Les Cahiers du GERAD ISSN: 0711–2440

Restless bandits: Indexability and computation of Whittle index

N. Akbarzadeh,
A. Mahajan

G–2020–34

June 2020

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2020
– Bibliothèque et Archives Canada, 2020

Legal deposit – Bibliothèque et Archives nationales du Québec, 2020
– Library and Archives Canada, 2020

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Restless bandits: Indexability and computation of Whittle index

Nima Akbarzadeh [a]
Aditya Mahajan [b,c]

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

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

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

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

June 2020
Les Cahiers du GERAD
G–2020–34

Abstract: Restless bandits are a class of sequential resource allocation problems concerned with allocating one or more resources among several alternative processes where the evolution of the process depends on the resource allocated to them. Such models capture the fundamental trade-offs between exploration and exploitation. In 1988, Whittle developed an index heuristic for restless bandit problems which has emerged as a popular solution approach due to its simplicity and strong empirical performance. The Whittle index heuristic is applicable if the model satisfies a technical condition known as indexability. In this paper, we present two general sufficient conditions for indexability and identify simpler to verify refinements of these conditions. We then present a general algorithm to compute Whittle index for indexable restless bandits. Finally, we present a detailed numerical study which affirms the strong performance of the Whittle index heuristic.

Keywords: Multi-armed bandits, restless bandits, Whittle index, indexability, stochastic scheduling, resource allocation

Acknowledgments: This research was funded in part by the Innovation for Defence Excellence and Security (IDEaS) Program of the Canadian Department of National Defence through grant CFPMN2–037.

1 Introduction

Restless bandits are a class of sequential resource allocation problems concerned with allocating one or more resources among several alternative processes where the evolution of the process depends on the resource allocated to them. Such models arise in various applications such as machine maintenance [18], congestion control [7], healthcare [11, 9], finance [15], channel scheduling [20], smart grid [1], and others.

Restless bandits are a generalization of classical multi-armed bandits [12], where the processes remain frozen when resources are not allocated to them. Gittins [13] showed that when a single resource is allocated among multiple resources, the optimal policy has a simple structure: compute an index for each process and allocate the resource to the process with the largest (or the lowest) index. In contrast, the general restless bandit problem is PSPACE-hard [24]. Whittle [29] showed that index-based policies are optimal for the Lagrangian relaxation of the restless bandit problem and argued that the corresponding index, now called Whittle index, is a reasonable heuristic for restless bandit problems. Subsequently, it has been found that the Whittle index heuristic is optimal under some conditions [28] and performs well in practice [5, 14, 16].

The Whittle index heuristic is applicable if a technical condition known as *indexability* is satisfied. The condition appears to be a natural condition which should be satisfied by all models, but that is not the case [29]. Sufficient conditions for indexability have been investigated under specific modeling assumptions (two state fully or partially observed restless bandits [20, 7]; monotone bandits [15, 5, 7]; models with *right-skeip free* transitions [18, 14]; models with monotone or convex cost/reward [14, 5, 7, 6, 30, 8]; models satisfying partial conservation laws [21, 22]). Indexability for models arising in specific applications has been investigated in [16, 18, 14, 15, 6, 30, 8]. Our first main contribution is to provide general sufficient conditions for indexability, which are presented in Section 3. These sufficient conditions are based on an alternative characterization of passive set, which might be useful in general as well. We also present refinements of these sufficient conditions that are simpler to verify.

Whittle index can be computed by conducting a binary search over penalty for active action (or a subsidy for passive action) [4, 27] but such a binary search is computationally expensive because a dynamic program needs to be solved at each step. Methods to compute a generalization of Whittle index known as marginal productivity index are presented in [23, 22]. Our second main contribution is to present a general algorithm to compute Whittle index for indexable restless bandits, which is developed in Section 4. The key idea to compute the Whittle index is to iteratively sort the states in increasing order of their Whittle index.

We generalize the results for monotone bandits [15, 5, 7] to what we call stochastic monotone bandits (Section 5). We show that stochastic monotone bandits are indexable and the Whittle index can be computed in closed form. We also investigate a special case of our sufficient conditions in detail: restless bandits with controlled restarts (Section 5). Such models have been considered in [4, 27] and may be viewed as generalizations of the restart models [18, 14]. We use ideas from renewal theory to simplify the computation of the Whittle index for such models.

A detailed numerical study comparing the performance of the Whittle index policy with that of the optimal policy (for small models) and the myopic policy (for larger models) is presented in Section 6. In general, the performance of Whittle index policy is comparable to the optimal policy and considerably better than the myopic policy.

Notation Uppercase letters (X, Y, etc.) denote random variables, lowercase letters (x, y, etc.) denote their realization, and 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, $\mathbb{E}[\cdot]$ denotes the expectation of a random variable. $\mathbb{Z}$ and $\mathbb{R}$ denote the sets of integers and real numbers. Given a matrix P, P_{ij} denotes its (i, j)-th element.

