Les Cahiers du GERAD

Maintenance of a collection of machines under partial observability: Indexability and computation of Whittle index

A. Akbarzadeh, A. Mahajan

Maintenance of a collection of machines under partial observability: Indexability and computation of Whittle index

Nima Akbarzadeh [a, b]

Aditya Mahajan [a, b]

[a] GERAD, Montréal (Québec), Canada, H3T 2A7

[b] McGill University – Department of Electrical and Computer Engineering, Montréal (Québec), Canada, H3A 0E9

nima.akbarzadeh@mail.mcgill.ca
aditya.mahajan@mcgill.ca

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.

Keywords: Machine maintenance, partially observable system, scheduling, restless bandits, Whittle index

1 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 [8, 25]. 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 availability [9]. For these reasons, scheduling of machine maintenance has been a problem of interest for several decades [14, 34, 39, 50].

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 [10, 23, 29, 32, 46, 47, 55, 56, 57]. Maintenance of a collection of machines is considered in [16, 21, 31, 42, 51].

The second category arises in applications where the condition of the machine cannot be directly observed or is observed via noisy sensors. Examples include semiconductor manufacturing [12], maintenance of wind turbines [11] or automated demand response devices [2]. Maintenance of a single machine with a single or multiple components is considered in [6, 11, 12, 15, 24, 53]. Maintenance of a collection of machines is considered in [2]. [12] provides a brief 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 [38]. 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 [54], 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 reduces to 100 sub-problems with 3 states.

The restless bandit framework is widely used in scheduling problems arising in various applications including telecommunication networks [1, 4, 22, 26, 27, 30, 35], patient prioritization [49], machine maintenance [2, 21], sensor management [48] and game theory [41].

RMAB models have been considered for both fully-observable [5, 19, 20, 21, 43, 52, 54] and partially-observable [1, 2, 22, 26, 27, 30, 35, 41, 48] 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 [1, 2, 22, 26, 27, 35, 48]. Such models are usually investigated under an additional technical assumption that the states are positively correlated [1, 26, 27]. However, there are only a few papers which consider a general state space under partial observations [4, 13, 30, 41] 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 [1, 2, 22, 26, 27, 35, 48].
- 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 [5] 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 orangization of the paper is as follows. In Section 2, we formulate the machine maintenance problem for two different observation models. In Section 3, we present various simplifications of the model. In Section 4, we present a short overview of restless bandits. In Section 5, we show the restless bandit problem is indexable for both models. In Section 6, we present algorithms to compute Whittle index. In Section 7, we present a detailed numerical study which compares the performance of Whittle index policy with the optimal and myopic policies. In Section 8, we present the conclusions and discussions.

1.1 Notations and definitions

We use $\mathbb{I}$ as the indicator function, $\mathbb{E}$ as the expectation operator, $\mathbb{P}$ as the probability function, $\mathbb{R}$ as the set of real numbers, $\mathbb{Z}$ as the set of integers and $\mathbb{Z}_{\geq 0}$ as the set of positive integers. Calligraphic alphabets are used to denote sets, bold variables are used for the vector of variables. 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 $z \in \mathcal{X}$.

2 Model and problem formulation

2.1 System model

Consider a system operator who has to maintain a collection of n machines, which we index by the set $\mathcal{N} := \{1, \ldots, n\}$. 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.

2.2 Problem formulation

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

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

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 $\boldsymbol{Y}_t = (Y_t^1, \ldots, Y_t^n) \in \prod_{i \in \mathcal{N}} \mathcal{Y}^i$ 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.,

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

where

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

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 $\boldsymbol{g} = (g_1, g_2, \ldots)$ 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

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

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 servicepersons, 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).

2.3   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 3.1.

However, such a transformation suffers from the curse of dimensionality. The belief state space takes values in the simplex $\Delta(\mathcal{X})$ which is double exponential in n. In Section 3.1 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 3.2, 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 4. In Section 5, we show that both models are indexable and in Section 6, we present efficient algorithms to compute Whittle index.

3   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 [7]. We will discuss the details below.

3.1 Belief state formulation for Problem 1

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

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

Note that Θ_t is a random variable that takes values in $\Delta(\mathcal{X})$. Using standard results in POMDPs [7], 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 $\Pi_t^i \in \Delta(\mathcal{X}^i)$ on the state of machine i at time t as follows: for any, $x_t^i \in \mathcal{X}^i$, let

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

Similar to Θ_t, Π_t^i is also a distribution-valued random variable. Let $\boldsymbol{\Pi}_t := (\Pi_t^1, \ldots, \Pi_t^n)$.

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

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

and for model B, the belief update rule is

$$\Pi_{t+1}^i = \begin{cases} \Pi_t^i P^i, & \text{if } A_t^i = 0, \\ \delta_{X_{t+1}^i} \text{ where } X_{t+1}^i \sim Q^i, & \text{if } A_t^i = 1. \end{cases} \tag{6}$$

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

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

Moreover, since the machines are independent, we have

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

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

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

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

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

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

Proof. Equation (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. □

3.2 Information state for Problem 1

Although the transformation presented in Section 3.1 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 [38]. Inspired by the approach taken in [45], we introduce a new information state which is countable and, at the same time, is equivalent to the belief state.

