Partially observable restless bandits with restarts: indexability and computation of Whittle index

Nima Akbarzadeh, Aditya Mahajan
McGill University

***Abstract*— We consider restless bandits with restarts, where the state of the active arms resets according to a known probability distribution while the state of the passive arms evolves in a Markovian manner. We assume that the state of the arm is observed after it is reset but not observed otherwise. We show that the model is indexable and propose an efficient algorithm to compute the Whittle index by exploiting the qualitative properties of the optimal policy. A detailed numerical study of machine repair models shows that Whittle index policy outperforms myopic policy and is close to optimal policy.**

I. 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]–[12].

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 [13].

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 [14]–[17] and, in general, it performs close to optimal in many instances [10], [18]–[21].

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 [10], [19], [20], [22]–[25]. Second, for some models, there are closed-form expressions to compute the Whittle index [3]–[6], [10], [21], [24], [26] 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 [18], [27]. Recently, [26] 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 [10], [13], [15], [24], [26], [28] 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], and some time, it is also assumed that the two states are positively correlated [3]–[5]. There are very few results for general state space models under partial observability [6], [7], [12], [29], [30], and, for such models, indexability is often verified numerically.

We focus on a class of partially observable restless multi-armed bandits where choosing an arm resets its state. This property was considered in [21] and has applications in healthcare and machine maintenance [10], [30].

The main contributions of our paper are as follows:

- We investigate partially observable restless bandits with restart and show that the model is indexable.
- We provide a refinement of the adaptive greedy algorithm of [26] 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 II, 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 III, we present a short overview of restless bandits. In Section IV, we show the restless bandit problem is indexable for both models and present a general formula to compute the index. In Section V, we present a countable state representation of the belief state and use it to develop methods to compute Whittle index. In Section VI, we present a detailed numerical study which compares the performance of Whittle index policy with two baseline policies. Finally, we conclude in Section VII.

A. 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 $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ 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 $z \in \mathcal{X}$.

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

II. Model and Problem Formulation

A. 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; $c : \mathcal{X} \times \{0,1\} \to \mathbb{R}_{\geq 0}$ 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., $P_{x\cdot}(1) = Q$, 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 assume that the operator observes the state of the arm after it has been reset, i.e.,

$$Y^i_{t+1} = \begin{cases} \mathfrak{E} & \text{if } A^i_t = 0 \\ X^i_{t+1} & \text{if } A^i_t = 1 \end{cases}, \quad i \in \mathcal{N}, \tag{1}$$

We use $\mathcal{Y}^i = \mathcal{X}^i \cup \{\mathfrak{E}\}$ to denote the observation alphabet for arm i.

B. Partially-observable Restless Multi-armed Bandit Problem

A partially-observable restless multi-armed bandit (PO-RMAB) problem is a collection of n independent arms $(\mathcal{X}^i, \{0,1\}, \{P^i(a)\}_{a\in\{0,1\}}, c^i, \pi^i_0)$, $i \in \mathcal{N} := \{1, \ldots, n\}$.

Let $\boldsymbol{\mathcal{X}} := \prod_{i\in\mathcal{N}} \mathcal{X}^i$, $\boldsymbol{\mathcal{A}}(m) := \big\{(a^1, \ldots, a^n) \in \{0,1\}^n : \sum_{i\in\mathcal{N}} a^i \leq m\big\}$, and $\boldsymbol{\mathcal{Y}} := \prod_{i\in\mathcal{N}} \mathcal{Y}^i$ denote the combined state, action, and observation spaces, respectively. Also, let $\boldsymbol{X}_t = (X^1_t, \ldots X^n_t) \in \boldsymbol{\mathcal{X}}$, $\boldsymbol{A}_t = (A^1_t, \ldots, A^n_t) \in \boldsymbol{\mathcal{A}}(m)$, and $\boldsymbol{Y}_t = (Y^1_t, \ldots Y^n_t) \in \boldsymbol{\mathcal{Y}}$ 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