For the totally ordered sets, $\mathcal{X}_{\geq k}$ denotes the set of states greater than or equal to state k and $\mathcal{X}_{<k}$ denotes the set of states lower than state k.

2 Restless bandits: problem formulation and solution concept

2.1 Restless Bandit Process

A discrete-time restless bandit process (RBP) is a controlled Markov process $(\mathcal{X}, \{0,1\}, \{P(a)\}_{a\in\{0,1\}}, c, x_0)$ where $\mathcal{X}$ denotes the state space which is a finite or countable set. $\{0,1\}$ denotes the action space. The action 0 is called the *passive* action and the action 1 is the *active* action. $P(a)$, $a \in \{0,1\}$, denotes the transition matrix when action a is chosen. $c : \mathcal{X} \times \{0,1\} \to \mathbb{R}$ denotes the cost function and x_0 denotes the initial state. We use X_t and A_t to denote the action of the process at time t. The process evolves in a controlled Markov manner, i.e., for any realization $x_{0:t+1}$ of $X_{0:t+1}$ and $a_{0:t+1}$ of $A_{0:t+1}$, we have $\mathbb{P}(X_{t+1} = x_{t+1} | X_{0:t} = x_{0:t}, A_{0:t} = a_{0:t}) = \mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t, A_t = a_t)$, which we denote by $P_{x_t x_{t+1}}(a_t)$.

2.2 Restless Multi-armed Bandit Problem

A restless multi-armed bandit problem is a collection of n independent RBPs $(\mathcal{X}^i, \{0,1\}, \{P^i(a)\}_{a\in\{0,1\}}, c^i)$, $i \in \mathcal{N} := \{1, \dots, n\}$. A decision maker observes the state of all RBPs, may choose the active action for only $m < n$ of them, and incurs a cost equal to the sum of the cost incurred by each bandit process.

Let $\boldsymbol{\mathcal{X}} := \prod_{i\in\mathcal{N}} \mathcal{X}^i$ and $\boldsymbol{\mathcal{A}}(m) := \left\{\boldsymbol{a} = (a^1, \dots, a^n) \in \mathcal{A}^n : \sum_{i\in\mathcal{N}} a^i = m\right\}$ to denote the joint state space and the feasible action space, respectively. Let $\boldsymbol{X}_t := (X_t^1, \dots X_t^n)$ and $\boldsymbol{A}_t = (A_t^1, \dots, A_t^n)$ denote the joint state and actions at time t. As the RBPs are independent, for any realization $\boldsymbol{x}_{0:t}$ of $\boldsymbol{X}_{0:t}$ and $\boldsymbol{a}_{0:t}$ of $\boldsymbol{A}_{0:t}$, we have $\mathbb{P}\left(\boldsymbol{X}_{t+1} = \boldsymbol{x}_{t+1} | \boldsymbol{X}_{0:t} = \boldsymbol{x}_{0:t}, \boldsymbol{A}_{0:t} = \boldsymbol{a}_{0:t}\right) = \prod_{i=1}^n \mathbb{P}\left(X_{t+1}^i = x_{t+1}^i | X_t^i = x_t^i, A_t^i = a_t^i\right)$. When the system is in state $\boldsymbol{x}_t = (x_t^1, \dots, x_t^n)$ and the decision-maker chooses action $\boldsymbol{a}_t = (a_t^1, \dots, a_t^n)$, the system incurs a cost $\bar{c}(\boldsymbol{x}_t, \boldsymbol{a}_t) := \sum_{i\in\mathcal{N}} c^i(x_t^i, a_t^i)$. The decision-maker chooses his actions using a time-homogeneous Markov policy $\boldsymbol{g} : X \to A(m)$, i.e., chooses $\boldsymbol{A}_t = \boldsymbol{g}(\boldsymbol{X}_t)$. The performance of any Markov policy $\boldsymbol{g}$ is given by

$$J^{(\boldsymbol{g})}(\boldsymbol{x}_0) := (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \bar{c}(\boldsymbol{X}_t, \boldsymbol{g}(\boldsymbol{X}_t)) \middle| \boldsymbol{X}_0 = \boldsymbol{x}_0\right],$$

where $\beta \in (0,1)$ is the discount factor, $\boldsymbol{x}_0$ is the initial system state and the expectation is taken with respect to the joint distribution of all system variables induced by the policy.

We are interested in the following optimization problem.

Problem 1 *Given the discount factor $\beta \in (0,1)$, the total number n of arms, the number m of active arms, RBPs $(\mathcal{X}^i, \{0,1\}, \{P^i(a)\}_{a\in\{0,1\}}, c^i, x_0^i)$, $i \in \mathcal{N}$ and initial state $\boldsymbol{x}_0 \in X$, choose a Markov policy $\boldsymbol{g} \colon X \to A(m)$ that minimizes $J^{(\boldsymbol{g})}(\boldsymbol{x}_0)$.*

Problem 1 is a multi-stage stochastic control problem and one can obtain an optimal solution using dynamic programming. However, the dynamic programming solution is intractable for large n since the cardinality of the state space is $\prod_{i\in\mathcal{N}} |\mathcal{X}^i|$, which grows exponentially with n. In the next section, we describe a heuristic known as *Whittle index* to efficiently obtain a suboptimal solution of the problem.