For model A, define

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

and for model B define

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

Assumption 1 *For model A, $\pi_0^i \in \mathcal{R}_A^i$ and for model B, $\pi_0^i \in \mathcal{R}_B^i$.*

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

$$K_t^i = \begin{cases} 0, & \text{if } A_{t-1}^i = 1 \\ K_{t-1}^i + 1, & \text{if } A_{t-1}^i = 0. \end{cases} \tag{9}$$

Similarly, for model B, define a process $\{S_t^i, K_t^i\}_{t \geq 0}$ as follows. The initial state (s_0^i, k_0^i) is such that $\pi_0^i = \delta_{s_0^i} (P^i)^{k_0^i}$ and for $t > 0$, K_t^i evolves according to (9) and S_t^i evolves according to

$$S_t^i = \begin{cases} X_{t-1}^i \text{ where } X_{t-1}^i \sim Q^i, & \text{if } A_{t-1}^i = 1 \\ S_{t-1}^i, & \text{if } A_{t-1}^i = 0. \end{cases} \tag{10}$$

Figure 1 depicts how S_t^i jumps to one of the "corner" points of the simplex upon active action.

Note that once the first observation has been taken in both models, K_t^i denotes the time elapsed since the last observation of machine i and, in addition in model B, S_t^i denotes the last observed states of machine i. Let $\boldsymbol{S}_t := (S_t^1, \dots S_t^n)$ and $\boldsymbol{K}_t := (K_t^1, \dots K_t^n)$. The relation between the belief state Π_t^i and variables S_t^i and K_t^i is characterized in the following lemma.
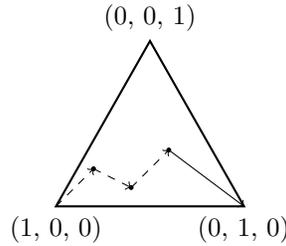
Figure 1: Belief state dynamics for a 3-state machine i in the simplex $\Delta(\{1, 2, 3\})$**.** Dashed arrows show a sample realizations of the belief state evolution under $A_t^i = 0$ for three time steps and the solid arrow shows a sample realization of the belief state evolution under $A_t^i = 1$

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

(i) *For model A, for any $i \in \mathcal{N}$ and any t, $\Pi_t^i \in \mathcal{R}_A^i$. In particular, $\Pi_t^i = Q^i (P^i)^{K_t^i}$.*

(ii) *For model B, for any $i \in \mathcal{N}$ and any t, $\Pi_t^i \in \mathcal{R}_B^i$. In particular, $\Pi_t^i = \delta_{S_t^i} (P^i)^{K_t^i}$.*

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

$$\bar{c}^i(K_t^i, A_t^i) := \bar{c}^i((Q^i P^i)^{K_t^i}, A_t^i) = \sum_{x \in \mathcal{X}^i} [(Q^i P^i)^{K_t^i}]_x c^i(x, A_t^i). \tag{11}$$

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

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

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

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

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

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

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.

4 Overview of restless multi-armed bandits

4.1 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

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

4.2 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 [17]. 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

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

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

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

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, [18] 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, [54] argued that a similar index policy should perform well for general RMAB provided a technical condition known as Whittle indexability is satisfied. Subsequently, [52] and [33] 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 [19, 21]. Furthermore, [36] 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 [37].

4.3 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

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

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

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

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 [40]. $\qquad\square$

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 $\lambda_1, \lambda_2 \in \mathbb{R}$,*

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

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.

4.4 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 [33, 52] and performs close-to-optimal for many other cases [19, 21].

4.5 Sufficient condition for indexability

To verify indexability of our model, we use a recently proposed sufficient condition for indexability [5] 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 ([5]) *If a RB satisfies Condition (R), then it is indexable.*

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

5 Indexability of models A and B

In this section, we show that both models A and B are indexable. For ease of notation, we will drop the superscript i from all relative variables for the rest of this and the next sections.

5.1 Indexability of model A

Model A with the information state representation given in Section 3.2 may be viewed as a RB with state space $\mathbb{Z}_{\geq 0}$, the action space $\{0, 1\}$, the cost function $\bar{c}$ defined in (11), the initial state distribution $\pi_0^A(x) = \mathbb{I}_{\{y=0\}}$ and the transition probability matrices $\{P^A, Q^A\}$ where $P_{xy}^A = \mathbb{I}_{\{y=x+1\}}$ and $Q_{xy}^A = \mathbb{I}_{\{y=0\}}$. Thus, model A may be represented as RB $(\mathbb{Z}_{\geq 0}, \{0, 1\}, \{P^A, Q^A\}, \bar{c}, \pi_0^A)$.

