Maintenance of a collection of machines under partial observability: A restless bandit approach

Nima Akbarzadeh, *Student Member, IEEE,* and Aditya Mahajan, *Senior Member, IEEE*

***Abstract*—We consider the problem of scheduling maintenance for a collection of machines under partial observations when the state of each machine deteriorates stochastically in a Markovian manner. We consider two observational models: first, the state of each machine is not observable at all, and second, the state of each machine is observable only if a service-person visits them. The agent takes a maintenance action, e.g., machine replacement, if he is chosen for the task. We model both problems as restless multi-armed bandit problem and propose the Whittle index policy for scheduling the visits. We show that both models are indexable. For the first model, we derive a closed-form expression for the Whittle index. For the second model, we propose an efficient algorithm to compute the Whittle index by exploiting the qualitative properties of the optimal policy. We present detailed numerical experiments which show that for multiple instances of the model, the Whittle index policy outperforms myopic policy and can be close-to-optimal in different setups.**

***Index Terms*—Scheduling, Machine Maintenance, Partially observable system, Restless bandits, Whittle index**

I. Introduction

Machines are subject to degradation. Malfunctioning of machines is a major cause of reduction of production capacity, deterioration of quality of service, and increase in downtime, monetary cost and lost work [1], [2]. In critical environments such as computer servers, aircrafts, power generators, medical devices, infrastructure tools, etc., malfunctioning of machines can cause catastrophic failures. In such environments, maintenance plays an essential role to ensure system reliabiilty [3]. For these reasons, scheduling of machine maintenance has been a problem of interest for several decades [4]–[7].

Broadly speaking, the literature on scheduling of machine maintenance may be classified in two categories: fully-observable and partially-observable models. Each category can be further separated in two subcategories: maintenance of a single machine or a collection of machines.

The first category considers a class of problems where the condition of the machine is accessible or can be directly monitored at all times. In such models, the decision-maker is responsible to maintain a machine or a collection of machines by regular or opportunistic inspections and taking an approriate maintenance action (repair, replacement, no actions etc.). Maintenance of a single machine with a single or multiple components is considered in [8]–[16]. Maintenance of a collection of machines is considered in [17]–[21].

N. Akbarzadeh and A. Mahajan are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada, nima.akbarzadeh@mail.mcgill.ca, aditya.mahajan@mcgill.ca

This work of Nima Akbarzadeh was supported in part by Fonds de Recherche du Quebec-Nature et technologies (FRQNT) fellowship.

The second category arises in applications where the condition of the machine cannot be directly observed or is observed via noisy sensors. Examples include automated demand response devices [22], semiconductor manufacturing [23], or maintenance of wind turbines [24]. Maintenance of a single machine with a single or multiple components is considered in [23]–[28]. Maintenance of a collection of machines is considered in [22]. See [23] for a literature overview of the recent development in this category.

We are interested in maintenance of a collection of machines under partial observations. Specifically, we 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. Due to manufacturing mistakes, all the machines may not be in pristine state when installed. If a machine is left un-monitored, then the state of the machine deteriorates and after a while, it ruins. 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. The company has $m < n$ service-persons. Therefore, the company has to schedule when a service-person should visit each machine to minimize the cumulative long-term cost.

The problem of identifying the optimal scheduling policies suffers from curse of dimensionality [29]. For example, if a single server-person is responsible to maintain 100 machines and each machine has three states. Then the total number of states is 3^{100} which is astronomically large. As a result, obtaining the optimal scheduling policy for such problems is, in general, intractable. When the state of the machines are not observed, the problem becomes even harder.

We model this setup as a restless multi-armed bandit (RMAB) and show that a heuristic policy, known as Whittle index [30], is applicable in this model. RMAB is a class of sequential decision making problem where a decision maker confronts n arms and can pick m arms ($m < n$) at each time. Each arm is modeled as a Markov chain and its dynamics differs when the arm is active (i.e., the arm is chosen) or passive (i.e., the arm is not chosen). There is a per-step cost associated with each arm which depends on state of the arms and the action. The decision maker's goal is to minimize the expected discounted cost of all arms accumulated over an infinite time horizon. The Whittle index exists if the RMAB problem satisfies a technical condition known as indexability. When the condition is satisfied, we can decompose the n-dimensional problem into n 1-dimensional problem and solve each problem separately.

This substantially reduces the computational complexity. For instance, in the example mentioned earlier, the optimization problem with 3^{100} states alters to 100 problems with 3 states.

The restless bandit framework is widely used in scheduling problems arising in various applications including telecommunication networks [31]–[37], patient prioritization [38], machine maintenance [18], [22], sensor management [39] and game theory [40].

RMAB models have been considered for both fully-observable [18], [30], [41]–[45] and partially-observable [22], [31]–[36], [39], [40] setups. The partially observable models are conceptually and computationally more challenging and most of the literature restricts attention to models where each alternative has two states [22], [31]–[35], [39]. Such models are usually investigated under an additional technical assumption that the states are positively correlated [33]–[35]. However, there are only a few papers which consider a general state space under partial observations [36], [37], [40], [46] and these often resort to numerical methods to verify indexability.

The main contributions of our paper are as follows:

- We model the machine maintenance problem with partial observation as a RMAB. For two different observation models, we show that the RMAB is indexable. Unlike much of the prior work on RMAB under partial observations, we do not restrict attention to binary states [22], [31]–[35], [39].
- We provide a closed-form expression to compute Whittle index for the first observation model and present an improved version of the adpative greedy algorithm to compute Whittle index presented in [43] for the second observation model.
- 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 oragnization of the paper is as follows. In Section II, we formulate the machine maintenance problem for two different observation models. In Section III, we present various simplifications of the model. In Section IV, we present a short overview of restless bandits. In Section V, we show the restless bandit problem is indexable for both models and present algorithms to compute Whittle index. In Section VI, we present a detailed numerical study which compares the performance of Whittle index policy with the optimal and myopic policies. In Section VII, we present the conclusions and discussions.

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 positive 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 for machine index 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 $f(x_1, y_2) - f(x_1, y_1) \geq f(x_2, y_2) - f(x_2, y_1)$. Given an ordered set $\mathcal{X}$, 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 $z \in \mathcal{X}$.

II. Model and Problem Formulation

A. System Model

Consider a system operator who has to maintain a collection of n machines, which we index by the set $\mathcal{N} \coloneqq \{1, \ldots, n\}$. Machine $i \in \mathcal{N}$ has an operating state which belongs to the set $\mathcal{X}^i = \{1, 2, \ldots, |\mathcal{X}^i|\}$. State 1 denotes the pristine state and a higher value indicate a more degraded state. There is a state-dependent cost associated with running the machine, which is captured by the cost function $\phi^i : \mathcal{X}^i \to [0, \infty)$. The function ϕ^i is an increasing function with $\phi^i(1) = 0$. The state of each machine deteriorates with time in a Markovian manner. We model this by assuming that the transition probability matrix P^i is upper triangular and stochastic monotone. Upper triangularity implies that the state of the machine deteriorates over time. Stochastic monotonicity implies that a machine in a bad state deteriorates faster than a machine in a good state.

There are $m < n$ service-persons and the operator may send a service-person to service machine $i \in \mathcal{N}$ at cost ρ^i. At each time, a service-person can be sent to only one machine and a machine can be serviced by only one service-person. Note that the operator may decide not to send a service-person to service any machine and this decision does not incur any cost. We assume the machines are cheap to manufacture, so when a service-person services a machine, he simply replaces the machine by a new one. However, the new machine does not necessarily start in a pristine state. Either due to quality management during manufacturing or storage, the state of a new machine of type $i \in \mathcal{N}$ is distributed according to the probability mass function Q^i. Note that when a service-person visits a machine, the system does not incur a running cost. The act of sending a service-person to machine i is denoted by action 1, and not sending a service-person is denoted by action 0 otherwise. Thus, the per-step cost incurred when the machine is in state $x \in \mathcal{X}^i$ and action $a \in \{0, 1\}$ is applied is given by

$$c^i(x, a) = (1 - a)\phi^i(x) + a\rho^i.$$

We assume the operator cannot observe the state of the machines and consider two observation models for the service-person.