2.3 Indexability and the Whittle index

Consider a RBP $(\mathcal{X}, \{0,1\}, \{P(a)\}_{a\in\{0,1\}}, c, x_0)$. For any $\lambda \in \mathbb{R}$, we consider a Markov decision process $\{\mathcal{X}, \{0,1\}, \{P(a)\}_{a\in\{0,1\}}, c_\lambda, x_0\}$, where

$$c_\lambda(x, a) := c(x, a) + \lambda a, \quad \forall x \in \mathcal{X}, \forall a \in \{0,1\}. \tag{1}$$

The parameter λ may be viewed as a penalty for taking active action. The performance of any time-homogeneous policy $g : \mathcal{X} \to \{0,1\}$ is

$$J_\lambda^{(g)}(x_0) := (1-\beta)\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t c_\lambda(X_t, g(X_t)) \middle| X_0 = x_0\right]. \tag{2}$$

Consider the following optimization problem.

Problem 2 *Given the RBP* $(\mathcal{X}, \{0,1\}, \{P(a)\}_{a\in\{0,1\}}, c_\lambda, x_0)$ *and the discount factor* $\beta \in (0,1)$*, choose a Markov policy* $g : \mathcal{X} \to \{0,1\}$ *to minimize* $J^{(g)}_\lambda(x_0)$.

Problem 2 is also a Markov decision process and one can obtain an optimal solution using dynamic programming. Let $V_\lambda : \mathcal{X} \to \mathbb{R}$ be the unique fixed point of the following:

$$V_\lambda(x) = \min\bigl\{H_\lambda(x,0), H_\lambda(x,1)\bigr\}, \quad \forall x \in \mathcal{X}, \tag{3}$$

where

$$H_\lambda(x,a) = (1-\beta)c_\lambda(x,a) + \beta \sum_{y\in\mathcal{X}} P_{xy}(a)V_\lambda(y), \quad a \in \{0,1\}. \tag{4}$$

Let $g_\lambda(x)$ denote the minimizer of the right hand side of (3) where we set $g_\lambda(x) = 1$ if $H_\lambda(x,0) = H_\lambda(x,1)$. Then, from Markov decision theory [25], we know that the time-homogeneous policy g_λ is optimal for Problem 2.

Let Π_λ denote the set of states where passive action is optimal, i.e.,

$$\Pi_\lambda := \{x \in \mathcal{X} : g_\lambda(x) = 0\}. \tag{5}$$

We call Π_λ as the *passive set*.

Definition 1 (Indexability) *An RBP is indexable if* Π_λ *is increasing in* λ*, i.e., for any* $\lambda_1, \lambda_2 \in \mathbb{R}$*,* $\lambda_1 \le \lambda_2$ *implies* $\Pi_{\lambda_1} \subseteq \Pi_{\lambda_2}$.

Definition 2 (Whittle index) *The Whittle index of state* x *of an indexable RBP is the smallest value of* λ *for which* x *is part of the passive set* Π_λ*, i.e.,* $w(x) = \inf\{\lambda \in \mathbb{R} : x \in \Pi_\lambda\}$.

Alternatively, the Whittle index $w(x)$ is a value of the penalty λ for which the optimal policy is indifferent between taking active and passive action when the RBP is in state x.

2.4 Whittle index heuristic

A restless multi-armed bandit problem is said to be indexable if all RBPs are indexable. For indexable problems, the Whittle index heuristic is as follows: *Compute the Whittle indices of all arms offline. Then, at each time, obtain the Whittle indices of the current state of all bandits and play bandits 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, 14, 16]. However, there are only a few general conditions to check indexability for general models.

2.5 Alternative characterizations of passive set

We now present alternative characterizations of passive set, which is important for the sufficient conditions of indexability that we provide later.

Let Σ denote the family of all stopping times with respect to the natural filtration associated with $\{X_t\}_{t\ge0}$. Given an initial state $x \in \mathcal{X}$ and a stopping time $\tau \in \Sigma$, let h_τ denote the (history dependent) policy that takes passive action up to time $\tau - 1$, active action at time τ, and follows the optimal policy g_λ after that. Let

$$M(x,\tau) := \mathbb{E}[\beta^\tau | X_0 = x], \quad L(x,\tau) := \mathbb{E}\Bigl[\sum_{t=0}^{\tau-1} \beta^t c(X_t, 0) + \beta^\tau c(X_\tau, 1) \Bigm| X_0 = x\Bigr],$$

and

$$W_\lambda(x) := (1-\beta)\lambda + \beta \sum_{y\in\mathcal{X}} P_{xy}(1) V_\lambda(x). \tag{6}$$

We now present different characterizations of the passive set.