Note that for this model, the modified per-step cost is

$$\bar{c}_\lambda(k, a) = \bar{c}(k, a) + \lambda a, \ k \in \mathbb{Z}_{\geq 0}, \ \forall a \in \mathcal{A}. \tag{17}$$

where $\bar{c}(k, a) = \sum_{x \in \mathcal{X}} [QP^k]_x c(x, a)$. The performance of any time-homogeneous decision policy $g^A : \mathbb{Z}_{\geq 0} \to \{0, 1\}$ given the initial information state k is defined by

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

Let g_λ^A be the optimal policy for a RB under model A obtained by dynamic program given in Proposition 4 with penalty λ. By definition, the passive set for model A is

$$\mathcal{W}_\lambda^A = \left\{k \in \mathbb{Z}_{\geq 0} : g_\lambda^A(k) = 0\right\}.$$

For state $k \in \mathbb{Z}_{\geq 0}$, let $w^A(k)$ denote the Whittle index of this model.

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

Proof. Model A is a RMAB which satisfies the restart property **(R)**. Hence, the problem is indexable by Proposition 5. $\qquad\square$

5.2 Indexability of model B

Similar to model A, we view model B with the information state representation given in Section 3.2 as a RB with the state space $\mathcal{X} \times \mathbb{Z}_{\geq 0}$, the action space $\{0, 1\}$, the cost function $\bar{c}$ defined in (12), the initial state distribution

$$[\pi_0^B]_{(s,k)} = \begin{cases} 0, & \text{if } k \neq 0 \\ [\pi_0]_s, & \text{if } k = 0 \end{cases}$$

and the transition probability matrices $\{P^B, Q^B\}$ where the transition probability from information state (s, k) to (r, l) are given by

$$P_{(s,k)(r,l)}^B = \begin{cases} 0, & \text{if } r \neq s \\ [\delta_s P^k]_l, & \text{if } r = s \end{cases}$$

for passive action and

$$Q_{(s,k)(r,l)}^B = \begin{cases} 0, & \text{if } l \neq 0 \\ Q_r, & \text{if } l = 0 \end{cases}.$$

for active action.

Therefore, model B is a RB given by $(\mathcal{X} \times \mathbb{Z}_{\geq 0}, \{0, 1\}, \{P^B, Q^B\}, \bar{c}, \pi_0^B)$.

Note that for this model, the modified per-step cost is

$$\bar{c}_\lambda(s, k, a) = \bar{c}(s, k, a) + \lambda a, \ s \in \mathcal{X}, \ k \in \mathbb{Z}_{\geq 0}, \ \forall a \in \mathcal{A}. \tag{19}$$

where $\bar{c}(s,k,a) = \sum_{x \in \mathcal{X}}[\delta_s P^k]_x c(x,a)$, defined in (12). The performance of any time-homogeneous decision policy $g^B : \mathbb{Z}_{\geq 0} \to \{0,1\}$ given the initial information state (s,k) is defined by

$$J_\lambda^{(g^B)}(s,k) := (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(S_t, K_t, g^B(S_t, K_t))\Big| S_0 = s, K_0 = k\right]. \tag{20}$$

Let g_λ^B be the optimal policy for a RB under model B obtained by dynamic program given in Proposition 4 with penalty λ. By definition, the passive set for model B is

$$\mathcal{W}_\lambda^B = \left\{(s,k) \in \mathcal{X} \times \mathbb{Z}_{\geq 0} : g_\lambda^B(s,k) = 0\right\}$$

For information state $(s,k) \in \mathcal{X} \times \mathbb{Z}_{\geq 0}$, let $w^B(s,k)$ denote the Whittle index of this model.

Proposition 7 *Problem 3 for model B is indexable for any arm $i \in \mathcal{N}$.*

Proof. Similar to model A, model B is also a RMAB which satisfies the restart property **(R)**. Hence, the problem is indexable by Proposition 5. □

6 Computation of Whittle index

In the previous section, we showed that both models are indexable. In this section, 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 [1, 3, 5, 19, 20, 21, 26, 27, 30]. 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 [36, 37]. Recently, [5] 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 [5].

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.

6.1 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:*

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

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

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

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

The proof is given in Appendix A1.

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

6.2 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 $g^{(\boldsymbol{\theta}^B)}$.

6.2.1 Model A

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

$$
\begin{aligned}
J_\lambda^{(\theta^A)}(k) &:= (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(K_t, g^{(\theta^A)}(K_t)) \,\Big|\, K_0 = k\right] \\
&= (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \big(\bar{c}(K_t, g^{(\theta^A)}(K_t)) + \lambda g^{(\theta^A)}(K_t)\big) \,\Big|\, K_0 = k\right] \\
&:= D^{(\theta^A)}(k) + \lambda N^{(\theta^A)}(k),
\end{aligned}
\tag{21}
$$

where

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

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

$D^{(\boldsymbol{\theta}^A)}(k)$ 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

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

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

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

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

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

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

The proof is given in Appendix A2.