- **Model A:** When deploying a new machine, the service-person does not observe the state of the machine being deployed.
- **Model B:** When deploying a new machine, the service-person observes the state of the machine being deployed.

In each of these cases, we are interested in choosing a decision strategy for the operator to minimize the expected discounted cost of running the system for an infinite time horizon. We formally state the problem in the next section.

B. Problem Formulation

Let $\boldsymbol{\mathcal{X}} \coloneqq \prod_{i\in\mathcal{N}} \mathcal{X}^i$ denote the state space of all machines. We use $\boldsymbol{X}_t = (X_t^1, \dots X_t^n) \in \boldsymbol{\mathcal{X}}$ to denote the state of all machines at time t and $\boldsymbol{A}_t = (A_t^1, \dots, A_t^n) \in \boldsymbol{\mathcal{A}}(m)$ denote the action taken by the operator at time t, where

$$\boldsymbol{\mathcal{A}}(m) \coloneqq \left\{ (a^1, \dots, a^n) \in \{0,1\}^n : \sum_{i\in\mathcal{N}} a^i \le m \right\} \tag{1}$$

denotes the set of all feasible actions. If component A_t^i of $\boldsymbol{A}_t$ is 1, it means that the operator sends a service-person to machine i; if $A_t^i = 0$ it means that the operator does not send a service-person to machine i.

Let $\boldsymbol{Y}_t = (Y_t^1, \dots, Y_t^n) \in \prod_{i\in\mathcal{N}} \mathcal{Y}^i$ denote the observation of the service-person at time t. For model A, $\mathcal{Y}^i = \{\mathfrak{E}\}$ and $Y_t^i = \mathfrak{E}$ which indicates that the operator never gets any observation about the state of the machine. For model B, $\mathcal{Y}^i = \mathcal{X}^i \cup \mathfrak{E}$, where

$$Y_{t+1}^i = \begin{cases} \mathfrak{E} & \text{if } \ A_t^i = 0 \\ X_{t+1}^i & \text{if } \ A_t^i = 1 \end{cases}, \quad i \in \mathcal{N}, \tag{2}$$

which indicates that when a machine is replaced, the service-person observes the state of the new machine.

The state of each machine evolves independently in a controlled Markov manner, i.e.,

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

where

$$\mathbb{P}(X_{t+1}^i = x_{t+1}^i | X_t^i = x_t^i, A_t^i = a_t^i) = \begin{cases} P^i_{x_t^i x_{t+1}^i} & \text{if } a_t^i = 0 \\ Q^i_{x_{t+1}^i} & \text{if } a_t^i = 1. \end{cases}$$

The decision at time t is chosen according to

$$\boldsymbol{A}_t = \boldsymbol{g}_t(\boldsymbol{Y}_{0:t-1} \boldsymbol{A}_{0:t-1}), \tag{3}$$

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

We assume initial state of machine i is random and distributed according to pmf π_0^i. Let $\boldsymbol{\pi}_0 = \bigotimes_{i\in\mathcal{N}} \pi_0^i$ denote the initial state distribution of all machines. The performance of policy g is given by

$$J^{(\boldsymbol{g})}(\boldsymbol{\pi}_0) \coloneqq (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \sum_{i\in\mathcal{N}} c^i(X_t^i, A_t^i) \middle| \begin{matrix} X_0^i \sim \pi_0^i, \\ \forall i \in \mathcal{N} \end{matrix} \right], \tag{4}$$

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

Let $(\mathcal{X}^i, \{0,1\}, P^i, Q^i, c^i, \pi_0^i)$ denote all the model parameters for machine i. Formally, the optimization problem of interest is as follows:

Problem 1. *Given a discount factor $\beta \in (0,1)$, the total number n of machines and m of service-persons, and the system model $\{(\mathcal{X}^i, \{0,1\}, P^i, Q^i, c^i, \pi_0^i)\}_{i\in\mathcal{N}}$, choose a policy $\boldsymbol{g} \in \boldsymbol{\mathcal{G}}$ that minimizes $J^{(\boldsymbol{g})}(\boldsymbol{\pi}_0)$ given by* (4)*.*

C. Roadmap of the results

Problem 1 is a partially observable Markov decision process (POMDP). One conceptual challange in POMDPs is that the actions have to be chosen as a function of the entire history of observations at the decision maker. Such history dependent policies are difficult to search and implement. The standard approach used in the literature to handle this conceptual difficulty is to transform the partially observed MDP into a fully observed MDP by using the decision maker's posterior belief on the unobserved state of the system given the history of observations as the information state of the system. We present this transformation in Section III-A.

However, such a transformation suffers from the curse of dimensionality. The belief state space takes values in the simplex $\mathcal{P}(\boldsymbol{\mathcal{X}})$ which is double exponential in n. In Section III-A we exploit the structure of the model to present a simpler belief which has a dimension that is exponential in n. Even this simpler belief is continuous valued which makes the resulting dynamic programming is difficult to solve.

In Section III-B, we exploit the structure of the reachable set of the simpler belief to propose an alternative information state which is countable. A countable information sate is advantageous because it can be approximated by a finite information state via truncation. Although the resulting finite state dynamic programs is significantly simpler than the naive beleif state based MDPs, it still suffers from curse of dimensionality and can be only solved for small values of n.

To circumvent the curse of dimensionality, we model the countable state MDP as a restless multi-armed bandit (RMAB) in Section IV. In Section V, we show that both models are indexable and present efficient algorithms to compute Whittle index.

III. Preliminary results: POMDP characterization and identification of a simpler information state

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 [47]. We will discuss the details below.

A. Belief state formulation for Problem 1

Define the 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 [47], we have the following.

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

Note that we do not give the details of the dynamic program based on the belief state, Θ_t because we pursue a different solution approach.

For the model under consideration, it is possible to simplify the belief state. For that matter, we define the decision maker's belief $\Pi^i_t \in \mathcal{P}(\mathcal{X}^i)$ on the state of machine i at time t as follows: for any, $x^i_t \in \mathcal{X}^i$, let

$$\Pi^i_t(x^i_t) \coloneqq \mathbb{P}(X^i_t = x^i_t \,|\, Y^i_{0:t-1}, A^i_{0:t-1}).$$

Similar to Θ_t, Π^i_t is also a distribution-valued random variable. Let $\mathbf{\Pi}_t \coloneqq (\Pi^1_t, \dots, \Pi^n_t)$.

The belief state of machine i evolves according to a controlled Markov process. In particular, for model A, the belief update rule is

$$\Pi^i_{t+1} = \begin{cases} \Pi^i_t P^i, & \text{if } A^i_t = 0, \\ Q^i, & \text{if } A^i_t = 1, \end{cases} \tag{5}$$

and for model B, the belief update rule is

$$\Pi^i_{t+1} = \begin{cases} \Pi^i_t P^i, & \text{if } A^i_t = 0, \\ \delta_{X^i_{t+1}} \text{ where } X^i_{t+1} \sim Q^i, & \text{if } A^i_t = 1. \end{cases} \tag{6}$$

Furthermore, from the definition of the belief state we get that

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

Moreover, since the machines are independent, we have

$$\mathbb{E}[c_t(\boldsymbol{X}_t, \boldsymbol{A}_t) | \boldsymbol{Y}_{0:t-1}, \boldsymbol{A}_{0:t-1}] = \sum_{i=1}^{n} \bar{c}^i(\Pi^i_t, A^i_t) =: \bar{\boldsymbol{c}}(\mathbf{\Pi}_t, \boldsymbol{A}_t). \tag{7}$$

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*

$$\Theta_t(\boldsymbol{x}) = \prod_{i \in \mathcal{X} N} \Pi^i_t(x^i), \quad \textit{a.s.}. \tag{8}$$

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

Note that as before, we do not present the details of the dynamic program because we pursue a different solution approach.