Proposition 1 *The following characterizations of the passive set are equivalent.*

- $\Pi_\lambda^{(a)} = \{x \in \mathcal{X} : g_\lambda(x) = 0\}$
- $\Pi_\lambda^{(b)} = \{x \in \mathcal{X} : H_\lambda(x, 0) < H_\lambda(x, 1)\}$
- $\Pi_\lambda^{(c)} = \{x \in \mathcal{X} : \exists \sigma \in \Sigma, \sigma \neq 0, \textit{ such that } J_\lambda^{(h_\sigma)}(x) < J_\lambda^{(h_0)}(x)\}$
- $\Pi_\lambda^{(d)} = \{x \in \mathcal{X} : \exists \sigma \in \Sigma, \sigma \neq 0, \textit{ such that } (1-\beta)\left(L(x,\sigma) - c(x,1)\right) < W_\lambda(x) - \mathbb{E}[\beta^\sigma W_\lambda(X_\sigma) | X_0 = x]\}$

See Appendix A for proof.

3 Sufficient conditions for indexability

In this section, we identify sufficient conditions for a RBP to be indexable.

3.1 Preliminary results

Consider a RBP $(\mathcal{X}, \{0,1\}, \{P(a)\}_{a\in\{0,1\}}, c, x_0)$. For any policy $g\colon \mathcal{X} \to \{0,1\}$ and $\lambda \in \mathbb{R}$, we can write

$$J_\lambda^{(g)}(x) = D^{(g)}(x) + \lambda N^{(g)}(x), \tag{7}$$

where

$$D^{(g)}(x) := (1-\beta)\mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t c(X_t, g(X_t)) \Big| X_0 = x\Big] \tag{8}$$

$$N^{(g)}(x) := (1-\beta)\mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t g(X_t) \Big| X_0 = x\Big] \tag{9}$$

are the expected discounted total cost and the expected number of activations under policy g starting at initial state x. $D^{(g)}(\cdot)$ and $N^{(g)}(\cdot)$ can be computed using policy evaluation formulas. In particular, define $P^{(g)}\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and $c^{(g)}\colon \mathcal{X} \to \mathbb{R}$ as follows: $P^{(g)}_{xy} = P_{xy}(g(x))$ and $c_\lambda^{(g)}(x) = c_\lambda(x, g(x)) = c^{(g)}(x, g(x)) + \lambda g(x)$ for any $x, y \in \mathcal{X}$. We also view g as an element in $\{0,1\}^{|\mathcal{X}|}$. Then, using policy evaluation formula for infinite horizon MDPs [25], we obtain

$$D^{(g)}(x) = (1-\beta)\big[(I-\beta P^{(g)})^{-1} c^{(g)}\big](x) \text{ and } N^{(g)}(x) = (1-\beta)\big[(I-\beta P^{(g)})^{-1} g\big](x). \tag{10}$$

The following two results follow immediately from (7).

Lemma 1 *For any $x \in \mathcal{X}$, $V_\lambda(x)$ is increasing and continuous in λ.*

Proof. The result follows from observing that $V_\lambda(x) = \min_{g:\mathcal{X}\to\{0,1\}} J_\lambda^{(g)}(x)$ and Equation (7) implies that $J_\lambda^{(g)}(x)$ is continuous and increasing in λ. □

Lemma 2 *For any $\lambda_1, \lambda_2 \in \mathbb{R}$,*

$$(\lambda_2 - \lambda_1) N^{(g_{\lambda_2})}(x) \le V_{\lambda_2}(x) - V_{\lambda_1}(x) \le (\lambda_2 - \lambda_1) N^{(g_{\lambda_1})}(x), \quad \forall x \in \mathcal{X}.$$

Consequently, $N^{(g_\lambda)}(x)$ is (weakly) decreasing in λ.

Proof. Recall that $V_\lambda(x) = J_\lambda^{(g_\lambda)}(x) \le J_\lambda^{(g_{\lambda'})}(x)$ for any $\lambda' \neq \lambda$. Thus,

$$\begin{aligned} V_{\lambda_2}(x) - V_{\lambda_1}(x) &= J_{\lambda_2}^{(g_{\lambda_2})}(x) - J_{\lambda_1}^{(g_{\lambda_1})}(x) \le J_{\lambda_2}^{(g_{\lambda_1})}(x) - J_{\lambda_1}^{(g_{\lambda_1})}(x) \\ &\overset{(a)}{=} (\lambda_2 - \lambda_1) N^{(g_{\lambda_1})}(x), \end{aligned} \tag{11}$$

where (a) follows from (7). Similarly, we have

$$\begin{aligned} V_{\lambda_2}(x) - V_{\lambda_1}(x) &= J_{\lambda_2}^{(g_{\lambda_2})}(x) - J_{\lambda_1}^{(g_{\lambda_1})}(x) \ge J_{\lambda_2}^{(g_{\lambda_2})}(x) - J_{\lambda_1}^{(g_{\lambda_2})}(x) \\ &\overset{(a)}{=} (\lambda_2 - \lambda_1) N^{(g_{\lambda_2})}(x), \end{aligned} \tag{12}$$

where (a) follows from (7). The result follows from combining the above inequalities. □