6.2.2 Model B

Let $J_\lambda^{(\boldsymbol{\theta}^B)}(s, k)$ 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.,

$$
\begin{aligned}
J_\lambda^{(\boldsymbol{\theta}^B)}(s, k) &= (1 - \beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}_\lambda(S_t, K_t, g^{(\boldsymbol{\theta}^B)}(S_t, K_t)) \,\Big|\, (S_0, K_0) = (s, k)\right] \\
&= (1 - \beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}(S_t, K_t, g^{(\boldsymbol{\theta}^B)}(S_t, K_t)) + \beta^t \lambda g^{(\boldsymbol{\theta}^B)}(S_t, K_t) \,\Big|\, (S_0, K_0) = (s, k)\right] \\
&:= D^{(\boldsymbol{\theta}^B)}(s, k) + \lambda N^{(\boldsymbol{\theta}^B)}(s, k),
\end{aligned}
\tag{22}
$$

where

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

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

$D^{(\boldsymbol{\theta}^B)}(s, k)$ 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 $\boldsymbol{J}_\lambda^{(\boldsymbol{\theta}^B)}(0) = (J_\lambda^{(\boldsymbol{\theta}^B)}(1, 0), \ldots, J_\lambda^{(\boldsymbol{\theta}^B)}(|\mathcal{X}|, 0))$ and vectors $\boldsymbol{D}^{(\boldsymbol{\theta}^B)}(0)$ and $\boldsymbol{N}^{(\boldsymbol{\theta}^B)}(0)$ in a similar manner. Then, from (21), $\boldsymbol{J}_\lambda^{(\boldsymbol{\theta}^B)}(0) = \boldsymbol{D}^{(\boldsymbol{\theta}^B)}(0) + \lambda \boldsymbol{N}^{(\boldsymbol{\theta}^B)}(0)$. Let's also define

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

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

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*

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

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

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

$$
\boldsymbol{D}^{(\boldsymbol{\theta}^B)}(0) = (I - Z^{(\boldsymbol{\theta}^B)})^{-1} \boldsymbol{L}^{(\boldsymbol{\theta}^B)}(0),
$$

$$
\boldsymbol{N}^{(\boldsymbol{\theta}^B)}(0) = (I - Z^{(\boldsymbol{\theta}^B)})^{-1} \boldsymbol{M}^{(\boldsymbol{\theta}^B)}(0).
$$

The proof is given in Apendix A3.

6.3 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 $\mathbb{N}_\ell := \{0, \ldots, \ell\}$ and $V_{\ell,\lambda} : \mathcal{X} \times \mathbb{N}_\ell \to \mathbb{R}$ be the unique fixed point of equation*

$$V_{\ell,\lambda}(k) = \min_{a \in \{0,1\}} H_\lambda(k, a), \ \hat{g}_{\ell,\lambda}(k) = \arg \min_{a \in \{0,1\}} H_{\ell,\lambda}(k, a) \tag{23}$$

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

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

(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 $k \in \{0, \ldots, \ell - 1\}$, the cost incurred by $\hat{g}_{\ell,\lambda}(\cdot)$ is the same as $g_\lambda^A(\cdot)$ for information states $\{k, \ldots, \ell\}$. The per-step cost incurred by $\hat{g}_{\ell,\lambda}(\cdot)$ differs from $g_\lambda^A(\cdot)$ for information states $\{\ell + 1, \ldots\}$ by at most $c_{\text{diff},\lambda}^{\max}$.

(ii) The sequence of finite-state models described above is an *augmentation type approximation sequence* (see [44, Definition 2.5.3]). As a result, a limit point of $\hat{g}_\lambda^*$ exists and the final result holds by [44, Proposition B.5, Theorem 4.6.3].

□

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

$$V_{\ell,\lambda}(s, k) = \min_{a \in \{0,1\}} H_\lambda(s, k, a), \ \hat{g}_{\ell,\lambda}(s, k) = \arg \min_{a \in \{0,1\}} H_{\ell,\lambda}(s, k, a) \tag{24}$$

where

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

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 $c_{diff,\lambda}^{\max} = \max\{\max_x \phi(x), \rho + \lambda\} - \min\{\min_x \phi(x), \rho + \lambda\}$, then*

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

(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 $k \in \{0, \ldots, \ell - 1\}$, the cost incurred by $\hat{g}_{\ell, \lambda}(\cdot, \cdot)$ is the same as $g_\lambda^B(\cdot, \cdot)$ for information states $\{(s, l)\}_{l=k}^\ell$. The per-step cost incurred by $\hat{g}_{\ell, \lambda}(\cdot, \cdot)$ differs from $g_\lambda^B(\cdot, \cdot)$ for later realized information states by at most $c_{\text{diff}, \lambda}^{\max}$. Thus, the bound would hold.

(ii) The sequence of finite-state models described above is an *augmentation type approximation sequence* (see [44, Definition 2.5.3]). As a result, a limit point of $\hat{g}_\lambda^*$ exists and the final result holds [44, Proposition B.5, Theorem 4.6.3].

$\square$

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 $\mathcal{K} = \mathbb{N}_\ell$.

6.4 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.

