf{K}_t, \mathbf{A}_t) := \sum_{i=1}^n \bar{c}(S_t, K_t, A_t)$.

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

Proof. This result immediately follows from Lemma 1, (16) and (17). \square

Next, consider the following assumption on the per-step cost function.

Assumption 2. *Let $c(x, a) = (1 - a)\phi(x) + a\rho(x)$ where $\phi : \mathcal{X} \rightarrow [0, \phi_{\max})$ and $\rho : \mathcal{X} \rightarrow [0, \rho_{\max})$ are increasing functions in \mathcal{X} and $c(x, a)$ is submodular in (x, a) .*

Under Assumption 2, we derive structural properties of the optimal policies for models A and B. Then, we show how the performance measure can be decomposed and computed. Next, we apply a finite state approximation to restrict the set of possible information states and make the computations feasible, and ultimately, we provide the Whittle index formula for model A and present an adaptive greedy algorithm to compute the Whittle indices for model B.

5.2. Structural properties of the optimal policy

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 Assumption 2, 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*

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

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*

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

We use θ^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 θ^B instead of $g^{(\theta^A)}$ and $g^{(\theta^B)}$.

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.,

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

where

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

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

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

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

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

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

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.,

$$J_\lambda^{(\theta^B)}(s, k) = (1 - \beta) \mathbb{E} \left[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(S_t, K_t, g^{(\theta^B)}(S_t, K_t)) \mid (S_0, K_0) = (s, k) \right]$$

$$=: D^{(\theta^B)}(s, k) + \lambda N^{(\theta^B)}(s, k), \quad (19)$$

where

$$D^{(\theta^B)}(s, k) := (1 - \beta) \mathbb{E} \left[\sum_{t=0}^{\infty} \beta^t \bar{c}(S_t, K_t, g^{(\theta^B)}(S_t, K_t)) \mid (S_0, K_0) = (s, k) \right],$$

$$N^{(\theta^B)}(s, k) := (1 - \beta) \mathbb{E} \left[\sum_{t=0}^{\infty} \beta^t g^{(\theta^B)}(S_t, K_t) \mid (S_0, K_0) = (s, k) \right].$$

$D^{(\theta^B)}(s, k)$ 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's define vector $\mathbf{J}_\lambda^{(\theta^B)}(0) = (J_\lambda^{(\theta^B)}(1, 0), \dots, J_\lambda^{(\theta^B)}(|\mathcal{X}|, 0))$ and vectors $\mathbf{D}^{(\theta^B)}(0)$ and $\mathbf{N}^{(\theta^B)}(0)$ in a similar manner. Then, from (18), $\mathbf{J}_\lambda^{(\theta^B)}(0) = \mathbf{D}^{(\theta^B)}(0) + \lambda \mathbf{N}^{(\theta^B)}(0)$. Let's also define

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

$$M^{(\theta^B)}(s, k) := (1 - \beta) \beta^{\theta_s^B-k}.$$

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

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,

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

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 $\mathbb{N}_\ell := \{0, \dots, \ell\}$ and $V_{\ell, \lambda} : \mathbb{N}_\ell \rightarrow \mathbb{R}$ be the unique fixed point of equation

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

where

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

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

$$|V_\lambda(k) - V_{\ell, \lambda}(k)| \leq \frac{\beta^{\ell - k + 1} \text{span}(c_\lambda)}{1 - \beta}.$$

(ii) For all $k \in \mathbb{Z}_{\geq 0}$, $\lim_{\ell \rightarrow \infty} V_{\ell, \lambda}(k) = V_\lambda(k)$. 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.

Theorem 6. (Model B.) Given $\ell \in \mathbb{N}$, let $\mathbb{N}_\ell := \{0, \dots, \ell\}$ and $V_{\ell, \lambda} : \mathcal{X} \times \mathbb{N}_\ell \rightarrow \mathbb{R}$ 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

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

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 $0 \leq k \leq \ell$,

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

(ii) For all $(s, k) \in \mathcal{X} \times \mathbb{Z}_{\geq 0}$, $\lim_{\ell \rightarrow \infty} V_{\ell,\lambda}(s, k) = V_\lambda(s, k)$. 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.

