Restless bandits with controlled restarts: Indexability and computation of Whittle index

Nima Akbarzadeh and Aditya Mahajan

***Abstract*—Motivated by applications in machine repair, queueing, surveillance, and clinic care, we consider a scheduling problem where a decision maker can reset m out of n Markov processes at each time. Processes that are reset, restart according to a known probability distribution and processes that are not reset, evolve in a Markovian manner. Due to the high complexity of finding an optimal policy, such scheduling problems are often modeled as restless bandits. We show that the model satisfies a technical condition known as indexability. For indexable restless bandits, the Whittle index policy, which computes a function known as Whittle index for each process and resets the m processes with the lowest index, is known to be a good heuristic. The Whittle index is computed by solving an auxiliary Markov decision problem for each arm. When the optimal policy for this auxiliary problem is threshold based, we use ideas from renewal theory to derive closed form expression for the Whittle index. We present detailed numerical experiments which suggest that Whittle index policy performs close to the optimal policy and performs significantly better than myopic policy, which is a commonly used heuristic.**

I. Introduction

A. Motivation

In this paper, we investigate scheduling problems where a decision maker can reset m out of n Markov processes at each time. Processes that are reset, restart according to a known probability distribution and processes that are not reset, evolve in a Markovian manner. Such problems arise in a variety of applications such as queueing [1], surveillance [2], smart grid [3], machine maintenance [4], [5] and clinical care [6]. Such scheduling problems are Markov decision processes where the state space is exponential in the number of alternatives. So, computing an exact solution is often intractable and one has to resort to heuristics to identify a good solution. Such scheduling problems of interest belong to a class of models known as restless bandits [7] which are generalization of multi-armed bandits [8]–[10].

Multi-armed bandits [8]–[10] are sequential decision making problems where a decision maker has to *activate* m out of n alternatives at each time. The alternatives that are activated evolve in a Markovian manner. Those that are not activated (i.e., are *passive*) either remain frozen or evolve in a Markovian manner (different from the active ones). Each process generates a reward or incurs a cost that depends on its state and the active or passive action.

When $m = 1$ and passive arms remain frozen, the model is known as the classical multi-armed bandits. For this model, Gittins [8] showed that the optimal policy has a simple structure: at each time, compute an index (known as Gittins index) for each arm and play the arm with the highest index.

The general case when the passive action arms also evolve is known as *restless bandits* [7]. Gittins index policy not optimal for such models, but it was argued by Whittle that if the model satisfies a technical condition known as *indexability*, then a modification of the Gittins index known as *Whittle index* is still a reasonable heuristic. Whittle index policy is a popular approach for restless bandits because: (i) its complexity is linear in the number of alternatives and (ii) it often performs close to optimal in practice [5], [11], [12].

Two steps need to be carried out for using the Whittle index policy: (i) check whether the model is Whittle indexable, and if so, (ii) find a low-complexity method to compute the Whittle index for each alternative. Although some results are available for specific models [4], [5], [13], [14], there is no general framework for checking Whittle indexability or for computing the Whittle index. These steps are often carried out on a case-by-case basis by exploiting the specific features of the model.

Motivated by the models in [1]–[6], we propose a model for what we call restless bandits with controlled restarts. We show that irrespective of the choice of the model parameters, the problem of restless bandits with controlled restarts is always indexable. For indexable models, the Whittle index ix computed by solving an auxiliary Markov decision problem. When the optimal policy for this auxiliary problem is threshold based, we use ideas from renewal theory to derive closed form expression for the Whittle index.

B. Notation

Uppercase letters (X, Y, etc.) denote random variables, the corresponding lowercase letters (x, y, etc.) denote their realization, and the corresponding script letters ($\mathcal{X}$, $\mathcal{Y}$, etc.) denote their state spaces. Subscripts denote time: so, X_t denotes a system variable at time t and $X_{1:t}$ is a short-hand for the system variables $(X_1, \dots, X_t)$. $\mathbb{P}(\cdot)$ denotes the probability of an event and $\mathbb{E}[\cdot]$ denotes the expectation of a random variable.

$\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{R}_{\geq 0}$ denote the sets of integers, real numbers, and positive real numbers, respectively. $\mathbb{I}$ denotes the indicator function, Given a matrix P, $P_{i,j}$ denotes its (i,j)-th element.

Given ordered sets $\mathcal{X}$ and $\mathcal{Y}$, a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is called submodular[1] 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

$$f(x_1, y_2) - f(x_1, y_1) \geq f(x_2, y_2) - f(x_2, y_1).$$

[1] Submodular functions satisfy the following useful property [15]. Given ordered sets $\mathcal{X}$ and $\mathcal{Y}$ and a submodular function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the function $g(x) = \min_{y \in \mathcal{Y}} f(x, y)$ is (weakly) increasing in x provided the $\arg\min$ exists.

II. Model and Problem Formulation

Consider a discrete time system with n *arms* and let $\mathcal{N} = \{1, \dots, n\}$ denote the set of arms. Each arm i, $i \in \mathcal{N}$, is a controlled Markov process with state space $\mathcal{X}^i$ and action space $\{0, 1\}$. For ease of exposition, we assume that $\mathcal{X}^i$ is a finite set. Let $X_t^i \in \mathcal{X}^i$ denote the state of arm i and $A_t^i \in \{0, 1\}$ denote the action applied to arm i at time t. Furthermore, let $\boldsymbol{X}_t$ denote $(X_t^1, \dots, X_t^n)$, $\boldsymbol{A}_t$ denote $(A_t^1, \dots, A_t^n)$, and $\boldsymbol{\mathcal{X}}$ denote $\mathcal{X}^1 \times \ldots \times \mathcal{X}^n$. We assume that the arms evolve in a Markovian manner independently from each other, i.e., for any $\boldsymbol{x}_t = \left(x_t^1, \dots, x_t^n\right)$ and $\boldsymbol{a}_t \coloneqq \left(a_t^1, \dots, a_t^n\right)$, we have

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{X}_{t+1} = \boldsymbol{x}_{t+1} | \boldsymbol{X}_{1:t} = \boldsymbol{x}_{1:t}, \boldsymbol{A}_{1:t} = \boldsymbol{a}_{1:t}\right) \\ = \prod_{i=1}^{n} \mathbb{P}\left(X_{t+1}^i = x_{t+1}^i | X_t^i = x_t^i, A_t^i = a_t^i\right).\end{aligned}$$

When $a_t^i = 0$, we say that arm i is *passive* at time t; when $a_t^i = 1$, we say that arm i is *active* at time t. Arm $i \in \mathcal{N}$ evolves as follows: for any $x, y \in \mathcal{X}^i$ and $a \in \{0, 1\}$ we have

$$\mathbb{P}(X_{t+1}^i = y | X_t^i = x, A_t^i = a) = \begin{cases} P_{xy}^i, & \text{if } a = 0 \\ Q_y^i, & \text{if } a = 1 \end{cases}$$

Thus, when arm i is passive, it evolves in a Markov manner according to transition probabilities P^i; when arm i is active, the state of arm i resets according to probability mass function Q^i, which we call it as *reset pmf*.