3.2 Sufficient conditions for indexability

Theorem 1 *Define $\mathcal{H} = \{(g,h) : g, h \colon \mathcal{X} \to \{0,1\}$ such that for all $x \in \mathcal{X}, N^{(g)}(x) \ge N^{(h)}(x)\}$. Each of the following is a sufficient condition for Whittle indexability:*

a. For any $g, h \in \mathcal{H}$, we have that for every $x, z \in \mathcal{X}$,

$$\sum_{y \in \mathcal{X}} \Big\{ \big[\beta P_{zy}(1) - P_{xy}(1)\big]^+ N^{(g)}(y) - \big[P_{xy}(1) - \beta P_{zy}(1)\big]^+ N^{(h)}(y) \Big\} \le \frac{(1-\beta)^2}{\beta}. \tag{13}$$

b. For any $g, h \in \mathcal{H}$, we have that for every $x \in \mathcal{X}$,

$$\sum_{y \in \mathcal{X}} \Big\{ \big[P_{xy}(0) - P_{xy}(1)\big]^+ N^{(g)}(y) - \big[P_{xy}(1) - P_{xy}(0)\big]^+ N^{(h)}(y) \Big\} \le \frac{1-\beta}{\beta}. \tag{14}$$

See Appendix B for the proof. The sufficient conditions of Theorem 1 are difficult to verify. Simpler sufficient conditions are stated below.

Proposition 2 *Each of the following is a sufficient condition for (13).*

a. $\max_{x,z \in \mathcal{X}} \sum_{y \in \mathcal{X}} \big[\beta P_{zy}(1) - P_{xy}(1)\big]^+ \le (1-\beta)^2/\beta$.
b. $P_{xy}(1) = P_{zy}(1)$, *for any* $x, z \in \mathcal{X}$.

In addition, each of the following is a sufficient condition for (14).

c. $\max_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} [P_{xy}(0) - P_{xy}(1)]^+ \le (1-\beta)/\beta$.
d. $\beta \le 0.5$.

See Appendix C for proof.

Some remarks

1. The sufficient conditions of Theorem 1 and Proposition 2 a, c, d may be viewed as bounds on the discount factor β for which the RBP is indexable. Numerical experiments to explore such a property are presented in [22].
2. We refer to models that satisfy the sufficient condition of Proposition 2.b as *restless bandits with controlled restarts*. Such models arise in various scheduling problems (e.g., machine maintenance, surveillance, etc.) where taking the active action resets the state according to known probability distribution. Specific instances of such models are considered in [4, 27]. The special case when the active action resets to a specific (pristine) state are considered in [18, 14]. Models where the passive action resets the bandit have been considered in [19, 8, 7].

4 Algorithm to compute Whittle index

Given an indexable RBP, a naive method to compute Whittle index at state x is to do a binary search over the penalty λ to find the critical penalty $w(x)$ such that for $\lambda \in (-\infty, w(x))$, $g_\lambda(x) = 0$ and for $\lambda \in [w(x), \infty)$, $g_\lambda(x) = 1$. Although such an approach has been used in the literature [26, 3], it is not efficient as it requires a separate binary search for each state. In this section, we present a more efficient algorithm to compute Whittle index.

Definition 3 *For any policy g, let $\texttt{ones}(g)$ denote the set of states where $g(x) = 1$. Given any state $y \in \texttt{ones}(g)$, let $\texttt{succ}(g, y)$ denote a policy which chooses the passive action at y and chooses the same action as g at all other states.*

As an example, consider $\mathcal{X} = \{1, 2, 3\}$ and $g = (0, 1, 1)$. Then, $\texttt{ones}(g) = \{2, 3\}$; $\texttt{succ}(g, 2) = (0, 0, 1)$ and $\texttt{succ}(g, 3) = (0, 1, 0)$.

Suppose $|\mathcal{X}| = K$. Define $\mathcal{K} = \{1, \dots, K\}$ and $\mathcal{K}^* = \{0, 1, \dots, K + 1\}$. Let $(x_1, \dots, x_K)$ be a permutation of states such that $w(x_1) \le w(x_2) \le \cdots \le w(x_K)$. For ease of notation, we define $\lambda_k = w(x_k)$ for $k \in \mathcal{K}$ and $\lambda_0 = -\infty$ and $\lambda_{K+1} = \infty$. Thus, $\{\lambda_k\}_{k \in \mathcal{K}^*}$ is a (weakly) increasing sequence.