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

5.5. 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 $\Lambda_k^A = \{k_0 \in \{0, 1, \dots, (\ell + 1) - 1\} : N^{(k)}(k_0) \neq N^{(k+1)}(k_0)\}$. Then, under Assumption 2, $\Lambda_k^A \neq \emptyset$ and for any $k_0 \in \Lambda_k^A$, the Whittle index of model A at information state $k \in \mathbb{N}_\ell$ is

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

Proof. Since model A is a restart model, the result follows from [3, Lemma 4]. \square

Theorem 7 gives us a closed-form expression to 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 $\lambda_1 < \lambda_2 < \dots < \lambda_{B_D}$ 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 $\mathcal{S} \subseteq \mathcal{X} \times \mathbb{N}_\ell$, define the policy $\bar{g}^{(\mathcal{S})} : \mathcal{X} \times \mathbb{N}_\ell \rightarrow \{0, 1\}$ as

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

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 $\Gamma_{b+1} = \mathcal{W}_{b+1} \setminus \mathcal{W}_b$. 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

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

Lemma 2. For $d \in \{0, \dots, B_D - 1\}$, we have the following:

1. For all $y \in \Gamma_{b+1}$, we have $w(y) = \lambda_{b+1}$.

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 .
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 $b = b + 1$.

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 [3, 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 $y \in \Phi_b$. \square

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 [3, Theorem 2]. Similar to Lemma 2, we consider $y \in \Phi_b$. \square

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 [3, 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 $x \in \tilde{\mathcal{X}}$, $\sum_{z \in \tilde{\mathcal{X}}_{\geq x}} \mu_z^1 \geq \sum_{z \in \tilde{\mathcal{X}}_{\geq x}} \mu_z^2$. 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 row of N . A basic property of stochastic dominance is the following.

Lemma 3. *If M^1 stochastically dominates M^2 and c is an increasing function defined on $\tilde{\mathcal{X}}$, then for all $x \in \tilde{\mathcal{X}}$, $\sum_{y \in \tilde{\mathcal{X}}} M_{xy}^1 c(y) \geq \sum_{y \in \tilde{\mathcal{X}}} M_{xy}^2 c(y)$.*

Proof. This is an induction from [27, Lemma 4.7.2]. \square

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 [3], 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 [3, Lemma 5].

Proposition 5. *The optimal policy of a stochastic monotone fully-observable restless bandit process is a threshold policy denoted by \tilde{g}_2 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.,*

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

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 $P_{xy}^A = \mathbb{I}_{\{y=x+1\}}$ and the transition probability matrix under active action for model A is $Q_{xy}^A = \mathbb{I}_{\{y=0\}}$. 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 $z \in \mathbb{N}_\ell$.
- (D3') As P stochastically dominates the identity matrix, we infer from [19, 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 increasing 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 3, $\bar{c}_\lambda(k, a)$ 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 3, $\bar{c}_\lambda(k, 0) - \bar{c}_\lambda(k, 1) = \sum_{x \in \mathcal{X}} [(QP)^k]_x (c_\lambda(x, 0) - c_\lambda(x, 1))$ 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.

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 4. *We have*

- a. $\bar{c}_\lambda(s, k, a)$ is increasing in k for any $s \in \mathcal{X}$ and $a \in \{0, 1\}$.
- b. Given a fixed λ , $V_\lambda(s, k)$ is increasing 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 increasing in k and x and as $c(x, a)$ is increasing in x , $\bar{c}(s, k, a)$ is increasing in k .

b. Let

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

where $V_\lambda^0(\cdot, \cdot) = 0$ for all $(s, k) \in \mathcal{X} \times \mathbb{Z}_{\geq 0}$.

Claim: $V_\lambda^j(s, k)$ is non-decreasing in k for any $s \in \mathcal{X}$ and $j \geq 0$.