When arm i in state x is passive it incurs a cost $c^i(x, 0)$; when it is active, it incurs a cost $c^i(x, 1)$. When the system is in state $\boldsymbol{x}_t$ and action $\boldsymbol{a}_t$ is taken, the system incurs a per-step cost given by

$$\sum_{i=1}^{n} c^i(x_t^i, a_t^i).$$

At each time, a decision-maker observes the state of all the arms and can reset m of them where $m < n$. Let $\boldsymbol{\mathcal{A}}(m)$ be a subset of actions where m arms are active, i.e.:

$$\boldsymbol{\mathcal{A}}(m) = \left\{ \boldsymbol{a} = (a^1, \dots, a^n) \in \{0, 1\}^n : \sum_{i=1}^{n} a^i = m \right\}.$$

The decision-maker uses a time-homogeneous and deterministic Markov policy $\boldsymbol{g} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{A}}(m)$ to choose its actions, i.e.,

$$\boldsymbol{A}_t = \boldsymbol{g}(\boldsymbol{X}_t).$$

The family of all such policies is denoted by $\boldsymbol{\mathcal{G}}$. The performance of any policy $\boldsymbol{g} \in \boldsymbol{\mathcal{G}}$ is quantified by the expected discounted cost given by

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

where $\beta \in (0, 1)$ is the discount factor and the expectation is taken with respect to the joint distribution induced on all system variables when $\boldsymbol{A}_t = \boldsymbol{g}(\boldsymbol{X}_t)$.

We are interested in the following problem.

Problem 1 *Given the discount factor β, the total number n of arms, the number m of active arms, the state space $\boldsymbol{\mathcal{X}}$, the transition matrices $\{P^i\}_{i \in \mathcal{N}}$, the reset pmfs $\{Q^i\}_{i \in \mathcal{N}}$, and the cost functions $\{c^i(\cdot, \cdot)\}_{i \in \mathcal{N}}$, choose a policy $\boldsymbol{g} \colon \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{A}}(m)$ that minimizes $J(\boldsymbol{g})$ given by* (1).

A. Specific instances of the model

There are several models that have been investigated in the literature that may be viewed as a restless bandits with controlled restarts.

1) Machine maintenance models where a repairman is responsible for maintaining several machines. Each machine has a state that stochastically deteriorates over time. The repairman sees the state of all machines and may repair a subset of those. There is a state-dependent cost associated with running and repairing the machine. Such models are considered in [4], [5].
2) Machine maintenance models as before but where the state of the machine is not observed. Such models have been considered in the context of sensor networks [2] and smart grids [3].
3) Scheduling multiple data queues over a shared communication channels, where there is a cost associated with holding packets in a queue and a cost associated with transmitting [1].

III. Indexability

A. Restless Bandits with activation cost

Problem 1 is a Markov decision process and can be solved using dynamic programming [15]. However, the dynamic programming solution suffers from the curse of dimensionality because sizes of the state space $\boldsymbol{\mathcal{X}}$ and action space $\boldsymbol{\mathcal{A}}(m)$ are exponential in the number of arms n.

In the special case, when only one arm can be activated at a time, (i.e., $m = 1$) and passive arms remain frozen (i.e., P^i is identity for all arms) Gittins [8] showed that the above n-dimensional problem can be solved by solving n one-dimensional problem. Whittle [7] showed that Gittins index solution is a good heuristic for the general restless case (i.e., P^i is not identity for all arms) when a technical condition known as *indexability* is satisfied.

Indexability is the property of individual arms. Given any $i \in \mathcal{N}$, consider arm i with the same dynamics as before but per-step cost given by

$$c_\lambda(x_t^i, a_t^i) \coloneqq c^i(x^i, a_t^i) + \frac{\lambda}{1 - \beta} a_t^i,$$

where $\lambda \in \mathbb{R}$ is a penalty[2] for activating the arm. Then, consider the following auxiliary optimization problem.

Problem 2 *Given an arm $i \in \mathcal{N}$, discount factor β, the state space $\mathcal{X}^i$, the transition probability matrix P^i, the reset probability mass function Q^i, the cost function $c^i(\cdot, \cdot)$ and the penalty $\lambda \in \mathbb{R}$, choose a policy $g^i : \mathcal{X}^i \to \{0, 1\}$ to minimize*

$$J^i(g^i) \coloneqq (1 - \beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t c_\lambda^i(X_t^i, A_t^i)\right]. \tag{2}$$

Problem 2 is also a Markov decision process and the optimal solution is given by the solution to the following dynamic

[2]In the standard restless bandit problem, one considers maximizing the discounted reward and modifies the per-step reward to include a subsidy for passive actions. In contrast, we consider minimizing the discounted cost, so we modify the per-step cost to include a penalty for active action.

program. Let $V_\lambda^i : \mathcal{X}^i \to \mathbb{R}$ be the unique fixed point of the following:

$$V_\lambda^i(x) = \min\{H_\lambda^i(x,0), H_\lambda^i(x,1)\},\ \forall x \in \mathcal{X}^i. \tag{3}$$

where

$$H_\lambda^i(x,0) = (1-\beta)c^i(x,0) + \beta \sum_{y\in\mathcal{X}^i} P_{xy}^i V_\lambda^i(y), \tag{4}$$

$$H_\lambda^i(x,1) = (1-\beta)c^i(x,1) + \lambda + \beta \sum_{y\in\mathcal{X}^i} Q_y^i V_\lambda^i(y). \tag{5}$$

Let $g_\lambda^i(x)$ denote the minimizer of the right hand side of (3) where we set $g_\lambda^i(x) = 1$ if $H_\lambda^i(x,0) = H_\lambda^i(x,1)$. Then, from Markov decision theory [15], we know that the time-homogeneous policy g_λ^i is optimal for (2).