Now for any $k \in \mathcal{K}^*$, define a policy g_k as follows: for any $\ell \in \mathcal{K}$, $g_k(x_\ell) = 0$ if $\ell \le k$ and 1 otherwise. Note that g_0 prescribes the active action at all states and g_{K+1} prescribes the passive action at all states. By construction, $x_{k+1} \in \texttt{ones}(g_k)$ and $g_{k+1} = \texttt{succ}(g_k, x_{k+1})$.

Indexability implies that for any $\lambda \in [\lambda_k, \lambda_{k+1})$ the policy g_k is optimal. Therefore, for any other policy h and any state $x \in \mathcal{X}$, $J^{(g_k)}_\lambda(x) \le J^{(h)}_\lambda(x)$ or, equivalently, $\lambda \le (D^{(h)}(x) - D^{(g_k)}(x))/(N^{(g_k)}(x) - N^{(h)}(x))$ with equality if $h = g_{k+1}$. Thus,

$$\lambda_{k+1} \le \frac{D^{(h)}(x) - D^{(g_k)}(x)}{N^{(g_k)}(x) - N^{(h)}(x)}, \tag{15}$$

with equality if $h = g_{k+1}$. This implies the following.

Theorem 2 *For any $k \in \mathcal{K}^*$, let $y \in \texttt{ones}(g_k)$ and $h = \texttt{succ}(g_k, y)$. Define*

$$\lambda^\circ_{g_k,y}(x) = \frac{D^{(h)}(x) - D^{(g_k)}(x)}{N^{(g_k)}(x) - N^{(h)}(x)}. \tag{16}$$

Then, $\lambda^\circ_{g_k,x_{k+1}}(x) = \lambda_{k+1}$ and does not depend on x. Moreover, for any $y \in \texttt{ones}(g_k)$ and $x \in \mathcal{X}$, $\lambda_{k+1} \le \lambda^\circ_{g_k,y}(x)$.

Proof. By construction, g_k is the optimal policy for $\lambda \in [\lambda_k, \lambda_{k+1})$ and g_{k+1} is the optimal policy for $\lambda \in [\lambda_{k+1}, \lambda_{k+2})$. Policies g_k and g_{k+1} differ only at state x_{k+1}. From Lemma 1, we know that $V_\lambda(x)$ is continuous in λ for all $x \in \mathcal{X}$. Thus, for all $x \in \mathcal{X}$,

$$\lim_{\lambda \uparrow \lambda_{k+1}} J^{(g_k)}_\lambda(x) = \lim_{\lambda \downarrow \lambda_{k+1}} J^{(g_{k+1})}_\lambda(x)$$

Thus, $J^{(g_k)}_{\lambda_{k+1}}(x) = J^{(g_{k+1})}_{\lambda_{k+1}}(x)$ and, therefore,

$$D^{(g_k)}(x) + \lambda_{k+1} N^{(g_k)}(x) = D^{(g_{k+1})}(x) + \lambda_{k+1} N^{(g_{k+1})}(x), \quad \forall x \in \mathcal{X}.$$

This implies that $\lambda^\circ_{g_k,x_{k+1}}(x) = \lambda_{k+1}$ and does not depend on x. The fact that $\lambda_{k+1} \le \lambda^\circ_{g_k,y}(x)$ follows from (15). □

Theorem 2 suggests a simple algorithm to identify the permutation $(x_1, \dots, x_K)$ and the corresponding Whittle indices. Recall that for any policy g, $D^{(g)}$ and $N^{(g)}$ can be computed using (10). We first identify x_1 as any $\arg\min_{y \in \texttt{ones}(g_0)} \lambda^\circ_{g_0,y}(x)$. Then $w(x_1) = \lambda^\circ_{g_0,x_1}(x)$.

Algorithm 1 Computing Whittle index of all states of an indexable RBP.

Input: RBP $(\mathcal{X}, \{0,1\}, P(a)_{a\in\{0,1\}}, c, x_0)$, Discount factor β.
let $k = 0$ and $g_0(x) = 1$ for all $x \in \mathcal{X}$
while $k < |\mathcal{X}|$ **do**
  let $\lambda = \min_{y\in\texttt{ones}(g_k)} \min_{x\in\mathcal{X}} \lambda^\circ_{g_k,y}(x)$.
  let $\texttt{next}(g_k) = \arg\min_{y\in\texttt{ones}(g_k)} \min_{x\in\mathcal{X}} \lambda^\circ_{g_k,y}(x)$.
  for $y \in \texttt{next}(g_k)$ **do**
    let $w(y) = \lambda$.
    let $g_{k+1} = \texttt{succ}(g_k, y)$.
    let $k \leftarrow k + 1$.
  end for
end while

Now assume that $(x_1, \ldots, x_k)$ have been identified and we are interested in identifying $(x_{k+1}, w(x_{k+1}))$. By Theorem 2, we have that $x_{k+1} \in \arg\min_{y\in\texttt{ones}(g_k)} \lambda^\circ_{g_k,y}(x)$ and $w(x_{k+1}) = \lambda^\circ_{g_k,x_{k+1}}(x)$. Continuing this way, we can identify $(x_1, w(x_1))$, $\ldots$, $(x_K, w(x_K))$. The detailed algorithm, where we take care of multiplicity of arg min, is presented in Algorithm 1.