We prove the claim by induction. By construction, $V_\lambda^0(s, k)$ is non-decreasing in k for any $s \in \mathcal{X}$. This forms the basis of induction. Now assume that $V_\lambda^j(s, k)$ 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

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

Therefore,

$$V_\lambda^{j+1}(s, \ell) = \min_a \{H_\lambda^j(s, \ell, a)\} \geq \min_a \{H_\lambda^j(s, k, a)\} = V_\lambda^{j+1}(s, k).$$

Thus, $V_\lambda^{j+1}(s, k)$ is non-decreasing in k for any $s \in \mathcal{X}$. This completes the induction step. $V_\lambda(s, k) = \lim_{j \rightarrow \infty} V_\lambda^j(s, k)$ 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 s^{th} row of P^k . Thus, $\delta_s P^{k+1}$ stochastically dominates $\delta_s P^k$ and by Lemma 3 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 increasing 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 5. *Suppose $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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 [27, Lemma 4.7.1]. □

Finally, we conclude that as $H_\lambda(s, k, a)$ is submodular in (k, a) for any $s \in \mathcal{X}$, then, based on Lemma 5 and as only two actions is available, the optimal policy is a threshold policy specified in the theorem statement.

6.2. Proof of Theorem 3

By the strong Markov property, we have

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

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

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

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

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

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

which results in

$$\begin{aligned}
\mathbf{D}^{(\theta^B)}(0) &= \mathbf{L}^{(\theta^B)}(0) + \mathbf{Z}^{(\theta^B)} \mathbf{D}^{(\theta^B)}(0), \\
\mathbf{N}^{(\theta^B)}(0) &= \mathbf{M}^{(\theta^B)}(0) + \mathbf{Z}^{(\theta^B)} \mathbf{N}^{(\theta^B)}(0)
\end{aligned}$$

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 $g_\lambda^A(\cdot)$ for information states $\{k, \dots, \ell\}$. The per-step cost incurred by $\hat{g}_{\ell, \lambda}(\cdot)$ differs from $g_\lambda^A(\cdot)$ for information states $\{\ell + 1, \dots\}$ by at most $\text{span}(c_\lambda)$.

(ii): The sequence of finite-state models described above is an *augmentation type approximation sequence*. As a result, a limit point of \hat{g}_λ^* exists and the final result holds by [30, 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 $g_\lambda^B(\cdot, \cdot)$ for information states $\{(s, l)\}_{l=k}^\ell$. The per-step cost incurred by $\hat{g}_{\ell, \lambda}(\cdot, \cdot)$ differs from $g_\lambda^B(\cdot, \cdot)$ 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}_λ^* exists and the final result holds [30, Proposition B.5, Theorem 4.6.3].

7. Numerical Analysis

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 un-monitored, then the state of the machine deteriorates and after a while, it ruins. However, the state of the machine is not observed. There is a cost associated with running the machine and the cost is a non-decreasing function of the state.

Furthermore, we assume the company cannot observe the state of the machines unless it sends a service-person to visit the machine. We assume that replacing the machine is relatively inexpensive, and when a service-person visits a machine, he simply replaces 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 $m < n$ service-persons. We are interested in determining a scheduling policy to decide which machines should be serviced at each time.

7.1. Policies Compared

We 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 $\ell < m$ machines have been selected. Then select machine $\ell + 1$ 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 $\ell = 0$.
 while $\ell \leq m$ **do**
 Compute $i_\ell^* \in \arg \min_{i \in \mathcal{Z}} \sum_{j \in \mathcal{Z} \setminus \{i\}} \bar{c}^j(S_t^j, K_t^j, 0) + \bar{c}^i(S_t^i, K_t^i, 1)$.
 Let $\mathcal{M} = \mathcal{M} \cup \{i_\ell^*\}$, $\mathcal{Z} = \mathcal{Z} \setminus \{i_\ell^*\}$.
 Set $\ell = \ell + 1$.
 Service the machines with indices collected in \mathcal{M} .
 Update K_t^i according to (14) and S_t^i according to (15) for all $i \in \mathcal{N}$.
 Set $t = t + 1$.

7.2. 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

TABLE 1: α_{OPT} for different choice of parameters in Experiment 1.

(a) Model A					(b) Model B				
ℓ	1	2	3	4	ℓ	1	2	3	4
α_{OPT}	100.0	100.0	100.0	100.0	α_{OPT}	100.0	99.72	99.81	99.57

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 $\mathcal{P}_\ell(p)$, $\ell \in \{1, 2, 3, 4\}$ where p is a parameter of the model. The details of all these models are presented in Appendix A. We assume each element of Q^i is sampled from $\text{Exp}(1)$, i.e., exponential distribution with the rate parameter of 1, and then normalized such that 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.