Let

$$\Pi_\lambda^i \coloneqq \{x^i \in \mathcal{X}^i : g_\lambda^i(x) = 0\} \tag{6}$$

denote the set of states where taking the passive action is optimal when the activation penalty is λ. This set is called the *passive set*. Arm i is said to be *indexable* if Π_λ^i is weakly increasing in λ, i.e., for any $\lambda_1, \lambda_2 \in \mathbb{R}$,

$$\lambda_1 < \lambda_2 \implies \Pi_{\lambda_1}^i \subseteq \Pi_{\lambda_2}^i.$$

When arm i is indexable, the Whittle index $w^i(x^i)$ at state x^i is defined as the smallest value of λ^i for which x^i belongs to the passive set Π_λ^i, i.e.,

$$w^i(x^i) \coloneqq \inf\{\lambda \in \mathbb{R} : x^i \in \Pi_\lambda^i\}. \tag{7}$$

Equivalently, the Whittle index $w^i(x^i)$ at state x^i is the smallest value of λ^i for which the optimal policy is indifferent between the active and the passive actions at state x^i.

A restless bandit problem is said to be indexable if all arms are indexable. For indexable problems, the whittle index heuristic is as follows: *at each time, compute the Whittle index of all arms and play the arms with the m smallest Whittle indices*.

As mentioned earlier, Whittle index policy is a popular approach for restless bandits because: (i) its complexity is linear in the number of alternatives and (ii) it often performs close to optimal in practice [5], [11], [12].

B. Indexability

In this section, we show that Problem 2 is Whittle indexable and derive an expression for the Whittle index.

Given an arm $i \in \mathcal{N}$, let Σ^i denote the family of all stopping times with respect to the natural filtration associated with $\{X_t^i\}_{t\geq 0}$. For any stopping time $\tau \in \Sigma^i$ and an initial state $x \in \mathcal{X}^i$, define

$$L^i(x,\tau) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t c(X_t^i, 0) + \beta^\tau c(X_\tau^i, 1) \;\middle|\; X_0^i = x\right],$$

$$B^i(x,\tau) \coloneqq \mathbb{E}[\beta^\tau | X_0^i = x].$$

Theorem 1 *Problem 1 is Whittle indexable and for any arm i, $i \in \mathcal{N}$, the Whittle index is given by*

$$w^i(x) = \inf\{\lambda \in \mathbb{R} : G^i(x) < W_\lambda^i\}$$

where

$$G^i(x) \coloneqq (1-\beta) \inf_{\tau\in\Sigma^i} \frac{L^i(x,\tau) - c^i(x,1)}{1 - B^i(x,\tau)}, \tag{8}$$

$$W_\lambda^i \coloneqq \lambda + \beta \sum_{y\in\mathcal{X}^i} Q_y V_\lambda^i(y). \tag{9}$$

□

To show that Problem 1 is indexable, we show that each arm is indexable. For that matter, we consider Problem 2 for each arm. For ease of notation, we drop the superscript i from all variables.

Lemma 1 *The following statements hold:*

1) *$V_\lambda(x)$ is strictly increasing in λ for any $x \in \mathcal{X}$.*
2) *W_λ is strictly increasing in λ.* □

PROOF These properties follow from the fact that $c_\lambda(x,a)$ is strictly increasing in λ. ■

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

$$\begin{aligned} C_\lambda(x,\tau) &= (1-\beta)\mathbb{E}^{h_\tau}\left[\sum_{t=0}^{\infty} \beta^t c_\lambda(X_t, A_t) \;\middle|\; X_0 = x\right] \\ &= (1-\beta)L(x,\tau) + \mathbb{E}[\beta^\tau W_\lambda | X_0 = x] \\ &= (1-\beta)L(x,\tau) + B(x,\tau)W_\lambda. \end{aligned} \tag{10}$$

Setting $\tau = 0$, we have

$$C_\lambda(x,0) = (1-\beta)c(x,1) + W_\lambda. \tag{11}$$

Lemma 2 *The following characterizations of the passive sets are equivalent to* (6)*.*

1) $\{x \in \mathcal{X} : H_\lambda(x,0) < H_\lambda(x,1)\}$.
2) $\{x \in \mathcal{X} : \exists \sigma \in \Sigma$ *such that* $C_\lambda(x,\sigma) < C_\lambda(x,0)\}$.
3) $\{x \in \mathcal{X} : G(x) < W_\lambda\}$. □

PROOF Characterization 1) follows from the dynamic program (3). Characterization 2) follows from the fact that $C_\lambda(x,0) = H_\lambda(x,1)$ and for $x \in \Pi_\lambda$, $C_\lambda(x,\sigma) = H_\lambda(x,0)$, where σ is the hitting time of the set $\mathcal{X}\backslash\Pi_\lambda$. Characterization 3) follows from characterization 2) and rearranging the terms using (10) and (11). ■

Now consider the characterization 3) in Lemma 2. $G(x)$ does not depend on λ while Lemma 1 shows that W_λ is strictly increasing in λ. Hence, Π_λ is increasing in λ. Thus arm i is indexable. The expression for the Whittle index in the Theorem 1 follows immediately from (7).

IV. Computation of Whittle index for threshold-based policies

In this section, we provide a closed form expression for the Whittle index when the state space is an ordered set and the model satisfies the following property.

(P) The optimal policy for Problem 2 is a threshold-based policy, i.e., for each $i \in N$, there exists a threshold $k^i \in \mathcal{X}^i$ such that

$$g^i_\lambda(x) \coloneqq \begin{cases} 0, & \text{if } x < k^i \\ 1, & \text{otherwise.} \end{cases}$$

In the sequel, we omit superscript i for ease of notation. We assume that the state space is given by $\mathcal{X} = \{1, \dots, \Omega\}$ and for any $k \in \mathcal{X}$, use the notation

$$\mathcal{X}_{<k} = \{x \in \mathcal{X} : x < k\} \quad \text{and} \quad \mathcal{X}_{\geq k} = \{x \in \mathcal{X} : x \geq k\}.$$

A. Sufficient conditions for optimality of threshold-based policies

We start by characterizing the sufficient conditions under which the optimal policy in Problem 2 is a threshold policy.