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

When the system is in state $\boldsymbol{x}_t$ and take action $\boldsymbol{a}_t$, the system incurs a cost $\boldsymbol{c}(\boldsymbol{x}_t, \boldsymbol{a}_t) := \sum_{i\in\mathcal{N}} c^i(x^i_t, a^i_t)$. The decision 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 the (history dependent) policy at time t. Let $\boldsymbol{g} = (g_1, g_2, \ldots)$ denote the policy for infinite time horizon and let $\boldsymbol{\mathcal{G}}$ denote the family of all such policies. Let $\boldsymbol{\pi}_0 = \bigotimes_{i\in\mathcal{N}} \pi^i_0$ denote the initial state distribution of all arms. Then, the performance of policy $\boldsymbol{g}$ is given by

$$J^{(\boldsymbol{g})}(\boldsymbol{\pi}_0) := (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \sum_{i\in\mathcal{N}} c^i(X^i_t, A^i_t) \middle| \begin{matrix} X^i_0 \sim \pi^i_0, \\ \forall i \in \mathcal{N} \end{matrix}\right], \tag{3}$$

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 $\{(\mathcal{X}^i, \{0,1\}, P^i(a), c^i, \pi^i_0, \mathcal{Y}^i)\}_{i\in\mathcal{N}}$ of each arm, and the observation model at the operator, choose a Markov policy $\boldsymbol{g} \in \boldsymbol{\mathcal{G}}$ that minimizes $J^{(\boldsymbol{g})}(\boldsymbol{\pi}_0)$ 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 [31].

C. Belief State

Let us define the operator's belief $\Pi^i_t \in \mathcal{P}(\mathcal{X}^i)$ on the state of arm i at time t as follows: for any, $x^i_t \in \mathcal{X}^i$, let $\Pi^i_t(x^i_t) := \mathbb{P}(X^i_t = x^i_t \,|\, Y^i_{0:t-1}, A^i_{0:t-1})$. Note that Π^i_t is a distribution-valued random variable. Also, define $\boldsymbol{\Pi}_t := (\Pi^1_t, \ldots, \Pi^n_t)$.

Then, for arm i, the evolution of the belief state is as follows:

$$\Pi^i_{t+1} = \begin{cases} \Pi^i_t P, & \text{if } A^i_t = 0, \\ \delta_{X^i_{t+1}} \text{ where } X^i_{t+1} \sim Q, & \text{if } A^i_t = 1. \end{cases} \tag{4}$$

The per-step cost function of the belief state Π^i_t when action A^i_t is taken is

$$\begin{aligned} \bar{c}(\Pi^i_t, A^i_t) &= \mathbb{E}[c^i_t(X^i_t, A^i_t) | Y^i_{0:t-1}, A^i_{0:t-1}] \\ &= \sum_{x\in\mathcal{X}^i} \Pi^i_t(x) c^i(x, A^i_t). \end{aligned}$$

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

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

Note that Θ_t is a random variable that takes values in $\mathcal{P}(\boldsymbol{\mathcal{X}})$. Using standard results in POMDPs [31], 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 $\boldsymbol{A}_t = g^{\text{belief}}_t(\Theta_t)$. 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 $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, we have

$$\Theta_t(\boldsymbol{x}) = \prod_{i \in \mathcal{N}} \Pi_t^i(x^i), \quad \text{a.s..} \tag{5}$$

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

Proof: Eq. (5) 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 DPs because they suffer from the curse of dimensionality. In particular, obtaining the optimal policy for PO-RMAB is PSPACE-hard [32]. So, we focus on the Whittle index heuristics to solve the problem.

III. 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

$$\bar{c}_\lambda(\pi, a) \coloneqq \bar{c}(\pi, a) + \lambda a, \quad \forall \pi \in \mathcal{P}(\boldsymbol{\mathcal{X}}), \forall a \in \{0,1\}, \lambda \in \mathbb{R}. \tag{6}$$

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

$$J_\lambda^{(g)}(\pi_0) \coloneqq (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(\Pi_t, g(\Pi_t)) \middle| X_0 \sim \pi_0\right]. \tag{7}$$

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 $g : \mathcal{P}(\boldsymbol{\mathcal{X}}) \to \{0,1\}$ to minimize $J_\lambda^{(g)}(\pi_0)$ given by (7).