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

$$w^A(k) = \min_{k_0 \in \Lambda_k^A} \frac{D^{(k+1)}(k_0) - D^{(k)}(k_0)}{N^{(k)}(k_0) - N^{(k+1)}(k_0)}. \tag{25}$$

Proof. The proof results from [5, Lemma 4]. $\square$

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

6.4.2 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 $\Lambda^* = \{\lambda_0, \lambda_1, \ldots, \lambda_{B_D}\}$ where $\lambda_1 < \lambda_2 < \ldots < \lambda_{B_D}$ denote the sorted distinct Whittle indices with $\lambda_0 = -\infty$. Let $\mathcal{W}_b := \{(s, k) \in \mathcal{X} \times \mathcal{K} : w(s, k) \leq \lambda_b\}$. 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 $\Gamma_b = \{(s, k) \in (\mathcal{X} \times \mathcal{K}) \backslash \mathcal{W}_b : (s, \max\{0, k - 1\}) \in \mathcal{W}_b\}$. Additionally, for any $b \in \{0, \ldots, B_D - 1\}$, 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

$$\mu_{b,y}(x, k) = \frac{D^{(h_{b,y})}(x, k) - D^{(h_b)}(x, k)}{N^{(h_b)}(x, k) - N^{(h_{b,y})}(x, k)}. \tag{26}$$

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

(i) For all $y \in \mathcal{W}_{b+1} \backslash \mathcal{W}_b$, we have $w(y) = \lambda_{b+1}$.

(ii) For all $y \in \Gamma_b$ 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. See [5, Lemma 3]. The only difference is that [5] 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:*

(i) For any $y \in \mathcal{W}_{b+1} \backslash \mathcal{W}_b$, the set $\Lambda_{b,y}$ is non-empty.

(ii) 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 [5, 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 1.

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

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

---

7   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.

7.1   Policies compared

7.1.1   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.

7.1.2 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 Algorithms 2 and 3, respectively.

Algorithm 2 Whittle index heuristic (Model A)

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

Algorithm 3 Whittle index heuristic (Model B)

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

7.1.3 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 Algorithm 4 and Algorithm 5, respectively.

Algorithm 4 Myopic heuristic (Model A)

1: $t = 0$.
2: **while** $t \geq 0$ **do**
3:     $\ell = 0$.
4:     **while** $\ell \leq m$ **do**
5:         $i_\ell^* \in \arg\min_{i \in \mathcal{Z}} \sum_{j \in \mathcal{Z} \setminus \{i\}} \bar{c}^j(K_t^j, 0) + \bar{c}^i(K_t^i, 1)$.
6:         let $\mathcal{M} = \mathcal{M} \cup \{i_\ell^*\}$, $\mathcal{Z} = \mathcal{Z} \setminus \{i_\ell^*\}$.
7:         $\ell = \ell + 1$.
8:     **end while**
9:     Service the machines with indices collected in $\mathcal{M}$.
10:    Update K_t^i according to (9) for all $i \in \mathcal{N}$.
11:    $t = t + 1$.
12: **end while**

7.2 Experiments and results

We conduct numerical experiments for both models A and B, and vary the number n of machines, the number m of service-persons and the parameters associated with each machine. There are three parameters associated with each machine: the deterioration probability matrix P^i, the reset pmf Q^i and the per-step cost $c^i(x, a)$. We assume the matrix P^i is chosen from a family of four types of structured transition matrices $\mathcal{P}_\ell(p)$, $\ell \in \{1, 2, 3, 4\}$ where p is a parameter of the model. The details of all these models are presented in Appendix A4. 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

---

Algorithm 5 Myopic heuristic (Model B)

1: $t = 0$.
2: **while** $t \geq 0$ **do**
3:      $\ell = 0$.
4:      **while** $\ell \leq m$ **do**
5:          $i_\ell^* \in \arg\min_{i \in \mathcal{Z}} \sum_{j \in \mathcal{Z} \setminus \{i\}} \bar{c}^j(S_t^j, K_t^j, 0) + \bar{c}^i(S_t^i, K_t^i, 1)$.
6:          let $\mathcal{M} = \mathcal{M} \cup \{i_\ell^*\}$, $\mathcal{Z} = \mathcal{Z} \setminus \{i_\ell^*\}$.
7:          $\ell = \ell + 1$.
8:      **end while**
9:      Service the machines with indices collected in $\mathcal{M}$.
10:     Update K_t^i according to (9) and S_t^i according to (10) for all $i \in \mathcal{N}$.
11:     $t = t + 1$.
12: **end while**

---

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

$$\alpha_{\text{OPT}} = 100 \times \frac{J(\text{OPT})}{J(\text{WIP})}. \tag{27}$$

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

Table 1: α_{opt} for different choice of parameters in Experiment 1

(a) Model A

ℓ	1	2	3	4
α_{OPT}	100	100	100	100

(b) Model A

ℓ	1	2	3	4
α_{OPT}	100	99.72	99.81	99.57

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

$$\varepsilon_{\text{MYP}} = 100 \times \frac{J(\text{MYP}) - J(\text{WIP})}{J(\text{MYP})}. \tag{28}$$

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 2 and 3, respectively.