Proposition 1 *Consider the following conditions.*

(C1) *P is stochastically monotone, i.e., for any $x, y \in \mathcal{X}$ such that $x < y$, we have*

$$\sum_{w \in \mathcal{X}_{\geq z}} P_{x,w} \leq \sum_{w \in \mathcal{X}_{\geq z}} P_{y,w}, \ \forall z \in \mathcal{X}.$$

(C2) *For any $a \in \{0, 1\}$, $c(x, a)$ is (weakly) increasing in x.*

(C3) *$c(x, a)$ is submodular.*

Under (C1)–(C3), there exists a threshold $k \in \mathcal{X} \cup \{\Omega + 1\}$ such that the optimal policy in Problem 2 is of the form

$$g(x) = \begin{cases} 0, & \textit{if } x \in \mathcal{X}_{<k} \\ 1, & \textit{otherwise.} \end{cases}$$ □

PROOF The Conditions (C1)–(C3) are the same as Properties (P1)–(P4) of [15, Theorem 4.7.4]. We can show that the model satisfies Property (P4) of [15, Theorem 4.7.]. See [16] for a complete proof. ■

B. Performance evaluation of threshold-based policies

Let $g^{(k)}$ be the threshold policy with threshold k, i.e.,

$$g^{(k)}(x) = \begin{cases} 0, & \text{if } x \in \mathcal{X}_{<k} \\ 1, & \text{otherwise.} \end{cases}$$

Let $C^{(k)}_\lambda$ be the total discounted cost incurred under policy $g^{(k)}$ and penalty λ where the initial state is distributed according to Q, i.e.,

$$\begin{aligned} C^{(k)}_\lambda &\coloneqq (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t c_\lambda(X_t, g^{(k)}(X_t)) \;\middle|\; X_0 \sim Q\right] \\ &= (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \big(c(X_t, g^{(k)}(X_t)) + \lambda g^{(k)}(X_t)\big) \;\middle|\; X_0 \sim Q\right] \\ &= D^{(k)} + \lambda N^{(k)}, \end{aligned} \tag{12}$$

where

$$\begin{aligned} D^{(k)} &\coloneqq (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t c(X_t, g^{(k)}(X_t)) \;\middle|\; X_0 \sim Q\right], \\ N^{(k)} &\coloneqq (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t g^{(k)}(X_t) \;\middle|\; X_0 \sim Q\right]. \end{aligned}$$

To compute the performance $C^{(k)}_\lambda$, we need to obtain $D^{(k)}$ and $N^{(k)}$ which can be computed as follows. Let τ_k denote the hitting time of the set $\mathcal{X}_{\geq k}$. Define $L^{(k)}$ and $M^{(k)}$ as the expected discounted cost and the expected discounted time until we hit $\mathcal{X}_{\geq k}$ starting from an initial state distributed according to Q, i.e.,

$$\begin{aligned} L^{(k)} &\coloneqq \mathbb{E}\left[\sum_{t=0}^{\tau_k} \beta^t c(X_t, g^{(k)}(X_t)) \;\middle|\; X_0 \sim Q\right] \\ M^{(k)} &\coloneqq \mathbb{E}\left[\sum_{t=0}^{\tau_k} \beta^t \;\middle|\; X_0 \sim Q\right] = \frac{1 - \mathbb{E}[\beta^{\tau_k+1} | X_0 \sim Q]}{1-\beta}. \end{aligned}$$

Theorem 2 *For all $k \in \mathcal{X} \cup \{\Omega + 1\}$,*

$$D^{(k)} = \frac{L^{(k)}}{M^{(k)}} \quad \textit{and} \quad N^{(k)} = \frac{1}{\beta M^{(k)}} - \frac{1-\beta}{\beta}.$$ □

PROOF The proof follows from standard ideas in renewal theory and it is omitted due to lack of space. See [16] for a complete proof. ■

Thus, computing $L^{(k)}$ and $M^{(k)}$ is sufficient for calculating $D^{(k)}$ and $N^{(k)}$ and, consequently, $C^{(k)}_\lambda$. In turn, $L^{(k)}$ and $M^{(k)}$ can be computed using standard formulas for truncated Markov chains. For that matter, let c_a denote the column vector of costs $c(\cdot, a)$, $a \in \{0, 1\}$ and Q the row vector of the restart pmf. For vector $\alpha \in \{c_0, c_1, Q\}$, let $\alpha^{(k)}$ denote the first $(k-1)$ elements and $\tilde{\alpha}^{(k)}$ the remaining $(\Omega - k)$ elements of α. Let P_k be a $(k-1) \times (k-1)$ dimensional upper-left submatrix of P and $\tilde{P}^{(k)}$ be the $(k-1) \times (\Omega - k + 1)$-dimensional upper-right submatrix of P.

Proposition 2 *For all $k \in \mathcal{X}$,*

$$\begin{aligned} L^{(k)} &= Q^{(k)} Z^{(k)} (c_0^{(k)} + \beta \tilde{P}^{(k)} \mathbf{1}_{\Theta - k}) + \tilde{Q}^{(k)} \tilde{c}_1^{(k)}, \\ M^{(k)} &= Q^{(k)} Z^{(k)} (\mathbf{1}_{k-1} + \beta \tilde{P}^{(k)} \tilde{c}_1^{(k)}) + \tilde{Q}^{(k)} \mathbf{1}_{\Omega - k + 1}, \end{aligned}$$

where $Z^{(k)} = (I_{k-1} - \beta P^{(k)})^{-1}$. □

PROOF The proof follows from the balance equations of the truncated Markov chains and is omitted due to lack of space. ■

C. Computation of the index

Next, we derive structural properties of $C^{(k)}_\lambda$.

Lemma 3 *The following statements hold:*

1) *$M^{(k)}$ is strictly increasing in k.*
2) *$C^{(k)}_\lambda$ is sub-modular in (k, λ).*
3) *Let $k_\lambda \coloneqq \arg\min_{k \in \mathcal{X}} C^{(k)}_\lambda$, i.e., the optimal threshold corresponding to penalty λ. Then, k_λ is increasing in λ.*
4) *$C^*_\lambda = C^{(k_\lambda)}_\lambda = \min_{k \in \mathcal{X}} C^{(k)}_\lambda$ is continuous in λ.* □

PROOF The monotonicity of $M^{(k)}$ follows from definition and together with Theorem 2 implies that $N^{(k)}$ is strictly decreasing in k. This, together with (12), implies that $C^{(k)}_\lambda$ is submodular in (k, λ). By the property of submodular functions mentioned in footnote 1, k_λ is increasing in λ. The continuity of C^*_λ follows from the fact that $C^{(k)}_\lambda$ is continuous in λ for each k. ■

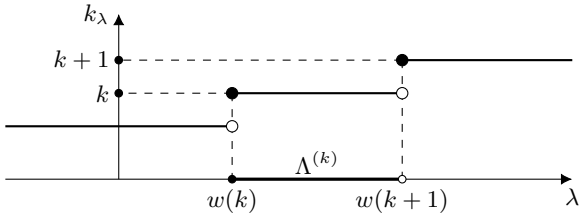


Fig. 1: k_λ as a function of λ.