Proof. Eq. (8) follows from the conditional independence of the machines, and the nature of the observation function. The structure of the optimal policies then follow immediately from Proposition 1. □

B. Information state for Problem 1

Although the transformation presented in Section III-A simplifies the problem by allowing the use of Markov decision theory and dynamic programming to solve the problem, the resulting fully observable system is computationally intractable in general [29]. Inspired by the approach taken in [48], we introduce a new information state which is countable and, at the same time, is equivalent to the belief state.

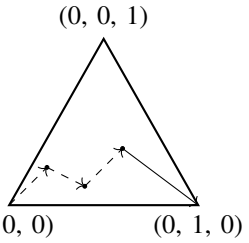

Fig. 1: Belief state dynamics for a 3-state machine i in the simplex $\mathcal{P}(\{1, 2, 3\})$. Dashed arrows show a sample realizations of the belief state evolution under $A^i_t = 0$ for three time steps and the solid arrow shows a sample realization of the belief state evolution under $A^i_t = 1$.

For model A, define

$$\mathcal{R}^i_A = \left\{ Q^i (P^i)^k : k \in \mathbb{Z}_{\geq 0} \right\},$$

and for model B define

$$\mathcal{R}^i_B = \left\{ \delta_s (P^i)^k : s \in \mathcal{X}^i, k \in \mathbb{Z}_{\geq 0} \right\}.$$

Assumption 1. *For model A, $\pi^i_0 \in \mathcal{R}^i_A$ and for model B, $\pi^i_0 \in \mathcal{R}^i_B$.*

For model A, define a process $\{K^i_t\}_{t \geq 0}$ as follows. The initial state k^i_0 is such that $\pi^i_0 = Q^i (P^i)^{k^i_0}$ and for $t > 0$, K^i_t is given by

$$K^i_t = \begin{cases} 0, & \text{if } A^i_{t-1} = 1 \\ K^i_{t-1} + 1, & \text{if } A^i_{t-1} = 0. \end{cases} \tag{9}$$

Similarly, for model B, define a process $\{S^i_t, K^i_t\}_{t \geq 0}$ as follows. The initial state (s^i_0, k^i_0) is such that $\pi^i_0 = \delta_{s^i_0} (P^i)^{k^i_0}$ and for $t > 0$, K^i_t evolves according to (9) and S^i_t evolves according to

$$S^i_t = \begin{cases} X^i_{t-1} \text{ where } X^i_{t-1} \sim Q^i, & \text{if } A^i_{t-1} = 1 \\ S^i_{t-1}, & \text{if } A^i_{t-1} = 0. \end{cases} \tag{10}$$

Note that once the first observation has been taken in both models, K^i_t denotes the time elapsed since the last observation of machine i and, in addition in model B, S^i_t denotes the last observed states of machine i. Let $\boldsymbol{S}_t \coloneqq (S^1_t, \dots S^n_t)$ and $\boldsymbol{K}_t \coloneqq (K^1_t, \dots K^n_t)$. The relation between the belief state Π^i_t and variables S^i_t and K^i_t is characterized in the following lemma.

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

- *For model A, for any $i \in \mathcal{N}$ and any t, $\Pi^i_t \in \mathcal{R}^i_A$. In particular, $\Pi^i_t = Q^i (P^i)^{K^i_t}$.*
- *For model B, for any $i \in \mathcal{N}$ and any t, $\Pi^i_t \in \mathcal{R}^i_B$. In particular, $\Pi^i_t = \delta_{S^i_t} (P^i)^{K^i_t}$.*

Proof. The results immediately follow from (5)-(6) and (9)-(10). □

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

$$\bar{c}^i(K^i_t, A^i_t) \coloneqq \bar{c}^i((Q^i P^i)^{K^i_t}, A^i_t) = \sum_{x \in \mathcal{X}^i} [(Q^i P^i)^{K^i_t}]_x c^i(x, A^i_t). \tag{11}$$

and the total expected per-step cost incurred at time t may be written as

$$\bar{\boldsymbol{c}}(\boldsymbol{K}_t, \boldsymbol{A}_t) \coloneqq \sum_{i=1}^{n} \bar{c}^i(K_t^i, A_t^i).$$

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

$$\begin{aligned}\bar{c}^i(S_t^i, K_t^i, A_t^i) &\coloneqq \bar{c}^i(\delta_{S_t^i}(P^i)^{K_t^i}, A_t^i) \\ &= \sum_{x \in \mathcal{X}^i} [\delta_{S_t^i}(P^i)^{K_t^i}]_x c^i(x, A_t^i). \end{aligned} \tag{12}$$

and the total expected per-step cost incurred at time t may be written as

$$\bar{\boldsymbol{c}}(\boldsymbol{S}_t, \boldsymbol{K}_t, \boldsymbol{A}_t) \coloneqq \sum_{i=1}^{n} \bar{c}^i(S_t^i, K_t^i, A_t^i).$$

Proposition 3. *In Problem 1, there is no loss of optimality in restricting attention to decision policies of the form $\boldsymbol{A}_t = g_t^{info}(\boldsymbol{K}_t)$ for model A and of the form $\boldsymbol{A}_t = g_t^{info}(\boldsymbol{S}_t, \boldsymbol{K}_t)$ for model B.*

Proof. This result immediately follows from Lemma 1, (11) and (12). □

In the next section, we review the basic concepts of restless multi-armed bandits problem (RMAB) and later, we show how Problem 1 can be modeled as a RMAB.

IV. Overview of Restless Multi-armed bandits

A. Restless Bandit Process

A restless bandit process (RB) is a controlled Markov process $(\tilde{\mathcal{X}}, \{0,1\}, \{\tilde{P}, \tilde{Q}\}, \tilde{c}, \pi_0)$ where $\tilde{\mathcal{X}}$ denotes the state space, $\{0,1\}$ denotes the action space, $\tilde{P}$ and $\tilde{Q}$ denote the transition probability matrices under actions 0 and 1 respectively, $\tilde{c} : \tilde{\mathcal{X}} \times \{0,1\} \to \mathbb{R}$ denotes the per-step cost function, and π_0 is the initial state distribution. Conventionally, the action 0 is called the *passive* action and the action 1 is called the *active* action. Let $\{\tilde{X}_t\}_{t\geq 0}$ and $\{\tilde{A}_t\}_{t\geq 0}$ denote the sequence of observed states and the chosen actions. Then, for any $x', x \in \tilde{\mathcal{X}}$ and $a \in \{0,1\}$, we have

$$\mathbb{P}(\tilde{X}_{t+1} = x | \tilde{X}_t = x', \tilde{A}_t = a) = \begin{cases} \tilde{P}_{xx'} & \text{if } a = 0 \\ \tilde{Q}_{xx'} & \text{if } a = 1. \end{cases}$$

B. Restless Multi-armed Bandit Problem

A group of n independent RBs $(\tilde{\mathcal{X}}^i, \{0,1\}, \{\tilde{P}^i, \tilde{Q}^i\}, \tilde{c}^i, \tilde{\pi}_0^i)$, $i \in \mathcal{N}$ is called a restless multi-armed bandit (RMAB) problem. Each process is also called as an *arm* in the literature [49]. A decision-maker selects m arms ($m < n$) at each time instance. Let $\tilde{X}_t^i$ and $\tilde{A}_t^i$ denote the state of arm i and the action chosen for arm i at time t. Let $\{\tilde{\boldsymbol{X}}_t\}_{t\geq 0}$ and $\{\tilde{\boldsymbol{A}}_t\}_{t\geq 0}$ denote the sequence of observed states and the chosen actions for all arms. Additionally, let $\tilde{\boldsymbol{\mathcal{X}}} = \prod_{i=1}^{n} \tilde{\mathcal{X}}^i$ and let $\boldsymbol{\mathcal{A}}(m)$ be the same as the one defined in (1). Then, for any $\boldsymbol{x}, \boldsymbol{x'} \in \tilde{\boldsymbol{\mathcal{X}}}, \boldsymbol{a} \in \boldsymbol{\mathcal{A}}(m)$ where $\boldsymbol{x} = (x^1, \ldots, x^n)$ and similar for $\boldsymbol{x'}, \boldsymbol{a}$, we have

$$\begin{aligned}&\mathbb{P}(\tilde{\boldsymbol{X}}_{t+1} = \boldsymbol{x} | \tilde{\boldsymbol{X}}_t = \boldsymbol{x'}, \tilde{\boldsymbol{A}}_t = \boldsymbol{a}) = \\ &\prod_{i \in \mathcal{N}} \mathbb{P}(\tilde{X}_{t+1}^i = x^i | \tilde{X}_t^i = x'^i, a^i).\end{aligned}$$