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

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

$$V_\lambda(\pi) = \min\Big\{(1-\beta)\bar{c}_\lambda(\pi, 0) + \beta V_\lambda(\pi P), \\ (1-\beta)\bar{c}_\lambda(\pi, 1) + \beta \sum_{x\in\mathcal{X}} Q_x V_\lambda(\delta_x)\Big\} \tag{8}$$

Let $g_\lambda(\pi)$ denote the $\arg\min$ of the right hand side of (8). 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 [33]. ∎

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.,

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

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

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

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 $\mathcal{W}_\lambda$, 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 [14]–[17] and performs close to optimal for many other cases [10], [18]–[21].

IV. Indexability and the corresponding Whittle index

Given an arm, let Σ denote the family of all stopping times with respect to the natural filtration associated with $\{\Pi_t\}_{t\ge0}$. For any stopping time $\tau \in \Sigma$ and an initial belief state $\pi \in \Pi$, define

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

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

Theorem 1: The PO-RMAB defined in Section II 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

$$G(\pi) \coloneqq (1-\beta) \inf_{\tau\in\Sigma} \frac{L(\pi,\tau) - \bar{c}(\pi, 1)}{1 - B(\pi,\tau)}, \tag{9}$$

$$W_\lambda \coloneqq \lambda + \beta \sum_{x\in\mathcal{X}} Q_x V_\lambda(\delta_x). \tag{10}$$

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 [21, Lemma 2], we know that the passive set

$$\mathcal{W}_\lambda = \{\pi \in \Pi : G(\pi) < W_\lambda\}. \tag{11}$$

Note that $G(\pi)$ does not depend on λ while we showed that $\mathcal{W}_\lambda$ is strictly increasing in λ. Hence, $\mathcal{W}_\lambda$ is increasing in λ. Thus arm i is indexable. The expression for the Whittle index in the Theorem 1 follows immediately from (11). ∎

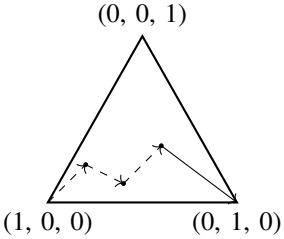

Fig. 1: Belief state dynamics for a 3-state arm i in the simplex $\mathcal{P}(\{1,2,3\})$. Dashed arrows show a sample realizations of the belief state evolution under $A_t = 0$ for three time steps and the solid arrow shows a sample realization of the belief state evolution under $A_t = 1$.

V. Whittle index computation

Computing the Whittle index using the belief state representation is intractable in general. Inspired by the approach taken in [34], we introduce a new information state which is equivalent to the belief state.

A. Information state

Assumption 1: The initial belief state $\pi_0 \in \mathcal{R} \coloneqq \{\delta_s P^k : s \in \mathcal{X}, k \in \mathbb{Z}_{\geq 0}\}$.

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

$$K_t = \begin{cases} 0, & \text{if } A_{t-1} = 1 \\ K_{t-1} + 1, & \text{if } A_{t-1} = 0 \end{cases} \tag{12}$$

and S_t evolves according to

$$S_t = \begin{cases} X_{t-1} \text{ where } X_{t-1} \sim Q, & \text{if } A_{t-1} = 1 \\ S_{t-1}, & \text{if } A_{t-1} = 0. \end{cases} \tag{13}$$

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

Lemma 1: Under Assumption 1, for any $i \in \mathcal{N}$ and any t, $\Pi_t \in \mathcal{R}$. In particular, $\Pi_t = \delta_{S_t} P^{K_t}$.

Proof: The results immediately follow from (4), (12), and (13). ∎

The expected per-step cost at time t may be written as

$$\bar{c}(S_t, K_t, A_t) \coloneqq \bar{c}(\delta_{S_t} P^{K_t}, A_t) = \sum_{x\in\mathcal{X}} [\delta_{S_t} P^{K_t}]_x c(x, A_t). \tag{14}$$

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

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

Proof: This result immediately follows from Lemma 1 and (14). ∎

Next, assume the following property holds:

(P) For every $\lambda \in \mathbb{R}$, there exists a vector $(\theta_{s,\lambda})_{s\in\mathcal{X}}$, where $\theta_{s,\lambda} \in \mathbb{Z}_{\geq -1}$ such that the semi-threshold policy

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

is optimal for Problem 2.