Since k_λ is increasing and takes integer values, it is staircase function in λ, as illustrated in Fig. 1. This property allows us to compute the Whittle index for Problem 2.

Theorem 3 *For Problem 2, under Property **(P)**, the Whittle index at state $k \in \mathcal{X}$ is*

$$w(k) = \frac{D^{(k+1)} - D^{(k)}}{N^{(k)} - N^{(k+1)}}. \tag{13}$$

□

PROOF Recall that the Whittle index is the smallest value of λ for which the optimal policy is indifferent between the active and the passive actions. Let $\Lambda^{(k)} = \{\lambda \in \mathbb{R} : k_\lambda = k\}$. See Fig 1 for an illustration. By definition, for any $\lambda \in \Lambda^{(k)}$ we have $C^*_\lambda = C^{(k)}_\lambda$, and for any $\lambda \in \Lambda^{k+1}$ we have $C^*_\lambda = C^{(k+1)}_\lambda$. From Lemma 3, part 4), C^*_λ is continuous in λ. Therefore,

$$C^{(k)}_{w(k)} = \lim_{\lambda \uparrow w(k)} C^*_\lambda = \lim_{\lambda \downarrow w(k)} C^*_\lambda = C^{(k+1)}_{w(k)}.$$

Thus, $C^{(k)}_{w(k)} = C^{(k+1)}_{w(k)}$ and

$$D^{(k)} + w(k)N^{(k)} = D^{(k+1)} + w(k)N^{(k+1)}$$

which implies (13). ■

Remark 1 Theorem 2 and Lemma 3, part 1), imply that $N^{(k)}$ is strictly decreasing in k. Hence, $N^{(k)} \neq N^{(k+1)}$) and the expression for Whittle index given in (13) is well-defined. □

V. NUMERICAL EXPERIMENTS

In this section, we perform numerical experiments on models which satisfy (C1)–(C3) of Proposition 1 and evaluate how well the Whittle index policy (WIP) performs compared to the optimal policy (OPT) as well as to a baseline policy known as the myopic policy (MYP) which is shown in Alg 1.

Algorithm 1 Myopic Heuristic

1: $t = 1$.
2: **while** $t \geq 1$ **do**
3: Set $k = 1$. $M = \emptyset$. $K = \mathcal{N}$.
4: Let $i^*_k = \arg\min_{i \in K} \sum_{j \in K \setminus \{i\}} c^j(X^j_t, 0) + c^i(X^i_t, 1)$.
5: Set $M = M \cup \{i^*_k\}$, $K = K \setminus \{i^*_k\}$.
6: If $k = m$ activate arms in M and stop. Else set $k = k+1$ and go to Line 4.
7: $t = t + 1$.
8: **end while**

TABLE I: Relative performance of WIP vs. OPT for Experiment 1.

(a) $m = 1$

ℓ	α_{OPT}
1	99.967
2	99.902
3	99.917
4	99.649

(b) $m = 2$

ℓ	α_{OPT}
1	100.00
2	99.997
3	99.999
4	99.972

A. Experimental Setup

The model has 3 components: the transition matrix P, the reset pmf Q and the cost function c. We choose these components as follows:

1) The choice of transition matrix: We have two setups for choosing the transition matrix. The first setup is a family of 4 types of structured stochastic monotone matrices, which we denote by $\mathcal{P}_\ell(p)$, $\ell \in \{1, \dots, 4\}$, where $p \in [0, 1]$ is a parameter of the model. The second setup is a randomly generated stochastic monotone matrices which we denote by $\mathcal{R}(d)$, where $d \in [0, 1]$ is a parameter of the model. The details of these models are presented in Appendix A.

2) The choice of reset pmf: In all our experiments, we use $Q = [1, 0, \dots, 0]$, i.e., choosing the restart action deterministically resets to the clear state.

3) The choice of the cost function: For all our experiments we choose $c(x, 0) = (x-1)^2$ and $c(x, 1) = 0.5(\Omega-1)^2$ where $\Omega = |\mathcal{X}|$.

B. Experimental details and result

We conduct different experiments to compare the performance of Whittle index with both the optimal policy and the myopic policy for different setups (described in Section V-A) and for different values of the size Ω of the state space, the number n of the arms, and the number m of active arms. For all experiments we choose the discount factor $\beta = 0.9$.

The performance of a policy is evaluated by Monte Carlo simulations over S trajectories is truncated at length T. In all our experiments, we choose $S = 5000$ and $T = 250$.

Experiment 1) *Comparison of Whittle index with the optimal policy for structured models:* The optimal policy is computed by solving the MDP for Problem 1. The state for this MDP is Ω^n. So, we can obtain the optimal policy only for small values of Ω and n. We choose $\Omega = 5$ and $n = 5$ and compare the two policies for model $\mathcal{P}_\ell(\cdot)$, $\ell \in \{1, \dots, 4\}$ and $m \in \{1, 2\}$.

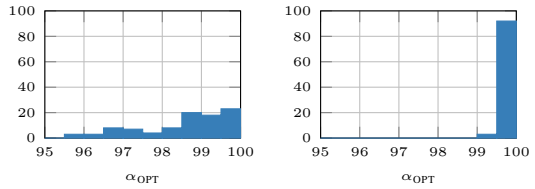
For a given value of n and ℓ, we generate the models for n arms as follows. Let $(p_1, \dots, p_n)$ denote n equispaced points in the interval $[0.35, 1]$. Then we choose $\mathcal{P}_\ell(p_i)$ as the transition matrix of arm i. Let

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

denote the relative performance (in percentage) of WIP compared to OPT.

The values of α_{OPT} for different values of ℓ and m are shown in Table I. The results for several simple models given in Table I show that WIP can be as good as OPT when $m = 2$ and slightly worse when $m = 1$.

Experiment 2) *Comparison of Whittle index with the optimal policy for randomly sampled models:* As before, we pick $\Omega = 5$ and $n = 5$ so that it is feasible to calculate the optimal policy. For each arm, we sample the transition matrix from $\mathcal{R}(5/\Omega)$. We repeat the experiment 100 times. The histogram of α_{OPT} over experiments for $m \in \{1, 2\}$ is plotted in Fig 4. Similar to the result of Experiment 1, WIP has a reasonable relative performance with respect to OPT.



(a) $m = 1$. **(b)** $m = 2$.