The instantaneous cost of the system is the sum of costs incurred by each RB. The decision at time t is chosen according to a time homogeneous Markov policy $\tilde{\boldsymbol{g}} : \tilde{\boldsymbol{\mathcal{X}}} \to \boldsymbol{\mathcal{A}}(m)$. Let $\tilde{\pi}_0^i$ denote the initial state distribution of arm i and $\tilde{\boldsymbol{\pi}}_0 = \prod_{i \in \mathcal{N}} \tilde{\pi}_0^i$. Then, the performance of policy $\tilde{\boldsymbol{g}}$ is measured by

$$\tilde{J}^{(\tilde{\boldsymbol{g}})}(\tilde{\boldsymbol{\pi}}_0) \coloneqq (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \sum_{i \in \mathcal{N}} \tilde{c}^i(\tilde{X}_t^i, \tilde{A}_t^i) \middle| \begin{matrix} \tilde{X}_0^1 \sim \tilde{\pi}_0^1, \\ \tilde{X}_0^n \sim \tilde{\pi}_0^n \end{matrix}\right], \tag{13}$$

where $\beta \in (0,1)$ denotes the discount factor. Finally, the restless bandit optimization problem is as follows.

Problem 2. *Given a discount factor $\beta \in (0,1)$, the models $\{(\tilde{\mathcal{X}}^i, \{0,1\}, \{\tilde{P}^i, \tilde{Q}^i\}, \tilde{c}^i, \tilde{\pi}_0^i)\}_{i \in N}$ of n arms and the number m of arms to be chosen at each time, choose a policy $\tilde{\boldsymbol{g}} : \tilde{\boldsymbol{\mathcal{X}}} \to \boldsymbol{\mathcal{A}}(m)$ that minimizes $\tilde{J}^{(\tilde{\boldsymbol{g}})}(\tilde{\boldsymbol{\pi}}_0)$ given by* (13).

Problem 2 is an MDP which can, in principle, be solved using dynamic programming. However, such a solution suffers from the curse of dimensionality because the state space of the dynamic program is exponential in the number of arms.

The RMAB problem simplifies significantly when arms remain frozen under passive action (i.e., $\tilde{Q}^i = I$). Under this assumption, [50] showed that the optimal solution is of the index type, i.e., we compute an index function $\nu^i : \tilde{\mathcal{X}}^i \to \mathbb{R}$ for each arm and at each time, play the arm which is in the state with highest index. Inspired by this result, [30] argued that a similar index policy should perform well for general RMAB provided a technical condition known as Whittle indexability is satisfied. Subsequently, [44] and [51] have identified different sufficient conditions under which Whittle index policy is optimal. There is a strong empirical evidence to suggest that in many applications, the Whittle index heuristic performs close-to-optimal in practice [18], [41]. Furthermore, [52] introduces PCL-indexability concept and shows that PCLs can act as sufficient conditions for indexability of restless bandit problems, and shows that an adaptive greedy algorithm can be used to compute the indices. This idea is further developed in [53].

C. Indexability and the Whittle index

Consider a RB $(\tilde{\mathcal{X}}, \{0,1\}, \{\tilde{P}, \tilde{Q}\}, \tilde{c}_\lambda, \tilde{\pi}_0)$ with a modified per-step cost function

$$\tilde{c}_\lambda(x, a) \coloneqq c(x, a) + \lambda a, \quad \forall x \in \tilde{\mathcal{X}}, \forall a \in \{0,1\}, \lambda \in \mathbb{R}. \tag{14}$$

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

$$J_\lambda^{(\tilde{g})}(\tilde{\pi}_0) \coloneqq (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t c_\lambda(X_t, \tilde{g}(X_t)) \middle| X_0 \sim \tilde{\pi}_0\right]. \tag{15}$$

Subsequently, consider the following optimization problem.

Problem 3. *Given an arm $(\tilde{\mathcal{X}}, \{0,1\}, \{\tilde{P}, \tilde{Q}\}_{a \in \{0,1\}}, \tilde{c}, \tilde{\pi}_0)$, the discount factor $\beta \in (0,1)$ and the penalty $\lambda \in \mathbb{R}$, choose a Markov policy $\tilde{g} : \tilde{\mathcal{X}} \to \{0,1\}$ to minimize $J_\lambda^{(\tilde{g})}(\tilde{\pi}_0)$ given by* (15).

Problem 3 is a Markov decision process where one may use dynamic program to obtain the optimal solution as follows.

Proposition 4. *Let $\tilde{V}_\lambda : \tilde{\mathcal{X}} \to \mathbb{R}$ be the unique fixed point of equation*

$$\tilde{V}_\lambda(x) = \min_{a\in\{0,1\}} \tilde{H}_\lambda(x,a) \tag{16}$$

where

$$\tilde{H}_\lambda(x,0) = (1-\beta)\bar{c}(x,0) + \beta \sum_{x'\in\tilde{\mathcal{X}}} \tilde{P}_{xx'}\tilde{V}_\lambda(x'),$$

$$\tilde{H}_\lambda(x,1) = (1-\beta)\bar{c}(x,1) + (1-\beta)\lambda + \beta \sum_{x'\in\tilde{\mathcal{X}}} \tilde{Q}_{xx'}\tilde{V}_\lambda(x').$$

Let $\tilde{g}_\lambda(x)$ denote the $\arg\min$ *of the right hand side of* (16) *where we set $\tilde{g}_\lambda(x) = 1$ if the two argument inside $\min\{\cdot,\cdot\}$ are equal. Then, the time-homogeneous policy $\tilde{g}_\lambda$ is optimal for Problem 3.*

Proof. The result follows immediately from Markov decision theory [54]. □

Finally, given penalty λ, define the passive set $\mathcal{W}_\lambda$ as the set of states where passive action is optimal for the modified RB, i.e.,

$$\mathcal{W}_\lambda := \{x \in \mathcal{X} : \tilde{g}_\lambda(x) = 0\}.$$

Definition 1 (Indexability)**.** *A RB 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 RMAB problem is indexable if all n RBs are indexable.

Definition 2 (Whittle index)**.** *The Whittle index of the state x of an arm is the smallest value of λ for which state x is part of the passive set $\mathcal{W}_\lambda$, i.e.,*

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

Equivalently, the Whittle index $w(x)$ is the smallest value of λ for which the optimal policy is indifferent between the active action and passive action when the information state of the machine is k.

D. Whittle Index Policy

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 [44], [51] and performs close-to-optimal for many other cases [18], [41].

E. Sufficient Condition for indexability

To verify indexability of our model, we use a recently proposed sufficient condition for indexability [43] which we sumerize below. A RMAB is said to have the *restart* property if each arm satisfies the following condition:

(R) The transition probability matrix under active action does not depend on the current state, i.e., $\tilde{Q}_{xy} = \tilde{Q}_{x'y}$ for any $x, x', y \in \tilde{\mathcal{X}}$.

Proposition 5 ([43])**.** *If a RB satisfies Condition **(R)**, then it is indexable.*

We use Proposition 5 to prove indexability for both of the models.

V. Indexability and Computation of Whittle index

In this section, we use the belief state to show that both models A and B are indexable. Afterwards, we use the infotmation states to compute the Whittle index for both models. For ease of notation, we will drop the superscript i from all relative variables for the rest of this and the next sections.