Experiment 1) Comparison of Whittle index with the optimal policy. In this experiment, we compare the performance of WIP with OPT. We assume $|\mathcal{X}| = 4$, $(\ell + 1) = 4$ 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 $\mathcal{P}_\ell(p_i)$ 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 prospective of the performances, we compute the relative performance of WIP with respect to OPT by computing

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

The closer α is to 100, the closer WIP is to OPT. The results of α_{OPT} for different choice of the parameters are shown in Table 1.

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 prospective of the performances, we compute the relative improvement of WIP with respect to MYP by computing

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

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

TABLE 2: ε_{MYP} for different choice of parameters of Model A in Experiment 2.

(a) Model A, $m = 1$					(b) Model A, $m = 5$				
ε_{MYP}	ℓ				ε_{MYP}	ℓ			
	1	2	3	4		1	2	3	4
20	1.42	3.20	2.04	6.47	20	0.15	0.27	0.22	1.59
n 40	2.45	5.62	4.82	7.09	n 40	1.09	1.28	1.13	3.79
60	2.68	4.40	4.33	5.30	60	1.38	2.17	2.14	7.27

TABLE 3: ε_{MYP} for different choice of parameters of Model B in Experiment 2.

(a) Model B, $m = 1$					(b) Model B, $m = 5$				
ε_{MYP}	ℓ				ε_{MYP}	ℓ			
	1	2	3	4		1	2	3	4
20	7.88	11.4	9.66	10.2	20	0.77	1.43	0.88	3.72
n 40	12.1	14.6	13.4	7.19	n 40	1.49	3.96	3.76	8.59
60	14.5	12.9	11.8	6.06	60	4.13	5.45	4.92	8.37

heterogeneous machines. The results of ε_{MYP} for different choice of the parameters for models A and B are shown in Tables 2 and 3, respectively.

7.3. 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 policies is more than model A.

In Experiment 2 where WIP is compared with MYP, we observe ε_{MYP} ranges from 0.15 to 14.5. 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 [3] 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.

Appendix A. 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,

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

Funding information

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

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

- [1] AALTO, S., LASSILA, P. AND OSTI, P. (2016). Whittle index approach to size-aware scheduling for time-varying channels with multiple states. *Queueing Systems* **83**, 195–225.
- [2] ABAD, C. AND IYENGAR, G. (2016). A near-optimal maintenance policy for automated DR devices. *IEEE Transactions on Smart Grid* **7**, 1411–1419.
- [3] AKBARZADEH, N. AND MAHAJAN, A. Conditions for indexability of restless bandits and an algorithm to compute Whittle index. *Journal of Applied Probability* (in print).
- [4] AKBARZADEH, N. AND MAHAJAN, A. (2019). Dynamic spectrum access under partial observations: A restless bandit approach. In *Canadian Workshop on Information Theory*. IEEE. pp. 1–6.
- [5] AKBARZADEH, N. AND MAHAJAN, A. (2019). Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *Conference on Decision and Control*. pp. 7294–7300.
- [6] ANSELL, P. S., GLAZEBROOK, K. D., NIÑO-MORA, J. AND O’KEEFFE, M. (2003). Whittle’s index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research* **57**, 21–39.
- [7] ARCHIBALD, T. W., BLACK, D. P. AND GLAZEBROOK, K. D. (2009). Indexability and index heuristics for a simple class of inventory routing problems. *Operations research* **57**, 314–326.
- [8] ASTROM, K. J. (1965). Optimal control of Markov processes with incomplete state information. *Journal of mathematical analysis and applications* **10**, 174–205.
- [9] AVRACHENKOV, K., AYESTA, U., DONCEL, J. AND JACKO, P. (2013). Congestion control of TCP flows in internet routers by means of index policy. *Computer Networks* **57**, 3463–3478.
- [10] AYESTA, U., ERAUSQUIN, M. AND JACKO, P. (2010). A modeling framework for optimizing the flow-level scheduling with time-varying channels. *Performance Evaluation* **67**, 1014–1029.
- [11] DANCE, C. R. AND SILANDER, T. (2019). Optimal policies for observing time series and related restless bandit problems. *J. Mach. Learn. Res.* **20**, 35–1.
- [12] GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)* 148–177.
- [13] GLAZEBROOK, K., HODGE, D. AND KIRKBRIDE, C. (2013). Monotone policies and indexability for bidirectional restless bandits. *Advances in Applied Probability* **45**, 51–85.
- [14] GLAZEBROOK, K. D., MITCHELL, H. M. AND 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**, 267–284.