B. Structural properties of the optimal policy

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

Theorem 2: A sufficient condition for Property (P) to hold is the following: 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 increasing functions in $\mathcal{X}$ and $c(x,a)$ is submodular in (x,a).
The proof is omitted due to space constraints. See [35] for details.

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

C. Performance of threshold based policies

We simplify the notation and denote a threshold-based policy by $\boldsymbol{\theta}$ instead of $g^{(\boldsymbol{\theta})}$.

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

$$\begin{aligned}(1-\beta)\mathbb{E}\Bigg[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(S_t, K_t, g^{(\boldsymbol{\theta})}(S_t, K_t)) \;\Big|\; \begin{matrix} S_0 = s \\ K_0 = k \end{matrix}\Bigg] \\ =: D^{(\boldsymbol{\theta})}(s,k) + \lambda N^{(\boldsymbol{\theta})}(s,k),\end{aligned} \tag{15}$$

where $D^{(\boldsymbol{\theta})}(s,k)$ is

$$(1-\beta)\mathbb{E}\Bigg[\sum_{t=0}^{\infty} \beta^t \bar{c}(S_t, K_t, g^{(\boldsymbol{\theta})}(S_t, K_t)) \;\Big|\; (S_0, K_0) = (s,k)\Bigg],$$

and $N^{(\boldsymbol{\theta})}(s,k)$ is

$$(1-\beta)\mathbb{E}\Bigg[\sum_{t=0}^{\infty} \beta^t g^{(\boldsymbol{\theta})}(S_t, K_t) \;\Big|\; (S_0, K_0) = (s,k)\Bigg].$$

We will show (see Theorem 4) that Whittle index can be computed as a function of $D^{(\boldsymbol{\theta})}(s,k)$ and $N^{(\boldsymbol{\theta})}(s,k)$. But first let's define vector $\boldsymbol{J}_\lambda^{(\boldsymbol{\theta})}(0) = (J_\lambda^{(\boldsymbol{\theta})}(1,0), \dots, J_\lambda^{(\boldsymbol{\theta})}(|\mathcal{X}|,0))$ and vectors $\boldsymbol{D}^{(\boldsymbol{\theta})}(0)$ and $\boldsymbol{N}^{(\boldsymbol{\theta})}(0)$ in a similar manner. Then, from (15), $\boldsymbol{J}_\lambda^{(\boldsymbol{\theta})}(0) = \boldsymbol{D}^{(\boldsymbol{\theta})}(0) + \lambda \boldsymbol{N}^{(\boldsymbol{\theta})}(0)$. Let's also define

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

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

Theorem 3: For any $(s,k) \in \mathcal{X} \times \mathbb{Z}_{\geq 0}$, we have

$$D^{(\boldsymbol{\theta})}(s,k) = L^{(\boldsymbol{\theta})}(s,k) + \beta^{\theta_s - k + 1} \sum_{r\in\mathcal{X}} Q_r D^{(\boldsymbol{\theta})}(r,0),$$

$$N^{(\boldsymbol{\theta})}(s,k) = M^{(\boldsymbol{\theta})}(s,k) + \beta^{\theta_s - k+1} \sum_{r \in \mathcal{X}} Q_r N^{(\boldsymbol{\theta})}(r,0).$$

Let $Z^{(\boldsymbol{\theta})}$ be a $|\mathcal{X}| \times |\mathcal{X}|$ matrix where $Z^{(\boldsymbol{\theta})}_{sr} = \beta^{\theta_s+1} Q_r$, for any $s, r \in \mathcal{X}$. Then, $\boldsymbol{D}^{(\boldsymbol{\theta})}(0) = (I - Z^{(\boldsymbol{\theta})})^{-1} \boldsymbol{L}^{(\boldsymbol{\theta})}(0)$, and $\boldsymbol{N}^{(\boldsymbol{\theta})}(0) = (I - Z^{(\boldsymbol{\theta})})^{-1} \boldsymbol{M}^{(\boldsymbol{\theta})}(0)$.

D. Whittle index

In this section, we provide an efficient algorithm to compute the Whittle index.