A. Indexability of models A and B

Proposition 6. *Problem 3 is indexable for models A and B for any arm $i \in \mathcal{N}$.*

Proof. Accoring to (5) for Model A and (6) for Model B, the RMAB defined based on the belief state satisfies the restart property **(R)**. Hence, the problem is indexable by Proposition 5. □

Next, we derive formulas and algorithms to compute the Whittle index.

For a Whittle indexable restless bandit problem, two approaches have been followed in the literature. For some specific models, it is possible to derive a closed-form expression for the Whittle index [18], [33]–[36], [41]–[43], [55]. However, in general, no such closed-form expression exists and the index needs to be computed numerically. For a subclass of RMABs 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 [52], [53]. Recently, [43] presented a generalization of adaptive greedy algorithm which is applicable to all indexable RMABs.

For model A, we derive a closed-form expression for the Whittle index. For model B, we analyze the structure of the optimal policy and propose a refinement of the modified adaptive greedy algorithm of [43].

The roadmap to this section is as follows. As the first step, 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.

B. Structural properties of the optimal policy

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

Theorem 1. *The following statements hold:*

1) *In model A, for any $\lambda \in \mathbb{R}$, the optimal policy $g^A_\lambda(k)$ is a threshold policy, i.e., there exists a threshold $\theta^A_\lambda \in$*

$\mathbb{Z} \geq -1$ *such that*

$$g^A_\lambda(k) = \begin{cases} 0, & k < \theta^A_\lambda \\ 1, & \textit{otherwise}. \end{cases}$$

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

$$g^B_\lambda(s,k) = \begin{cases} 0, & k < \theta^B_{s,\lambda} \\ 1, & \textit{otherwise}. \end{cases}$$

The proof is given in Appendix A.

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

C. Performance of threshold based policies

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

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

$$\begin{aligned} J^{(\theta^A)}_\lambda(k) &:= (1-\beta)\mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(K_t, g^{(\theta^A)}(K_t)) \;\Big|\; K_0 = k\Big] \\ &=: D^{(\theta^A)}(k) + \lambda N^{(\theta^A)}(k), \end{aligned} \tag{17}$$

where

$$\begin{aligned} D^{(\theta^A)}(k) &:= (1-\beta)\mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t c(K_t, g^{(\theta^A)}(K_t)) \;\Big|\; K_0 = k\Big], \\ N^{(\theta^A)}(k) &:= (1-\beta)\mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t g^{(\theta^A)}(K_t) \;\Big|\; K_0 = k\Big]. \end{aligned}$$

$D^{(\boldsymbol{\theta}^A)}(k)$ represents the expected total discounted cost while $N^{(\boldsymbol{\theta}^A)}(k)$ 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 6) that the Whittle index can be computed as a function of $D^{(\boldsymbol{\theta}^A)}(k)$ and $N^{(\boldsymbol{\theta}^A)}(k)$. First, we present a method to compute these two variables. Let

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

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

$$\begin{aligned} D^{(\theta^A)}(k) &= L^{(\theta^A)}(k) + \beta^{\theta^A-k+1}\frac{L^{(\theta^A)}(0)}{1-\beta^{\theta^A+1}}, \\ N^{(\theta^A)}(k) &= M^{(\theta^A)}(k) + \beta^{\theta^A-k+1}\frac{M^{(\theta^A)}(0)}{1-\beta^{\theta^A+1}}. \end{aligned}$$

The proof is given in Appendix B.

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

$$\begin{aligned} J^{(\boldsymbol{\theta}^B)}_\lambda(s,k) &= (1-\beta)\mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(S_t, K_t, g^{(\boldsymbol{\theta}^B)}(S_t,K_t)) \\ &\qquad\qquad \Big| \; (S_0,K_0) = (s,k)\Big] \\ &=: D^{(\boldsymbol{\theta}^B)}(s,k) + \lambda N^{(\boldsymbol{\theta}^B)}(s,k), \end{aligned} \tag{18}$$

where

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

$D^{(\boldsymbol{\theta}^B)}(s,k)$ and $N^{(\boldsymbol{\theta}^B)}(s,k)$ have the same interpretations as the ones for model A. We will show (see Theorem 7) that Whittle index can be computed as a function of $D^{(\boldsymbol{\theta}^B)}(s,k)$ and $N^{(\boldsymbol{\theta}^B)}(s,k)$. But first let's define vector $\boldsymbol{J}^{(\boldsymbol{\theta}^B)}_\lambda(0) = (J^{(\boldsymbol{\theta}^B)}_\lambda(1,0), \ldots, J^{(\boldsymbol{\theta}^B)}_\lambda(|\mathcal{X}|,0))$ and vectors $\boldsymbol{D}^{(\boldsymbol{\theta}^B)}(0)$ and $\boldsymbol{N}^{(\boldsymbol{\theta}^B)}(0)$ in a similar manner. Then, from (17), $\boldsymbol{J}^{(\boldsymbol{\theta}^B)}_\lambda(0) = \boldsymbol{D}^{(\boldsymbol{\theta}^B)}(0) + \lambda \boldsymbol{N}^{(\boldsymbol{\theta}^B)}(0)$. Let's also define

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

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

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

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

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

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

The proof is given in Apendix C.

D. Finite state approximation

For computing Whittle index, we provide a finite state approximation of Proposition 4 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 4 (Model A)**.** *Given $\ell \in \mathbb{N}$, let $\mathbb{N}_\ell := \{0, \dots, \ell\}$ and $V_{\ell,\lambda} : \mathbb{N}_\ell \to \mathbb{R}$ be the unique fixed point of equation*

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

where

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

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) Let $\Delta c_\lambda = \max\{\max_x \phi(x), \rho+\lambda\} - \min\{\min_x \phi(x), \rho+\lambda\}$, then for $k \in \mathbb{Z}_{\geq 0}$ such that $k \leq \ell$, we have

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

(ii) For all $k \in \mathbb{Z}_{\geq 0}$, $\lim_{\ell\to\infty} V_{\ell,\lambda}(k) = V_\lambda(k)$. Moreover, let $\hat{g}^*_\lambda(\cdot)$ be any limit point of $\{\hat{g}_{\ell,\lambda}(\cdot)\}_{\ell\geq 1}$. Then, the policy $\hat{g}^*_\lambda(\cdot)$ is optimal for Problem 3.

Proof. (i): Starting from information state $k \in \{0,\dots,\ell-1\}$, the cost incurred by $\hat{g}_{\ell,\lambda}(\cdot)$ is the same as $g^A_\lambda(\cdot)$ for information states $\{k,\dots,\ell\}$. The per-step cost incurred by $\hat{g}_{\ell,\lambda}(\cdot)$ differs from $g^A_\lambda(\cdot)$ for information states $\{\ell+1,\dots\}$ by at most Δc_λ.
(ii): The sequence of finite-state models described above is an *augmentation type approximation sequence* (see [56, Definition 2.5.3]). As a result, a limit point of $\hat{g}^*_\lambda$ exists and the final result holds by [56, Proposition B.5, Theorem 4.6.3]. □

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

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

where

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

We set $\hat{g}_{\ell,\lambda}(s,k) = 1$ if $H_{\ell,\lambda}(s,k,0) = H_{\ell,\lambda}(s,k,1)$. Then, we have the following:
(i) Let $\Delta c_\lambda = \max\{\max_x \phi(x), \rho+\lambda\} - \min\{\min_x \phi(x), \rho+\lambda\}$, then

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

(ii) For all $(s,k) \in \mathcal{X}\times\mathbb{Z}_{\geq 0}$, $\lim_{\ell\to\infty} V_{\ell,\lambda}(s,k) = V_\lambda(s,k)$. Let $\hat{g}^*_\lambda(\cdot,\cdot)$ be any limit point of $\{\hat{g}_{\ell,\lambda}(\cdot,\cdot)\}_{\ell\geq 1}$. Then, the policy $\hat{g}^*_\lambda(\cdot,\cdot)$ is optimal for Problem 3.