Fig. 4: Histogram of the relative performance α_{OPT} of WIP versus OPT for Experiment 2.

Experiment 3) *Comparison of Whittle index with the myopic policy for structured models:* We generate the structured models as in Experiment 1 but for $\Omega = 25$, $n \in \{25, 50, 75\}$, and $m \in \{1, 2, 5\}$. In this case, let

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

denote the relative improvement of WIP compared to MYP. The results of ε_{MYP} for different choice of the parameters are shown in Fig 2.

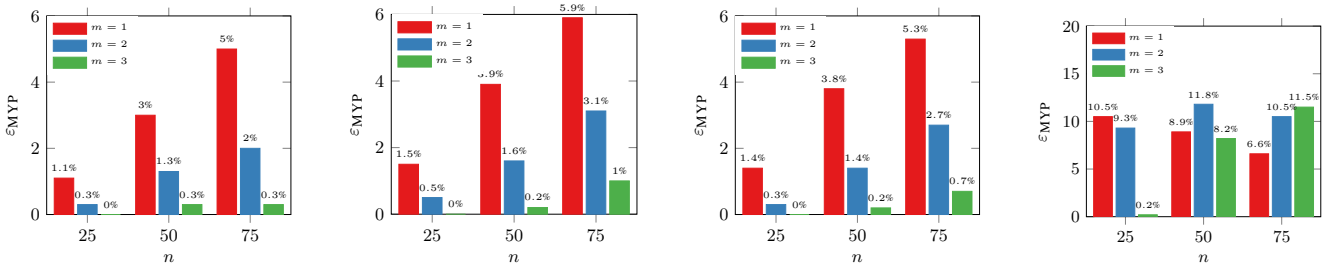
In Fig 2, we observe that WIP performs considerably better than MYP. In addition to that, performance of WIP is better with respect to MYP when $\ell = 4$ which is more complicated than models where $\ell \in \{1, 2, 3\}$. However, increasing m doesn't necessarily contribute to better ε_{MYP} as overlap between the choices of the two policies may increase. Note that as $\mathcal{P}_4(\cdot)$ is very different from the rest of the models, the trend of bars in Fig 2d with respect to n varies differently from the rest of the models.

Experiment 4) *Comparison of Whittle index with the myopic policy for randomly sampled models:* We generate 100 random models as described in Experiment 2 but for $\Omega = 25$ and larger values of n. For each case, ε_{MYP} is computed. The histogram of ε_{MYP} for different choices of the parameters are shown in Fig 3.

The result shows that on average, WIP performs considerably better than MYP and this improvement is guaranteed as the concentration of data for the sampled models is mostly on positive values of ε_{MYP}.

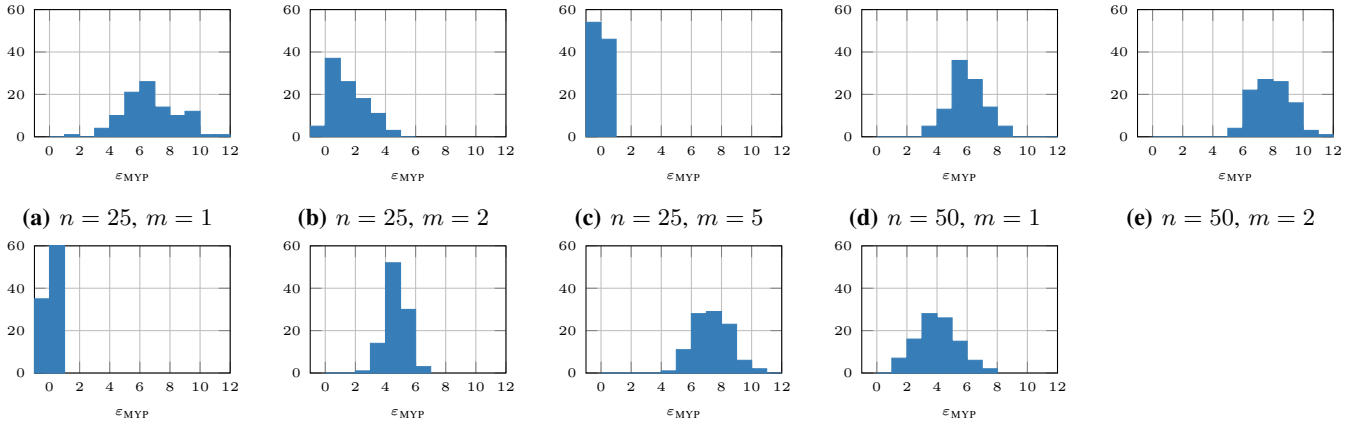
VI. Conclusion

In this paper, we present a model for restless bandit with controlled restarts. We show that the model is indexable. When the auxiliary problem to compute the Whittle index has a



(a) $\ell = 1$ **(b)** $\ell = 2$ **(c)** $\ell = 3$ **(d)** $\ell = 4$

Fig. 2: Relative improvement ε_{MYP} of WIP vs. MYP for $\Omega = 25$ when $\ell \in \{1, \ldots, 4\}$, $n \in \{25, 50, 75\}$, and $m \in \{1, 2, 5\}$.



(a) $n = 25$, $m = 1$ **(b)** $n = 25$, $m = 2$ **(c)** $n = 25$, $m = 5$ **(d)** $n = 50$, $m = 1$ **(e)** $n = 50$, $m = 2$

(f) $n = 50$, $m = 5$ **(g)** $n = 75$, $m = 1$ **(h)** $n = 75$, $m = 2$ **(i)** $n = 75$, $m = 5$

Fig. 3: Histogram of relative improvement ε_{MYP} of WIP vs. MYP for $\Omega = 25$ when $n \in \{25, 50, 75\}$, and $m \in \{1, 2, 5\}$.

threshold-based optimal strategy, we derive a closed form expression to compute the Whittle index. For the case when the Markov chain matrix under the passive action is stochastic monotone and the per-step cost is monotonically increasing and submodular, we present detailed numerical experiments which suggest that the Whittle index policy performs very close to the optimal policy and considerably better than other heuristics such as a myopic policy.

APPENDIX

A. Stochastic Monotone Matrix Generation

1) Structured models: Consider a Markov chain with n states. We consider four different class of stochastic monotone transition probability matrices, which we call $\mathcal{P}_\ell(p)$, $\ell = \{1, \ldots, 4\}$, where p is a model parameter.

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

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

$\mathcal{P}_1(p)$ is stochastic monotone if $p \in [1/3, 1]$.