Table 2: $\varepsilon_{\mathbf{myp}}$ for different choice of parameters of Model A in Experiment 2

(a) Model A, $m = 1$

	$\varepsilon_{\mathrm{MYP}}$	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, $m = 5$

	$\varepsilon_{\mathrm{MYP}}$	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

Table 3: $\varepsilon_{\mathbf{myp}}$ for different choice of parameters of Model B in Experiment 2

(a) Model B, $m = 1$

	$\varepsilon_{\mathrm{MYP}}$	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, $m = 5$

	$\varepsilon_{\mathrm{MYP}}$	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

7.3 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 $\varepsilon_{\mathrm{MYP}}$ ranges from 0.15 to 14.5. In a similar interpretation as Experiment 1, as model B is more complex than model A, $\varepsilon_{\mathrm{MYP}}$ for model B is higher than the ones model A given the same set of parameters. Furthermore, we observe that as n increases, $\varepsilon_{\mathrm{MYP}}$ also increases overall. Also, as m increases, $\varepsilon_{\mathrm{MYP}}$ 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.

8 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 2, 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

A1   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 A1 *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 a straightforward induction from [40, Lemma 4.7.2].   □

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

Proposition 8 *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.,*

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

A1.1   Proof of Theorem 1, Part (i)

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-A) Recall that $P^A_{xy} = \mathbb{I}_{\{y=x+1\}}$ and $Q^A_{xy} = \mathbb{I}_{\{y=0\}}$. Thus, P^A and Q^A are stochastic monotone matrices.

(D2-A) 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-A) As P stochastically dominates the identity matrix, we infer from [28, 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 A1, $\bar{c}_\lambda(k, a)$ is non-decreasing in k.

(D4-A) 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-A), QP^ℓ stochastically dominates QP^k for any $\ell > k \geq 0$. Therefore, by Lemma A1, $\bar{c}_\lambda(k, 0) - \bar{c}_\lambda(k, 1) = \sum_{x \in \mathcal{X}} [(QP)^k]_x (c_\lambda(x, 0) - c_\lambda(x, 1))$ is non-decreasing in k.

Therefore, according to Proposition 8, the optimal policy of an RB under model A is a threshold based policy.

A1.2 Proof of Theorem 1, Part (ii)

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

Lemma A2 *We have*

(i) $\bar{c}_\lambda(s,k,a)$ *is increasing in k for any $s \in \mathcal{X}$ and $a \in \{0,1\}$.*

(ii) *Given a fixed λ, $V_\lambda(s,k)$ is increasing in k for any $s \in \mathcal{X}$.*

(iii) $\bar{c}_\lambda(s,k,a)$ *is submodular in (k,a), for any $s \in \mathcal{X}$.*

(iv) $H_\lambda(s,k,a)$ *is submodular in (k,a), for any $s \in \mathcal{X}$.*

Proof. The proof of each part is as follows.

(i) 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-A) in Proposition 8, 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.

(ii) Let

$$H_\lambda^j(s,k,0) := (1-\beta)\bar{c}(s,k,0) + \beta V_\lambda^j(s,k+1),$$

$$H_\lambda^j(s,k,1) := (1-\beta)\bar{c}(s,k,1) + (1-\beta)\lambda + \beta \sum_r Q_r V_\lambda^j(r,0),$$

$$V_\lambda^{j+1}(s,k) := \min_{a \in \{0,1\}} \{H_\lambda^j(s,k,a)\},$$

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

Claim 1 $V_\lambda^j(s,k)$ *is non-decreasing in k for any $s \in \mathcal{X}$ and $j \geq 0$.*

We prove the claim by induction. By construction, $V_\lambda^0(s,k)$ is non-decreasing in k for any $s \in \mathcal{X}$. This forms the basis of induction. Now assume that $V_\lambda^j(s,k)$ is non-decreasing in k is for any $s \in \mathcal{X}$ and some $j \geq 0$. Consider $\ell > k \geq 0$. Then, by induction hypothesis we have

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

Therefore,

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

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

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

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

(iii) 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 sth row of P^k. Thus, $\delta_s P^{k+1}$ stochastically dominates $\delta_s P^k$ and by Lemma A1 we have

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

Therefore,

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

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

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

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

(iv) 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

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

$\square$

Lemma A3 *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 a straightforward induction from [40, Lemma 4.7.1]. $\square$

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 A3 and as only two actions is available, the optimal policy is a threshold policy specified in the theorem statement.

A2   Proof of Theorem 2

By the strong Markov property, we have

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

and

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

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

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

A3 Proof of Theorem 3

By the strong Markov property, we have

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

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

and

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

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

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

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

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

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

A4 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.

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

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

Matrix $\mathcal{P}_2(p)$: Similar to $\mathcal{P}_1(p)$ with $q_1 = (1-p)/2$ and $q_2 = (1-p)/2$.
Matrix $\mathcal{P}_3(p)$: Similar to $\mathcal{P}_1(p)$ with $q_1 = 2(1-p)/3$ and $q_2 = (1-p)/3$.
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] Samuli Aalto, Pasi Lassila, and Prajwal Osti. Whittle index approach to size-aware scheduling for time-varying channels with multiple states. Queueing Systems, 83(3–4):195–225, 2016.