Proof. (i): Starting from information state (s,k), given any $s \in \mathcal{X}$ and $k \in \{0,\dots,\ell-1\}$, the cost incurred by $\hat{g}_{\ell,\lambda}(\cdot,\cdot)$ is the same as $g^B_\lambda(\cdot,\cdot)$ for information states $\{(s,l)\}_{l=k}^{\ell}$. The per-step cost incurred by $\hat{g}_{\ell,\lambda}(\cdot,\cdot)$ differs from $g^B_\lambda(\cdot,\cdot)$ for later realized information states by at most Δc_λ. Thus, the bound would hold.
(ii): The sequence of finite-state models described above is an *augmentation type approximation sequence* (see [56, Definition 2.5.3]). As a result, a limit point of $\hat{g}^*_\lambda$ exists and the final result holds [56, Proposition B.5, Theorem 4.6.3]. □

Due to Theorems 4 and 5, we can restrict the countable part of the information state to a finite set, which we denote by $\mathcal{K}$. Note that $\mathcal{K} = \mathbb{N}_\ell$.

E. 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.

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 6. *Let $\Lambda^A_k = \{k_0 \in \{0,1,\dots,|\mathcal{K}|-1\} : N^{(k)}(k_0) \neq N^{(k+1)}(k_0)\}$. Then, $\Lambda^A_k \neq \emptyset$ and for any $k_0 \in \Lambda^A_k$, the Whittle index of model A at information state $k \in \mathcal{K}$ is*

$$w^A(k) = \min_{k_0\in\Lambda^A_k} \frac{D^{(k+1)}(k_0) - D^{(k)}(k_0)}{N^{(k)}(k_0) - N^{(k+1)}(k_0)}. \tag{19}$$

Proof. The proof results from [43, Lemma 4]. □

Theorem 6 gives us a closed-form expression to compute the Whittle index for model A.

2) Modified adaptive greedy algorithm for model B: Let $B = |\mathcal{X}||\mathcal{K}|$ 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 $\mathcal{W}_b := \{(s,k) \in \mathcal{X}\times\mathcal{K} : w(s,k) \leq \lambda_b\}$. For any subset $\mathcal{S} \subseteq \mathcal{X}\times\mathcal{K}$, define the policy $\bar{g}^{(\mathcal{S})} : \mathcal{X}\times\mathcal{K} \to \{0,1\}$ as

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

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

$$\mu_{b,y}(x,k) = \frac{D^{(h_{b,y})}(x,k) - D^{(h_b)}(x,k)}{N^{(h_b)}(x,k) - N^{(h_{b,y})}(x,k)}. \tag{20}$$

Lemma 2. *For $d \in \{0,\dots,B_D-1\}$, we have the following:*
1) *For all $y \in \mathcal{W}_{b+1}\backslash\mathcal{W}_b$, we have $w(y) = \lambda_{b+1}$.*
2) *For all $y \in \Gamma_b$ and $\lambda \in (\lambda_b, \lambda_{b+1}]$, we have $J^{(h_{b,y})}_\lambda(x) \geq J^{(h_b)}_\lambda(x)$ 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. See [43, Lemma 3]. The only difference is that [43] consider $y \in (\mathcal{X}\times\mathcal{K})\backslash\mathcal{W}_b$. However, since we know from Theorem 1 that the optimal policy is a threshold poliy with respect to the second dimension, we restrict y to belong to the set Γ_b. □

Theorem 7. *The following properties hold:*

1) *For any $y \in \mathcal{W}_{b+1} \backslash \mathcal{W}_b$, 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 $y \in \mathcal{W}_{b+1} \backslash \mathcal{W}_b$.*

Proof. See [43, Theorem 2] for the proof steps. Similar to Lemma 2, $(\mathcal{X} \times \mathcal{K}) \backslash \mathcal{W}_b$ is replaced with Γ_b. □

By Theorem 7, we can find the Whittle indices iteratively. By definition, $\mathcal{W}_0 = \emptyset$ and $\lambda_0 = -\infty$. Now suppose $\mathcal{W}_0 \subset \mathcal{W}_1 \subset \ldots \subset \mathcal{W}_b$ and $\lambda_0 < \lambda_1 < \ldots < \lambda_b$ have been identified. Then, we can obtain $\mathcal{W}_{b+1}$ and λ_{b+1} as follows.

1) For $h_b = \bar{g}^{(\mathcal{W}_b)}$, compute $N^{(h_b)}$ by using Theorem 3.
2) For all $y \in \Gamma_b$, compute $N^{(\bar{h}_{b,y})}$ where $\bar{h}_{b,y} = g^{(\mathcal{W} \cup \{y\})}$ by using Theorem 3 and compute $\Lambda_{b,y}$. Let $\mu^*_{b,y} = \min_{(s,k) \in \Lambda_{b,y}} \mu_{b,y}(s,k)$. Then, $\lambda_{b+1} = \min_{y \in \Gamma_b} \mu^*_{b,y}$ and $\mathcal{W}_{b+1} = \mathcal{W}_b \cup \left(\arg\min_{y \in \Gamma_b} \mu^*_{b,y} \right)$. Then, for all $x \in \arg\min_{y \in \Gamma_b} \mu^*_{b,y}$, $w(x) = \lambda_{b+1}$.

The Whittle index of all information states can be obtained By following the same procedure. This approach is summarized in Algorithm 2.

VI. Numerical Analysis

In this section, we compare the performance of Whittle index policy with the optimal policy and a baseline policy called the myopic policy for various setups. 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. For medium and large-scale models, we only compare with the myopic policy. For the sake of computations, we apply the finite state approximation of information state K_t in both of the models. Next, we briefly describe all of the mentioned algorithms.

A. Policies Compared

*1) Optimal policy (*OPT*):* To compute the optimal policy, we run the standard value iteration method for Problem 1 with respect to the truncated information states corresponding for each model.

Algorithm 2 Computing Whittle index of all information states of model B

Input: Machine $(\mathcal{X}, \{0,1\}, P, Q, c, \rho)$, discount factor β.
Initialize $b = 0$, $\mathcal{W}_b = \emptyset$.
while $\mathcal{W}_b \neq \mathcal{X} \times \mathcal{K}$ **do**
  Compute $\Lambda_{b,y}$ and $\mu^*_{b,y}$, $\forall y \in \Gamma_b$.
  Let $\lambda_{b+1} = \min_{y \in \Gamma_b} \mu^*_{b,y}$ and $\mathcal{W}_{b+1} = \mathcal{W}_b \cup \arg\min_{y \in \Gamma_b} \mu^*_{b,y}$.
  $w(z) = \lambda_{b+1}$, $\forall z \in \arg\min_{y \in \Gamma_b} \mu^*_{b,y}$.
  $b = b + 1$.
end while

*2) Whittle index (*WIP*):* The Whittle index policy is as follows: *at each time, obtain the Whittle index corresponding to current information state of all machines and service m of them which have the highest Whittle indices*. The algorithms for models A and B are shown in Alg. 3 and 4, respectively.

Algorithm 3 Whittle Index Heuristic (Model A)

Compute $w^i(k)$, for all $k \in \mathcal{K}$, and for all $i \in \mathcal{N}$ according to Theorem. 6.
$t = 0$.
while $t \geq 0$ **do**
  Service the machines with the m largest $w^i(K^i_t)$.
  Update K^i_t according to (9) for all $i \in \mathcal{N}$.
  $t = t + 1$.
end while

Algorithm 4 Whittle Index Heuristic (Model B)

Compute $w^i(s,k)$, $\forall k \in \mathcal{K}$, $\forall s \in \mathcal{X}$, and $\forall i \in \mathcal{N}$ according to Alg. 2.
$t = 0$.
while $t \geq 0$ **do**
  Service the machines with the m largest $w^i(S^i_t, K^i_t)$.
  Update K^i_t according to (9) and S^i_t according to (10) for all $i \in \mathcal{N}$.
  $t = t + 1$.
end while

*3) Myopic Policy (*MYP*):* The myopic heuristic that we consider in this section is as follows. At each time step, we sequentially selects m machines. First, we assume one machine has to be selected and we pick the machine which results in lowest per-step cost. Then, we set the machine aside and pick another machine which minimizes the per-step cost among the new collection of machines. This procedure continues until m machines are selected. Then, the selected machines are serviced. The algorithms for models A and B are shown in Alg. 5 and Alg. 6, respectively.