Matrix $\mathcal{P}_2(p)$: Similar to $\mathcal{P}_1(p)$ with $q_1 = (1-p)/2$ and $q_2 = (1-p)/2$. $\mathcal{P}_2(p)$ is stochastic monotone if $p \in [1/4, 1]$.

Matrix $\mathcal{P}_3(p)$: Similar to $\mathcal{P}_1(p)$ with $q_1 = (1-p)/3$ and $q_2 = (1-p)/6$. $\mathcal{P}_3(p)$ is stochastic monotone if $p \in [1/5, 1]$.

Matrix $\mathcal{P}_4(p)$: Let $q = (1-p)/(n-1)$. Then,

$$\mathcal{P}_4(p) = \begin{bmatrix} p & q & q & \ldots & q & q \\ q & p & q & \ldots & q & q \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ q & q & q & \ldots & p & q \\ q & q & q & \ldots & q & p \end{bmatrix}$$

$\mathcal{P}_4(p)$ is stochastic monotone if $p \in [1/n, 1]$.

2) Randomly generated model: Consider a Markov chain with n states and the transition probability matrix

$$P = \begin{bmatrix} P_{11} & P_{12} & P_{13} & \ldots & P_{1n} \\ P_{21} & P_{22} & P_{23} & \ldots & P_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_{n1} & P_{n2} & P_{n3} & \ldots & P_{nn} \end{bmatrix}.$$

Let $F_{ij} = \sum_{y=j}^{n} P_{iy}$. The necessary condition for P to be stochastic monotone is that for any $1 \le i \le l \le n$ and any $1 \le j \le n$, $F_{ij} \le F_{lj}$.

Initially, we generate P_{11} uniformly random between $[1-d, 1]$ where $d \in [0, 1]$. Variable d prevents the kernel to behave badly when the number of states increases. Then, we generate $P_{12}, P_{13}, \ldots, P_{1n}$ sequentially from P_{12} to P_{1n} where each mass is selected uniformly random from $[0, B_i]$ where $B_i = 1 - \sum_{l=1}^{i-1} P_{1l}$. As $F_{in} = P_{in}$ for any i, we select P_{in} sequentially for rows from 2 to n where each element is generated uniformly random from $[P_{(i-1)n}, \min\{1, P_{(i-1)n} + d\}]$. Then, for any row from 2 to n, we repeat the following procedure backwardly for columns from $n-1$ to 1. Consider row i and column j. We generate a uniformly random number from $[\text{LB}_{ij}, \text{UB}_{ij}]$ where $\text{LB}_{ij} = F_{(i-1)j} - F_{i(j+1)}$ and $\text{UB}_{ij} = \min\{\text{LB} + d, 1 - F_{i(j+1)}\}$ and set the generated number as P_{ij}. The lower bound is due to stochastic monotonicity property and the upper bound is due to definition of a probability mass function and variable d. Note that for the elements in the first column, the mentioned interval shrinks to $[1 - F_{i2}, 1 - F_{i2}]$ for row i which results in $P_{i1} = 1 - F_{i2}$.

REFERENCES

[1] Vivek S. Borkar, Gaurav S. Kasbekar, Sarath Pattathil, and Priyesh Shetty, "Opportunistic scheduling as restless bandits," *IEEE Transactions on Control of Network Systems*, 2017.

[2] Sofía 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.

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

[4] 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*, vol. 165, no. 1, pp. 267–284, 2005.

[5] 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.

[6] Sarang Deo, Seyed Iravani, Tingting Jiang, Karen Smilowitz, and Stephen Samuelson, "Improving health outcomes through better capacity allocation in a community-based chronic care model," *Operations Research*, vol. 61, no. 6, pp. 1277–1294, 2013.

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

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

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

[10] Aditya Mahajan and Demosthenis Teneketzis, "Multi-armed bandits," in *Foundations and Applications of Sensor Management*, pp. 121–151. Springer-Verlag, 2008.

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

[12] KD Glazebrook and HM Mitchell, "An index policy for a stochastic scheduling model with improving/deteriorating jobs," *Naval Research Logistics (NRL)*, vol. 49, no. 7, pp. 706–721, 2002.

[13] KD Glazebrook, DJ Hodge, and Christopher Kirkbride, "Monotone policies and indexability for bidirectional restless bandits," *Advances in Applied Probability*, vol. 45, no. 1, pp. 51–85, 2013.

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

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

[16] Nima Akbarzadeh and Aditya Mahajan, "Restless bandits with controlled restarts: Indexability and computation of whittle index," (extended version). http://www.cim.mcgill.ca/~adityam/projects/bandits/conference/2019-cdc-extended.pdf.

[17] Donald M Topkis, *Supermodularity and complementarity*, Princeton university press, 2011.

A. Proof of Lemma 1

(a) The per step cost of $c_\lambda(x,a)$ is increasing in λ for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$. Hence, according to (3), $V_\lambda(x)$ is strictly increasing in λ for any $x \in \mathcal{X}$.

(b) The result follows from the definition of W_λ in (9) and Part (a).

B. Proof of Lemma 2

We prove the equivalence as follows.

Eq. (6) $\iff$ **Relationship** (1)**:**

$$x \in \Pi_\lambda \overset{(a)}{\iff} g(x) = 0 \overset{(b)}{\iff} H_\lambda(x,0) < H_\lambda(x,1)$$

where (a) follows from (6) and (b) follows from the dynamic program (3).

Relationship (1) $\iff$ **Relationship** (2)**:** For necessity of the relation, let σ be the hitting time of the set $\mathcal{X} \setminus \Pi_\lambda$. By definition, for any $x \in \Pi_\lambda$, the policy h_σ coincides with the optimal policy. Therefore, $\forall x \in \Pi_\lambda : C_\lambda(x,\sigma) = H_\lambda(x,0)$. By Relationship (1), we have

$$\forall x \in \Pi_\lambda : C_\lambda(x,\sigma) = H_\lambda(x,0) < H_\lambda(x,1) = C_\lambda(x,0).$$

For sufficiency part, suppose there exists a stopping time σ such that $C_\lambda(x,\sigma) < C_\lambda(x,0)$. Therefore, we have

$$\begin{aligned} &V_\lambda(x) \le C_\lambda(x,\sigma) < C_\lambda(x,0) = H(x,1) \Rightarrow \\ &V_\lambda(x) = H(x,0) < H(x,1). \end{aligned}$$

Relationship (2) $\iff$ **Relationship** (3)**:** Combining (10), (11), and Relationship (2) we get that $x \in \Pi_\lambda$ if and only if there exists a positive stopping time $\sigma \in \Sigma$ such that

$$(1-\beta)L(x,\sigma) + B(x,\sigma)W_\lambda < (1-\beta)c(x,1) + W_\lambda$$

which is equivalent to

$$(1-\beta)\frac{L(x,\sigma) - c(x,1)}{1 - B(x,\sigma)} < W_\lambda.$$

One can assert $x \in \Pi_\lambda$ if and only if

$$\inf_{\sigma \in \Sigma} (1-\beta)\frac{L(x,\sigma) - c(x,1)}{1 - B(x,\sigma)} < W_\lambda.$$

Therefore, $\Pi_\lambda = \{x \in \mathcal{X} : G(x) < W_\lambda\}$.