Some remarks

1. The idea of computing the index by iteratively sorting the states according to their index is commonly used in the offline algorithms to compute Gittins index; for example, the largest-remaining-index algorithm, the state-elimination algorithm, the triangularization algorithm, and the fast-pivoting algorithm use variations of this idea. See [10] for details.
2. The term $\lambda^\circ_{g_k,y}(x)$ is equal to the marginal productivity index for general resource allocation problems [23]. The algorithm proposed above is similar in spirit to the adaptive greedy algorithm of [21].
3. Suppose computing $D^{(g)}$ and $N^{(g)}$ requires d computations.[1] Then, the worst case complexity of Algorithm 1 is $\sum_{k=1}^{K}(K-k)d = K(K-1)d/2$.

An illustrative example Consider a RBP with $\mathcal{X} = \{1,2,3,4\}$, $\beta = 0.75$,

$$P(0) = \begin{bmatrix} 0.2 & 0.3 & 0.2 & 0.3 \\ 0.1 & 0.3 & 0.5 & 0.1 \\ 0.2 & 0.1 & 0.3 & 0.4 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{bmatrix}, \quad P(1) = \begin{bmatrix} 0.3 & 0.2 & 0.0 & 0.5 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0.0 & 0.0 & 0.5 & 0.5 \\ 0.5 & 0.0 & 0.2 & 0.3 \end{bmatrix}, \quad c_0 = \begin{bmatrix} 1 \\ 2 \\ 5 \\ 4 \end{bmatrix}, \quad c_1 = \begin{bmatrix} 5 \\ 1 \\ 4 \\ 8 \end{bmatrix}.$$

It can be verified that the above model satisfies condition (c) of Proposition 2. Thus, the RBP is indexable.

To compute the Whittle index, we start with policy g_0 and compute $\lambda^\circ_{g_0,y}$ for all $y \in \texttt{ones}(g_0) = \{1,2,3,4\}$. The smallest value is -5.9815 at $y = 4$. Thus, $x_1 = 4$ and $w(4) = -5.9815$.

Now, $g_1 = \texttt{succ}(g_0, x_1) = (1,1,1,0)$. We compute $\lambda^\circ_{g_1,y}$ for all $y \in \texttt{ones}(g_1) = \{1,2,3\}$. The smallest value is -4.8728 for $y = 1$. Thus, $x_2 = 1$ and $w(1) = -4.8728$.

Next, $g_2 = \texttt{succ}(g_1, x_2) = (0,1,1,0)$. We compute $\lambda^\circ_{g_2,y}$ for all $y \in \texttt{ones}(g_2) = \{2,3\}$. The smallest value is 0.0886 for $y = 3$. Thus, $x_3 = 3$ and $w(3) = 0.0886$.

Finally, $g_3 = \texttt{succ}(g_2, x_3) = (0,1,0,0)$. The set $\texttt{ones}(g_3) = \{2\}$ is a singleton. We compute $\lambda^\circ_{g_3,y}$ for $y = 2$ and it equals 1.7274. Thus, $x_4 = 2$ and $w(4) = 1.7274$.

5 Some special cases

In this section, we show how the results developed in this paper can be refined for some special cases.

[1] The exact dependence of d on the size K of the state space depends on the structure of the transition matrix and the method used to solve the linear system in (10). Typically $d = O(K^3)$ for dense matrices and $O(K)$ for sparse matrices.

5.1 Stochastic monotone bandits

Consider a RBP $(\mathcal{X}, \{0,1\}, \{P(a)\}_{a\in\{0,1\}}, c, x_0)$ where the state space $\mathcal{X}$ is a totally ordered set. We say that the RBP is *stochastic monotone* if it satisfies the following conditions.

(D1) For any $a \in \{0,1\}$, $P(a)$ is stochastically monotone, i.e., for any $x, y \in \mathcal{X}$ such that $x < y$, we have $\sum_{w\in\mathcal{X}_{\geq z}} P_{xw}(a) \leq \sum_{w\in\mathcal{X}_{\geq z}} P_{yw}(a)$ for any $z \in \mathcal{X}$.

(D2) For any $z \in \mathcal{X}$, $S_{zx}(a) := \sum_{w\in\mathcal{X}_{\geq z}} P_{xw}(a)$ in submodular[2] in (x, a).

(D3) For any $a \in \{0,1\}$, $c(x, a)$ is (weakly) increasing in x.