1) Modified adaptive greedy algorithm. : As $K_t \in \mathbb{Z}_{\geq 0}$, let $\mathbb{N}_\ell := \{0, \ldots, \ell\}$ where $\ell \in \mathbb{N}$ denote the truncated space such that $K_t \in \mathbb{N}_\ell$. Let $B = |\mathcal{X}|(\ell+1)$ and $B_D (\leq B)$ denote the number of distinct Whittle indices. Let $\Lambda^* = \{\lambda_0, \lambda_1, \ldots, \lambda_{B_D}\}$ where $\lambda_1 < \lambda_2 < \ldots < \lambda_{B_D}$ 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 $\bar{g}^{(\mathcal{S})} : \mathcal{X} \times \mathbb{N}_\ell \to \{0,1\}$ as

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

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 $\Gamma_{b+1} = \mathcal{W}_{b+1} \backslash \mathcal{W}_b$. Additionally, for any $b \in \{0, \ldots, B_D - 1\}$, 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

$$\mu_{b,y}(x,k) = \frac{D^{(h_{b,y})}(x,k) - D^{(h_b)}(x,k)}{N^{(h_b)}(x,k) - N^{(h_{b,y})}(x,k)}. \tag{16}$$

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

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 $y \in \mathcal{W}_{b+1} \backslash \mathcal{W}_b$ and $\lambda = \lambda_{b+1}$.

Proof: The result follows from [26, 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 4: 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 [26, Theorem 2]. Similar to Lemma 2, we consider $y \in \Phi_b$. ∎

By Theorem 4, 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 [26, Algorithm 2].

Algorithm 1: Computing Whittle index of all information states

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 (16), $\forall y \in \Phi_b$.
  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 $b = b + 1$.

VI. Numerical Analysis

We conduct numerical experiments for a machine maintenance problem, and analyze how varying the number n of machines, the number m of service-persons and the parameters associated with each machine affects the performance.

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. 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. There is a cost associated with each state of the machine, and we are interested in determining a scheduling policy to decide which machines should be serviced at each time.

A. 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.

B. Experiments and Results

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 . We assume each element of Q^i is sampled from 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.

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 (12) and S_t^i according to (13) for all $i \in \mathcal{N}$.

  Set $t = t + 1$.

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

ℓ	1	2	3	4
α_{OPT}	100.0	99.72	99.81	99.57

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 $n = 3$, $m = 1$. In order to model heterogeneous machines, we consider the following. Let $(p_1, \ldots, p_n)$ 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 perspective of the performances, we compute the relative performance of WIP with respect to OPT by computing

$$\alpha_{\text{OPT}} = 100 \times \frac{J(\text{OPT})}{J(\text{WIP})}. \tag{17}$$

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 I.

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

$$\varepsilon_{\text{MYP}} = 100 \times \frac{J(\text{MYP}) - J(\text{WIP})}{J(\text{MYP})}. \tag{18}$$

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 are shown in Table II, respectively.

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

(a) $m = 1$

ε_{MYP}		ℓ			
		1	2	3	4
n	20	7.88	11.4	9.66	10.2
	40	12.1	14.6	13.4	7.19
	60	14.5	12.9	11.8	6.06

(b) $m = 5$

ε_{MYP}		ℓ			
		1	2	3	4
n	20	0.77	1.43	0.88	3.72
	40	1.49	3.96	3.76	8.59
	60	4.13	5.45	4.92	8.37

C. 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.

In Experiment 2 where WIP is compared with MYP, we observe ε_{MYP} ranges from 0.15% to 14.5% which in overall, shows that WIP outperforms MYP considerably.

Furthermore, we observe that as n increases, ε_{MYP} also increases in general. 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.

VII. Conclusion

We investigated partially observable restless bandits. Unlike most of the existing literature which restricts attention to models with binary state space, we do not impose such as assumption. 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. We used the threshold policy to present a refinement of the adaptive greedy algorithm of [26] 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

Consider a Markov chain with n 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,

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

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 $q_i = (1-p)/(n-i)$. Then,

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

REFERENCES

[1] R. Meshram, D. Manjunath, and A. Gopalan, "On the Whittle index for restless multiarmed hidden Markov bandits," *IEEE Trans. Autom. Control*, vol. 63, no. 9, pp. 3046–3053, 2018.

[2] S. Guha, K. Munagala, and P. Shi, "Approximation algorithms for restless bandit problems," *Journal of the ACM (JACM)*, vol. 58, no. 1, p. 3, 2010.