[2] Carlos Abad and Garud Iyengar. A near-optimal maintenance policy for automated DR devices. IEEE Transactions on Smart Grid, 7(3):1411–1419, 2016.

[3] N. Akbarzadeh and A. Mahajan. Restless bandits with controlled restarts: Indexability and computation of whittle index. In 2019 IEEE 58th Conference on Decision and Control (CDC), pages 7294–7300, 2019.

[4] Nima Akbarzadeh and Aditya Mahajan. Dynamic spectrum access under partial observations: A restless bandit approach. In Canadian Workshop on Information Theory, pages 1–6. IEEE, 2019.

[5] Nima Akbarzadeh and Aditya Mahajan. Conditions for indexability of restless bandits and an algorithm to compute whittle index. arXiv preprint arXiv:2008.06111v3, 2021.

[6] M Amalnik and Morteza Pourgharibshahi. An optimal maintenance policy for machine replacement problem using dynamic programming. Management science letters, 7(6):311–320, 2017.

[7] Karl J Astrom. Optimal control of Markov processes with incomplete state information. Journal of mathematical analysis and applications, 10(1):174–205, 1965.

[8] Vladimir Babishin and Sharareh Taghipour. Optimal maintenance policy for multicomponent systems with periodic and opportunistic inspections and preventive replacements. Applied Mathematical Modelling, 40(23–24):10480–10505, 2016.

[9] Mohamed Ben-Daya and MA Rahim. Effect of maintenance on the economic design of x-control chart. European Journal of Operational Research, 120(1):131–143, 2000.

[10] F Berthaut, Ali Gharbi, and Karem Dhouib. Joint modified block replacement and production/inventory control policy for a failure-prone manufacturing cell. Omega, 39(6):642–654, 2011.

[11] Eunshin Byon, Lewis Ntaimo, and Yu Ding. Optimal maintenance strategies for wind turbine systems under stochastic weather conditions. IEEE Transactions on Reliability, 59(2):393–404, 2010.

[12] Merve Celen and Dragan Djurdjanovic. Integrated maintenance and operations decision making with imperfect degradation state observations. Journal of Manufacturing Systems, 55:302–316, 2020.

[13] Christopher R Dance and Tomi Silander. Optimal policies for observing time series and related restless bandit problems. J. Mach. Learn. Res., 20:35–1, 2019.

[14] Bram De Jonge and Philip A Scarf. A review on maintenance optimization. European Journal of Operational Research, 2019.

[15] Alaa H Elwany, Nagi Z Gebraeel, and Lisa M Maillart. Structured replacement policies for components with complex degradation processes and dedicated sensors. Operations research, 59(3):684–695, 2011.

[16] Ahmed Gara-Ali, Gerd Finke, and Marie-Laure Espinouse. Parallel-machine scheduling with maintenance: Praising the assignment problem. European Journal of Operational Research, 252(1):90–97, 2016.

[17] John Gittins, Kevin Glazebrook, and Richard Weber. Multi-armed bandit allocation indices. John Wiley & Sons, 2011.

[18] John C. Gittins. Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society. Series B (Methodological), pages 148–177, 1979.

[19] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride. Some indexable families of restless bandit problems. Advances in Applied Probability, 38(3):643–672, 2006.

[20] KD Glazebrook, DJ Hodge, and Christopher Kirkbride. Monotone policies and indexability for bidirectional restless bandits. Advances in Applied Probability, 45(1):51–85, 2013.

[21] Kevin 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, 165(1):267–284, 2005.

[22] Sudipto Guha, Kamesh Munagala, and Peng Shi. Approximation algorithms for restless bandit problems. Journal of the ACM (JACM), 58(1):3, 2010.

[23] Khac Tuan Huynh, Anne Barros, Christophe Berenguer, and Inmaculada Torres Castro. A periodic inspection and replacement policy for systems subject to competing failure modes due to degradation and traumatic events. Reliability Engineering & System Safety, 96(4):497–508, 2011.

[24] Julie Simmons Ivy and Harriet Black Nembhard. A modeling approach to maintenance decisions using statistical quality control and optimization. Quality and Reliability Engineering International, 21(4):355–366, 2005.

[25] David L Kaufman and Mark E Lewis. Machine maintenance with workload considerations. Naval Research Logistics (NRL), 54(7):750–766, 2007.

[26] Kesav Kaza, Varun Mehta, Rahul Meshram, and SN Merchant. Restless bandits with cumulative feedback: Applications in wireless networks. In 2018 IEEE Wireless Communications and Networking Conference (WCNC), pages 1–6. IEEE, 2018.

[27] Kesav Kaza, Rahul Meshram, Varun Mehta, and Shabbir N Merchant. Sequential decision making with limited observation capability: Application to wireless networks. IEEE Transactions on Cognitive Communications and Networking, 5(2):237–251, 2019.