Algorithm 5 Myopic Heuristic (Model A)

$t = 0$.
while $t \geq 0$ **do**
  $\ell = 0$.
  while $\ell \leq m$ **do**
    $i^*_\ell \in \arg\min_{i \in \mathcal{Z}} \sum_{j \in \mathcal{Z} \backslash \{i\}} \bar{c}^j(K^j_t, 0) + \bar{c}^i(K^i_t, 1)$.
    let $\mathcal{M} = \mathcal{M} \cup \{i^*_\ell\}$, $\mathcal{Z} = \mathcal{Z} \setminus \{i^*_\ell\}$.
    $\ell = \ell + 1$.
  end while
  Service the machines with indices collected in $\mathcal{M}$.
  Update K^i_t according to (9) for all $i \in \mathcal{N}$.
  $t = t + 1$.
end while

B. Experiments and Results

We conduct numerical experiments for both models A and B, and vary the number n of machines, the number m of service-persons and the parameters associated with each machine.

Algorithm 6 Myopic Heuristic (Model B)

$t = 0$.
while $t \geq 0$ **do**
  $\ell = 0$.
  while $\ell \leq m$ **do**
    $i^*_\ell \in \arg\min_{i \in \mathcal{Z}} \sum_{j \in \mathcal{Z} \setminus \{i\}} \bar{c}^j(S^j_t, K^j_t, 0) + \bar{c}^i(S^i_t, K^i_t, 1)$.
    let $\mathcal{M} = \mathcal{M} \cup \{i^*_\ell\}$, $\mathcal{Z} = \mathcal{Z} \setminus \{i^*_\ell\}$.
    $\ell = \ell + 1$.
  end while
  Service the machines with indices collected in $\mathcal{M}$.
  Update K^i_t according to (9) and S^i_t according to (10) for all $i \in \mathcal{N}$.
  $t = t + 1$.
end while

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 D. 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 $T = 1000$ averaged over $S = 5000$ 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 $|\mathcal{X}| = 4$, $|\mathcal{K}| = 4$ and $n = 3$, $m = 1$ for both models A and B. In order to model heterogeneous machines, we consider the following. Let $(p_1, \ldots, p_n)$ 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 prospective of the performances, we compute the relative performance of WIP with respect to OPT by computing

$$\alpha_{\text{OPT}} = 100 \times \frac{J(\text{OPT})}{J(\text{WIP})}. \tag{21}$$

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 $|\mathcal{K}| = 40$, 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

$$\varepsilon_{\text{MYP}} = 100 \times \frac{J(\text{MYP}) - J(\text{WIP})}{J(\text{MYP})}. \tag{22}$$

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

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

(a) Model A

ℓ	1	2	3	4
α_{OPT}	100.0	100.0	100.0	100.0

(b) Model B

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

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

(a) Model A, $m = 1$

ε_{MYP}		ℓ			
		1	2	3	4
n	20	1.42	3.20	2.04	6.47
	40	2.45	5.62	4.82	7.09
	60	2.68	4.40	4.33	5.30

(b) Model A, $m = 5$

ε_{MYP}		ℓ			
		1	2	3	4
n	20	0.15	0.27	0.22	1.59
	40	1.09	1.28	1.13	3.79
	60	1.38	2.17	2.14	7.27

C. Discussion

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

In Experiment 2 where WIP is compared with MYP, we observe ε_{MYP} ranges from 0.15 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 overally. Also, as m

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

(a) Model B, $m = 1$

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) Model 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

increases, ε_{MYP} decreases in general. This means 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 considered the problem of scheduling the maintenance of a collection of machines under partial-observations using restless multi-armed bandit problem. We assume each machine has several states, the states of all machines are deteriorating over time and the state dynamics are Markovian. Obtaining the optimal scheduling policy in such a setup is NP-hard. We proposed to model the problem as a RMAB and use the Whittle index policy as a heuristic. This policy is applicable if a technical condition called as indexability is satisfied. We showed that under the assumptions made in Section II, both models are indexable. The Whittle index approach decomposes the problem into a sub-problem for each machine.

Instead of using the belief state formulation which is continuous, we identify a simpler countable information state. Using this information state allowed us to prove some structural results on the problem. We proved that for a single arm problem, the optimal policy is a threshold policy for the first model and a threshold policy with respect to a dimension in the second model. Using these a structural results, we provided a closed-form expression for the Whittle index of the first model and an improved version of adaptive greedy algorithm to compute the Whittle index for the second model.

Finally, we demonstrated 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. Proof of Theorem 1

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

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

Proof. This is an induction from [54, Lemma 4.7.2]. □

Consider a RB $\{(\tilde{\mathcal{X}}, \{0,1\}, \{\tilde{P}, \tilde{Q}\}, \tilde{c}, \tilde{\pi}_0)\}$. According to [43], we say a RB 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 [43, Lemma 5].

Proposition 7. *The optimal policy of a stochastic monotone RB is a threshold policy denoted by $\tilde{g}$, which is a policy which takes passive action for states below a threshold denoted by $\tilde{\theta}$ and active action for the rest of the states, i.e.,*

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

1) Proof of Theorem 1, Part 1: We show that each machine in model A is a stochastic monotone RB. Each condition of stochastic monotone RB is presented and proven for model A below.

(D1') The transition probability matrix under passive action for model A based on the information states is $P^A_{xy} = \mathbb{I}_{\{y=x+1\}}$ and the transition probability matrix under active action for model A is $Q^A_{xy} = \mathbb{I}_{\{y=0\}}$. Thus, P^A and Q^A are stochastic monotone matrices.
(D2') Since P^A is a stochastic monotone matrix and Q^A has constant rows, $\sum_{r \geq z} [P^A - Q^A]_{sr}$ is non-decreasing in s for any $z \in \mathcal{K}$.
(D3') As P stochastically dominates the identity matrix, we infer from [57, 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 (11) 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 $c(x,0) = \phi(x)$ which is increasing in x and $c(x,1) = \rho$ which is a constant, $c_\lambda(x,0) - c_\lambda(x,1)$ is non-decreasing in x. 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.

Therefore, according to Proposition 7, the optimal policy of an RB under model A is a threshold based policy.

2) Proof of Theorem 1, 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 7, 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

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

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

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

Therefore,

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

Thus, $V^{j+1}_\lambda(s,k)$ is non-decreasing in k for any $s \in \mathcal{X}$. This completes the induction step.

$$V_\lambda(s,k) = \lim_{j\to\infty} V^j_\lambda(s,k)$$

and monotonicity is preserved under limits, the induction proof is complete.

c. As $c(x,0) = \phi(x)$ which is increasing in x and $c(x,1) = \rho$, we infer $c(x,0) - c(x,1)$ is increasing in x. 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,

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} [\delta_s(P^k - P^{k+1})]_x c(x,1) \geq\\
&\sum_{x\in\mathcal{X}} [\delta_s(P^k - P^{k+1})]_x c(x,0)\\
&\Rightarrow \sum_{x\in\mathcal{X}} [\delta_s P^k]_x c(x,1) - \sum_{x\in\mathcal{X}} [\delta_s P^k]_x c(x,0) \geq\\
&\sum_{x\in\mathcal{X}} [\delta_s P^{k+1}]_x c(x,1) - \sum_{x\in\mathcal{X}} [\delta_s P^{k+1}]_x c(x,0)\\
&\Rightarrow \bar{c}(s,k,1) - \bar{c}(s,k,0) \geq \bar{c}(s,k+1,1) - \bar{c}(s,k+1,0).
\end{aligned}$$

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 $k \in \mathcal{K}$ and $a \in \{0,1\}$, we have

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

□

Lemma 5. *Suppose* $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *is a submodular function and for each* $x \in \mathcal{X}$, $min_{y\in\mathcal{Y}} f(x,y)$ *exists. Then,* $\max\{\arg\min_{y\in\mathcal{Y}} f(x,y)\}$ *is monotone non-decreasing in* x.