C. Proof of Theorem 3

By strong Markov property, we have

$$\begin{aligned} D^{(k)} = \mathbb{E}\Bigg[&(1-\beta)\sum_{t=0}^{\tau_k} \beta^t c(X_t, g^{(k)}(X_t)) \\ &+ \beta^{\tau_k+1} D^{(k)} \;\Big|\; X_0 \sim Q\Bigg] \\ &= (1-\beta)L^{(k)} + \mathbb{E}[\beta^{\tau_k+1} | X_0 \sim Q] D^{(k)}. \end{aligned}$$

Using $M^{(k)}$ definition, we have

$$\mathbb{E}[\beta^{\tau_k+1} | X_0 \sim Q] = 1 - (1-\beta)M^{(k)}.$$

Substituting this in (C) and rearranging the terms we get

$$D^{(k)} = \frac{L^{(k)}}{M^{(k)}}.$$

For $N^{(k)}$, by strong Markov property we have

$$\begin{aligned} N^{(k)} &= \mathbb{E}\Big[(1-\beta)\beta^{\tau_k} + \beta^{\tau_k+1} N^{(k)}) \;\Big|\; X_0 \sim Q\Big] \\ &= \mathbb{E}[\beta^{\tau_k} | X_0 \sim Q](1 - \beta + \beta N^{(k)}) \\ &= \frac{1-(1-\beta)M^{(k)}}{\beta}(1 - \beta + \beta N^{(k)}). \end{aligned}$$

Therefore,

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

D. Proof of Proposition 1

We have

$$\begin{aligned} L^{(k)} &= \mathbb{E}\Big[\sum_{t=0}^{\tau_k - 1} \beta^t c(X_t, 0) + \beta^{\tau_k} c(X_{\tau_k}, 1) \;\Big|\; X_0 \sim Q\Big] \\ &= Q^{(k)} L_{k1} + \tilde{Q}^{(k)} L_{k2} \end{aligned}$$

where

$$\begin{aligned} L_{k1} &= \mathbb{E}\Big[\sum_{t=0}^{\tau_k - 1} \beta^t c(X_t, 0) + \beta^{\tau_k} c(X_{\tau_k}, 1) \;\Big|\; X_0 \in \mathcal{X}_{<k}\Big] \\ L_{k2} &= \mathbb{E}[c(X_t, 1) | X_0 \in \mathcal{X}_{\ge k}]. \end{aligned}$$

Note that $d_k(x)$ represents the expected cost incurred when the state escapes the stopping set Π_k starting from state $x < k$. Therefore, to compute L_{k1}, we use the following procedure.

$$\begin{aligned} L_{k1} &= \mathbb{E}\Big[\sum_{t=0}^{\tau_k - 1} \beta^t c(X_t, 0) + \beta^{\tau_k} c(X_{\tau_k}, 1) \;\Big|\; X_0 \in \mathcal{X}_{<k}\Big] \\ &= \sum_{\tau=0}^{\infty} (\beta P^k)^\tau (c^{(k)} + \beta d^{(k)}) = Z^{(k)}(c^{(k)} + \beta d^{(k)}). \end{aligned}$$

It is also straightforward that $L_{k2} = \tilde{c}^{(k)}$. Hence, the statement given for $L^{(k)}$ holds.

Similar to $L^{(k)}$, for $M^{(k)}$ we have

$$M^{(k)} = \mathbb{E}\Big[\sum_{t=0}^{\tau_k} \beta^t \;\Big|\; X_0 \sim Q\Big] = Q^{(k)} M_{k1} + \tilde{Q}^{(k)} M_{k2}$$

where

$$\begin{aligned} M_{k1} &= \mathbb{E}\Big[\sum_{t=0}^{\tau_k - 1} \beta^t + \beta^{\tau_k} \;\Big|\; X_0 \in \mathcal{X}_{<k}\Big] \\ M_{k2} &= \mathbb{E}[1 | X_0 \in \mathcal{X}_{\ge k}]. \end{aligned}$$

Following the same analysis carried out in the previous part we have

$$\begin{aligned} M_{k1} &= \mathbb{E}\Big[\sum_{t=0}^{\tau_k - 1} \beta^t + \beta^{\tau_k} \;\Big|\; X_0 \in \mathcal{X}_{<k}\Big] \\ &= \sum_{\tau=0}^{\infty} (\beta P^k)^\tau (\mathbf{1}_{k-1} + \beta e^{(k)}) = Z^{(k)}(\mathbf{1}_{k-1} + \beta e^{(k)}). \end{aligned}$$

It is straightforward that $M_{k2} = \mathbf{1}_{\Omega - k + 1}$. Hence, the statement given for $M^{(k)}$ holds.

E. Proof of Lemma 3

(a) A sample path which begins from state x must escape the set $\mathcal{X}_{<k}$ before escaping set $\mathcal{X}_{<l}$. Thus $\tau_k < \tau_l$ everywhere and therefore $M^{(k)} < M^{(l)}$.

(b) To prove the first property, assume $k < l$ and we have

$$C_\lambda^{(l)} - C_\lambda^{(k)} = D^{(l)} - D^{(k)} - \lambda(N^{(k)} - N^{(l)}).$$

From Part (a), $M^{(k)}$ is strictly increasing in k. Therefore, due to Theorem 2, $N^{(k)}$ is strictly decreasing in k and hence the difference $C_\lambda^{(l)} - C_\lambda^{(k)}$ is strictly decreasing λ. Therefore, $C_\lambda^{(k)}$ is submodular in (k, λ).

(c) The result follows from Part (b) and [17, Theorem 2.8.2].

(d) Note that $C_\lambda^{(k)}$ is continuous in λ. Therefore, C_λ^* is a minimum of Ω functions that are continuous in λ and is, therefore, continuous in λ.