[28] Julian Keilson and Adri Kester. Monotone matrices and monotone Markov processes. Stochastic Processes and their Applications, 5(3):231–241, 1977.

[29] Murat Kurt and Jeffrey P Kharoufeh. Optimally maintaining a Markovian deteriorating system with limited imperfect repairs. European Journal of Operational Research, 205(2):368–380, 2010.

[30] Maialen Larrañaga, Mohamad Assaad, Apostolos Destounis, and Georgios S Paschos. Dynamic pilot allocation over Markovian fading channels: A restless bandit approach. In 2016 IEEE Information Theory Workshop (ITW), pages 290–294. IEEE, 2016.

[31] Chung-Yee Lee and Zhi-Long Chen. Scheduling jobs and maintenance activities on parallel machines. Naval Research Logistics (NRL), 47(2):145–165, 2000.

[32] Yu Liu and Hong-Zhong Huang. Optimal replacement policy for multi-state system under imperfect maintenance. IEEE Transactions on Reliability, 59(3):483–495, 2010.

[33] Christopher Lott and Demosthenis 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, 14(3):259–297, 2000.

[34] John J McCall. Maintenance policies for stochastically failing equipment: a survey. Management science, 11(5):493–524, 1965.

[35] Rahul Meshram, D Manjunath, and Aditya Gopalan. On the Whittle index for restless multiarmed hidden Markov bandits. 63(9):3046–3053, 2018.

[36] José Niño-Mora. Restless bandits, partial conservation laws and indexability. Advances in Applied Probability, 33(1):76–98, 2001.

[37] José Niño-Mora. Dynamic priority allocation via restless bandit marginal productivity indices. TOP, 15 (2):161–198, 2007.

[38] Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queuing network control. Mathematics of Operations Research, 24(2):293–305, 1999.

[39] Hoang Pham and Hongzhou Wang. Imperfect maintenance. European journal of operational research, 94 (3):425–438, 1996.

[40] Martin L. Puterman. Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons, 2014.

[41] Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In Int. Conf. on Autonomous Agents & Multiagent Systems, pages 123–131, 2016.

[42] Maher Rebai, Imed Kacem, and Kondo H Adjallah. Scheduling jobs and maintenance activities on parallel machines. Operational Research, 13(3):363–383, 2013.

[43] Diego Ruiz-Hernández, Jesús M Pinar-Pérez, and David Delgado-Gómez. Multi-machine preventive maintenance scheduling with imperfect interventions: A restless bandit approach. Computers & Operations Research, 119:104927, 2020.

[44] Linn I Sennott. Stochastic dynamic programming and the control of queueing systems, volume 504. John Wiley & Sons, 2009.

[45] David I Shuman, Ashutosh Nayyar, Aditya Mahajan, Yuriy Goykhman, Ke Li, Mingyan Liu, Demosthenis Teneketzis, Mahta Moghaddam, and Dara Entekhabi. Measurement scheduling for soil moisture sensing: From physical models to optimal control. Proceedings of the IEEE, 98(11):1918–1933, 2010.

[46] Sharareh Taghipour and Dragan Banjevic. Optimal inspection of a complex system subject to periodic and opportunistic inspections and preventive replacements. European Journal of Operational Research, 220(3):649–660, 2012.

[47] Zhigang Tian and Haitao Liao. Condition based maintenance optimization for multi-component systems using proportional hazards model. Reliability Engineering & System Safety, 96(5):581–589, 2011.

[48] Sofía S Villar. Indexability and optimal index policies for a class of reinitialising restless bandits. Probability in the engineering and informational sciences, 30(1):1–23, 2016.

[49] Sofía S Villar, Jack Bowden, and James 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, 30(2):199, 2015.

[50] Hongzhou Wang. A survey of maintenance policies of deteriorating systems. European journal of operational research, 139(3):469–489, 2002.

[51] Jian-Jun Wang, Ji-Bo Wang, and Feng Liu. Parallel machines scheduling with a deteriorating maintenance activity. Journal of the Operational Research Society, 62(10):1898–1902, 2011.

[52] Richard R Weber and Gideon Weiss. On an index policy for restless bandits. Journal of Applied Probability, 27(3):637–648, 1990.

[53] Chelsea C White. Optimal control-limit strategies for a partially observed replacement problem. International Journal of Systems Science, 10(3):321–332, 1979.

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

[55] Yisha Xiang. Joint optimization of x control chart and preventive maintenance policies: a discrete-time Markov chain approach. European Journal of Operational Research, 229(2):382–390, 2013.

[56] Yisha Xiang, C Richard Cassady, and Edward A Pohl. Optimal maintenance policies for systems subject to a Markovian operating environment. Computers & Industrial Engineering, 62(1):190–197, 2012.

[57] Xuejing Zhao, Mitra Fouladirad, Christophe Bérenguer, and Laurent Bordes. Condition-based inspection/replacement policies for non-monotone deteriorating systems with environmental covariates. Reliability Engineering & System Safety, 95(8):921–934, 2010.