[3] K. Kaza, R. Meshram, V. Mehta, and S. N. Merchant, "Sequential decision making with limited observation capability: Application to wireless networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 2, pp. 237–251, 2019.

[4] K. Kaza, V. Mehta, R. Meshram, and 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, and P. Osti, "Whittle index approach to size-aware scheduling for time-varying channels with multiple states," *Queueing Systems*, vol. 83, no. 3-4, pp. 195–225, 2016.

[6] M. Larrañaga, M. Assaad, A. Destounis, and 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 and 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, and J. Wason, "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*, vol. 30, no. 2, p. 199, 2015.

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

[10] K. D. Glazebrook, H. M. Mitchell, and P. S. Ansell, "Index policies for the maintenance of a collection of machines by a set of repairmen," *European Journal of Operational Research*, vol. 165, no. 1, pp. 267–284, 2005.

[11] S. S. Villar, "Indexability and optimal index policies for a class of reinitialising restless bandits," *Probability in the engineering and informational sciences*, vol. 30, no. 1, pp. 1–23, 2016.

[12] Y. Qian, C. Zhang, B. Krishnamachari, and M. Tambe, "Restless poachers: Handling exploration-exploitation tradeoffs in security domains," in *Int. Conf. on Autonomous Agents & Multiagent Systems*, 2016, pp. 123–131.

[13] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, no. A, pp. 287–298, 1988.

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

[15] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of Applied Probability*, vol. 27, no. 3, pp. 637–648, 1990.

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

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

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

[19] P. S. Ansell, K. D. Glazebrook, J. Niño-Mora, and M. O'Keeffe, "Whittle's index policy for a multi-class queueing system with convex holding costs," *Mathematical Methods of Operations Research*, vol. 57, no. 1, pp. 21–39, 2003.

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

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

[22] T. W. Archibald, D. P. Black, and K. D. Glazebrook, "Indexability and index heuristics for a simple class of inventory routing problems," *Operations research*, vol. 57, no. 2, pp. 314–326, 2009.

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

[24] K. Glazebrook, D. Hodge, and C. Kirkbride, "Monotone policies and indexability for bidirectional restless bandits," *Advances in Applied Probability*, vol. 45, no. 1, pp. 51–85, 2013.

[25] Z. Yu, Y. Xu, and L. Tong, "Deadline scheduling as restless bandits," *IEEE Trans. Autom. Control*, vol. 63, no. 8, pp. 2343–2358, 2018.

[26] N. Akbarzadeh and A. Mahajan, "Conditions for indexability of restless bandits and an $\mathcal{O}(k^3)$ algorithm to compute whittle index," *Advances in Applied Probability*, p. 1–29, 2022.

[27] J. Niño-Mora, "Restless bandits, partial conservation laws and indexability," *Advances in Applied Probability*, vol. 33, no. 1, pp. 76–98, 2001.

[28] D. Ruiz-Hernández, J. M. Pinar-Pérez, and D. Delgado-Gómez, "Multi-machine preventive maintenance scheduling with imperfect interventions: A restless bandit approach," *Computers & Operations Research*, vol. 119, p. 104927, 2020.

[29] C. R. Dance and T. Silander, "Optimal policies for observing time series and related restless bandit problems." *J. Mach. Learn. Res.*, vol. 20, pp. 35–1, 2019.

[30] A. Mate, J. Killian, H. Xu, A. Perrault, and M. Tambe, "Collapsing bandits and their application to public health intervention," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 15 639–15 650.

[31] K. J. Astrom, "Optimal control of Markov processes with incomplete state information," *Journal of mathematical analysis and applications*, vol. 10, no. 1, pp. 174–205, 1965.

[32] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, 1999.

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

[34] D. I. Shuman, A. Nayyar, A. Mahajan, Y. Goykhman, K. Li, M. Liu, D. Teneketzis, M. Moghaddam, and D. Entekhabi, "Measurement scheduling for soil moisture sensing: From physical models to optimal control," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1918–1933, 2010.

[35] N. Akbarzadeh and A. Mahajan, "Two families of indexable partially observable restless bandits and whittle index computation," *arXiv preprint arXiv:2104.05151*, 2021.