- [15] GLAZEBROOK, K. D., RUIZ-HERNANDEZ, D. AND KIRKBRIDE, C. (2006). Some indexable families of restless bandit problems. *Advances in Applied Probability* **38**, 643–672.
- [16] GUHA, S., MUNAGALA, K. AND SHI, P. (2010). Approximation algorithms for restless bandit problems. *Journal of the ACM (JACM)* **58**, 3.
- [17] KAZA, K., MEHTA, V., MESHAM, R. AND MERCHANT, S. (2018). Restless bandits with cumulative feedback: Applications in wireless networks. In *Wireless Communications and Networking Conference*. IEEE. pp. 1–6.
- [18] KAZA, K., MESHAM, R., MEHTA, V. AND MERCHANT, S. N. (2019). Sequential decision making with limited observation capability: Application to wireless networks. *IEEE Transactions on Cognitive Communications and Networking* **5**, 237–251.
- [19] KEILSON, J. AND KESTER, A. (1977). Monotone matrices and monotone Markov processes. *Stochastic Processes and their Applications* **5**, 231–241.
- [20] LARRAÑAGA, M., ASSAAD, M., DESTOUNIS, A. AND PASCHOS, G. S. (2016). Dynamic pilot allocation over Markovian fading channels: A restless bandit approach. In *Information Theory Workshop*. IEEE. pp. 290–294.
- [21] LIU, K. AND ZHAO, Q. (2010). Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. **56**, 5547–5567.
- [22] LOTT, C. AND 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**, 259–297.
- [23] MESHAM, R., MANJUNATH, D. AND GOPALAN, A. (2018). On the Whittle index for restless multiarmed hidden Markov bandits. **63**, 3046–3053.
- [24] NIÑO-MORA, J. (2001). Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability* **33**, 76–98.
- [25] NIÑO-MORA, J. (2007). Dynamic priority allocation via restless bandit marginal productivity indices. *TOP* **15**, 161–198.
- [26] PAPADIMITRIOU, C. H. AND TSITSIKLIS, J. N. (1999). The complexity of optimal queuing network control. *Mathematics of Operations Research* **24**, 293–305.
- [27] PUTERMAN, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
- [28] QIAN, Y., ZHANG, C., KRISHNAMACHARI, B. AND TAMBE, M. (2016). Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Int. Conf. on Autonomous Agents & Multiagent Systems*. pp. 123–131.
- [29] RUIZ-HERNÁNDEZ, D., PINAR-PÉREZ, J. M. AND DELGADO-GÓMEZ, D. (2020). Multi-machine preventive maintenance scheduling with imperfect interventions: A restless bandit approach. *Computers & Operations Research* **119**, 104927.

- [30] SENNOTT, L. I. (2009). *Stochastic dynamic programming and the control of queueing systems* vol. 504. John Wiley & Sons.
- [31] SHUMAN, D. I., NAYYAR, A., MAHAJAN, A., GOYKHMAN, Y., LI, K., LIU, M., TENEKETZIS, D., MOGHADDAM, M. AND ENTEKHABI, D. (2010). Measurement scheduling for soil moisture sensing: From physical models to optimal control. *Proceedings of the IEEE* **98**, 1918–1933.
- [32] VILLAR, S. S. (2016). Indexability and optimal index policies for a class of reinitialising restless bandits. *Probability in the engineering and informational sciences* **30**, 1–23.
- [33] VILLAR, S. S., BOWDEN, J. AND WASON, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics* **30**, 199.
- [34] WEBER, R. R. AND WEISS, G. (1990). On an index policy for restless bandits. *Journal of Applied Probability* **27**, 637–648.
- [35] WHITTLE, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability* **25**, 287–298.
- [36] YU, Z., XU, Y. AND TONG, L. (2018). Deadline scheduling as restless bandits. **63**, 2343–2358.