Proof. This is an induction from [54, 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.

B. Proof of Theorem 2

By the strong Markov property, we have

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

and

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

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

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

C. Proof of Theorem 3

By the strong Markov property, we have

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

and

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

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

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

and

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

which results in

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

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

D. 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,

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

REFERENCES

[1] D. L. Kaufman and M. E. Lewis, "Machine maintenance with workload considerations," *Naval Research Logistics (NRL)*, vol. 54, no. 7, pp. 750–766, 2007.

[2] V. Babishin and S. Taghipour, "Optimal maintenance policy for multicomponent systems with periodic and opportunistic inspections and preventive replacements," *Applied Mathematical Modelling*, vol. 40, no. 23-24, pp. 10 480–10 505, 2016.

[3] M. Ben-Daya and M. Rahim, "Effect of maintenance on the economic design of x-control chart," *European Journal of Operational Research*, vol. 120, no. 1, pp. 131–143, 2000.

[4] J. J. McCall, "Maintenance policies for stochastically failing equipment: a survey," *Management science*, vol. 11, no. 5, pp. 493–524, 1965.

[5] H. Pham and H. Wang, "Imperfect maintenance," *European journal of operational research*, vol. 94, no. 3, pp. 425–438, 1996.

[6] H. Wang, "A survey of maintenance policies of deteriorating systems," *European journal of operational research*, vol. 139, no. 3, pp. 469–489, 2002.

[7] B. De Jonge and P. A. Scarf, "A review on maintenance optimization," *European Journal of Operational Research*, 2019.

[8] Y. Liu and H.-Z. Huang, "Optimal replacement policy for multi-state system under imperfect maintenance," *IEEE Transactions on Reliability*, vol. 59, no. 3, pp. 483–495, 2010.

[9] M. Kurt and J. P. Kharoufeh, "Optimally maintaining a Markovian deteriorating system with limited imperfect repairs," *European Journal of Operational Research*, vol. 205, no. 2, pp. 368–380, 2010.

[10] X. Zhao, M. Fouladirad, C. Bérenguer, and L. Bordes, "Condition-based inspection/replacement policies for non-monotone deteriorating systems with environmental covariates," *Reliability Engineering & System Safety*, vol. 95, no. 8, pp. 921–934, 2010.

[11] K. T. Huynh, A. Barros, C. Berenguer, and I. T. Castro, "A periodic inspection and replacement policy for systems subject to competing failure modes due to degradation and traumatic events," *Reliability Engineering & System Safety*, vol. 96, no. 4, pp. 497–508, 2011.

[12] F. Berthaut, A. Gharbi, and K. Dhouib, "Joint modified block replacement and production/inventory control policy for a failure-prone manufacturing cell," *Omega*, vol. 39, no. 6, pp. 642–654, 2011.

[13] Z. Tian and H. Liao, "Condition based maintenance optimization for multi-component systems using proportional hazards model," *Reliability Engineering & System Safety*, vol. 96, no. 5, pp. 581–589, 2011.

[14] Y. Xiang, C. R. Cassady, and E. A. Pohl, "Optimal maintenance policies for systems subject to a Markovian operating environment," *Computers & Industrial Engineering*, vol. 62, no. 1, pp. 190–197, 2012.

[15] S. Taghipour and D. Banjevic, "Optimal inspection of a complex system subject to periodic and opportunistic inspections and preventive replacements," *European Journal of Operational Research*, vol. 220, no. 3, pp. 649–660, 2012.

[16] Y. Xiang, "Joint optimization of x control chart and preventive maintenance policies: a discrete-time Markov chain approach," *European Journal of Operational Research*, vol. 229, no. 2, pp. 382–390, 2013.

[17] C.-Y. Lee and Z.-L. Chen, "Scheduling jobs and maintenance activities on parallel machines," *Naval Research Logistics (NRL)*, vol. 47, no. 2, pp. 145–165, 2000.

[18] 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.

[19] J.-J. Wang, J.-B. Wang, and F. Liu, "Parallel machines scheduling with a deteriorating maintenance activity," *Journal of the Operational Research Society*, vol. 62, no. 10, pp. 1898–1902, 2011.

[20] M. Rebai, I. Kacem, and K. H. Adjallah, "Scheduling jobs and maintenance activities on parallel machines," *Operational Research*, vol. 13, no. 3, pp. 363–383, 2013.

[21] A. Gara-Ali, G. Finke, and M.-L. Espinouse, "Parallel-machine scheduling with maintenance: Praising the assignment problem," *European Journal of Operational Research*, vol. 252, no. 1, pp. 90–97, 2016.

[22] 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.

[23] M. Celen and D. Djurdjanovic, "Integrated maintenance and operations decision making with imperfect degradation state observations," *Journal of Manufacturing Systems*, vol. 55, pp. 302–316, 2020.

[24] E. Byon, L. Ntaimo, and Y. Ding, "Optimal maintenance strategies for wind turbine systems under stochastic weather conditions," *IEEE Transactions on Reliability*, vol. 59, no. 2, pp. 393–404, 2010.

[25] C. C. White, "Optimal control-limit strategies for a partially observed replacement problem," *International Journal of Systems Science*, vol. 10, no. 3, pp. 321–332, 1979.

[26] J. S. Ivy and H. B. Nembhard, "A modeling approach to maintenance decisions using statistical quality control and optimization," *Quality and Reliability Engineering International*, vol. 21, no. 4, pp. 355–366, 2005.

[27] A. H. Elwany, N. Z. Gebraeel, and L. M. Maillart, "Structured replacement policies for components with complex degradation processes and dedicated sensors," *Operations research*, vol. 59, no. 3, pp. 684–695, 2011.

[28] M. Amalnik and M. Pourgharibshahi, "An optimal maintenance policy for machine replacement problem using dynamic programming," *Management science letters*, vol. 7, no. 6, pp. 311–320, 2017.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride, "Some indexable families of restless bandit problems," *Advances in Applied Probability*, vol. 38, no. 3, pp. 643–672, 2006.

[42] 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.

[43] N. Akbarzadeh and A. Mahajan, "Conditions for indexability of restless bandits and an algorithm to compute Whittle index," *arXiv preprint arXiv:2008.06111v3*, 2021.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

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

[51] 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.

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

[53] ——, "Dynamic priority allocation via restless bandit marginal productivity indices," *TOP*, vol. 15, no. 2, pp. 161–198, 2007.

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

[55] 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.

[56] L. I. Sennott, *Stochastic dynamic programming and the control of queueing systems*. John Wiley & Sons, 2009, vol. 504.

[57] J. Keilson and A. Kester, "Monotone matrices and monotone Markov processes," *Stochastic Processes and their Applications*, vol. 5, no. 3, pp. 231–241, 1977.

PLACE PHOTO HERE

Nima Akbarzadeh (S'17) is a PhD student in the Electrical and Computer Engineering, McGill University, Canada. He received the B.Sc. degree in Electrical and Computer Engineering from Shiraz University, Iran, in 2014, the M.Sc. in Electrical and Electronics Engineering from Bilkent University, Turkey, in 2017. He is a recipient of 2020 FRQNT PhD Scholarship. His research interests include stochastic control, reinforcement learning and multi-armed bandits.

PLACE PHOTO HERE

Aditya Mahajan (S'06-M'09-SM'14) is Associate Professor in the the department of Electrical and Computer Engineering, McGill University, Montreal, Canada. He serves as Associate Editor of Springer Mathematics of Control, Signal, and Systems. He was an Associate Editor of the IEEE Control Systems Society Conference Editorial Board from 2014 to 2017. He is the recipient of the 2015 George Axelby Outstanding Paper Award, 2014 CDC Best Student Paper Award (as supervisor), and the 2016 NecSys Best Student Paper Award (as supervisor). His principal research interests include learning and control of centralized and decentralized stochastic systems.