(D4) $c(x, a)$ is submodular in (x, a).

For ease of notation, we use $\mathcal{X}^* = \mathcal{X} \cup \{\star\}$, where $\star$ is an element which is smaller than all elements of $\mathcal{X}$. Under (D1)–(D4), we have the following.

Lemma 3 *For a stochastic monotone RBP, the optimal policy g_λ for any $\lambda \in \mathbb{R}$ is threshold based, i.e., there exists a threshold $\ell_\lambda \in \mathcal{X}^*$ such that the policy $g^{(\ell_\lambda)}(x) = 0$ if $x \leq \ell_\lambda$ and 1 otherwise. If there are multiple such thresholds, ℓ_λ denotes the largest threshold.*

Proof. Conditions (D1)–(D4) are the same as the properties of [25, Theorem 4.7.4], which implies an optimal policy exists and it is threshold based. □

Proposition 3 *Consider a stochastic monotone RBP which satisfies the following condition.*

(D5) *For any $x \in \mathcal{X}$, $N^{(g^{(\ell)})}$ is (weakly) decreasing in ℓ.*

Then, ℓ_λ is increasing with λ. Therefore, a stochastic monotone RBP is indexable.

Proof. We first show that for any $\ell \in \mathcal{X}^*$, $J_\lambda^{(g^{(\ell)})}(x)$ is submodular in (ℓ, λ) for all $x \in \mathcal{X}$. In particular, for any $k < \ell$, we have

$$J_\lambda^{(g^{(\ell)})}(x) - J_\lambda^{(g^{(k)})}(x) = D_\lambda^{(g^{(\ell)})}(x) - D_\lambda^{(g^{(k)})}(x) + \lambda(N_\lambda^{(g^{(\ell)})}(x) - N_\lambda^{(g^{(k)})}(x)).$$

Now (D5) implies that the difference $J_\lambda^{(g^{(\ell)})}(x) - J_\lambda^{(g^{(k)})}(x)$ is decreasing in λ. Therefore, $J_\lambda^{(g^{(\ell)})}(x)$ is submodular in (ℓ, λ). Consequently, from [25, Theorem 2.8.2], $\ell_\lambda = \max\{\ell' \in \arg\min_{\ell\in\mathcal{X}^*} J_\lambda^{(g^{(\ell)})}(x)\}$ is increasing in λ.

The fact that ℓ_λ is increasing in λ implies that the set $\Pi_\lambda = \{x \in \mathcal{X} : g_\lambda(x) = 0\}$ is increasing in λ. □

Combining Lemma 3 and Proposition 3, we get that under (D1)–(D5), the model satisfies the following property:

(P) There exists a (weakly) increasing family of thresholds $\{\ell_\lambda\}_{\lambda\in\mathbb{R}}$ such that the threshold policy $g^{(\ell_\lambda)}$ is optimal for Problem 2.

For specific models, condition (P) may hold under weaker set of assumptions. In fact, several models where (P) holds have been considered in the literature [5, 17, 7, 15, 27, 4].

Condition (P) implies that the Whittle index $w(x)$ is (weakly) increasing in x. Therefore, we can directly compute the Whittle index of state x_k in closed form:

$$w(x_k) = \frac{D^{(g^{(x_{k+1})})}(x) - D^{(g^{(x_k)})}(x)}{N^{(g^{(x_k)})}(x) - N^{(g^{(x_{k+1})})}(x)}.$$

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

5.2 Restless bandits with controlled restarts

Consider restless bandits with controlled restarts (i.e., models where $P_{xy}(1)$ does not depend on x). By Proposition 2c, such models are indexable. In this section, we explain how to simplify the computation of the Whittle index for such models. For ease of notation, we use P_{xy} to denote $P_{xy}(0)$ and Q_y to denote $P_{xy}(1)$.

Define

$$\mathsf{D}^{(g)} = \sum_{x\in\mathcal{X}} Q_x D^{(g)}(x) \quad \text{and} \quad \mathsf{N}^{(g)} = \sum_{x\in\mathcal{X}} Q_x N^{(g)}(x).$$

Now, following the discussion of Section 4, we can show that in addition to (15), we can establish that

$$\lambda_{k+1} \le \frac{\mathsf{D}^{(h)} - \mathsf{D}^{(g_k)}}{\mathsf{N}^{(g_k)} - \mathsf{N}^{(h)}}, \tag{17}$$

with equality if $h = g_{k+1}$. Therefore, the result of Theorem 2 continues to holds when $\lambda^{\circ}_{g_k,y}(x)$ replaced by

$$\lambda^*_{g_k,y} = \frac{\mathsf{D}^{(h)} - \mathsf{D}^{(g_k)}}{\mathsf{N}^{(g_k)} - \mathsf{N}^{(h)}}.$$

Therefore, we can replace $\lambda^{\circ}_{g_k,y}(x)$ in Algorithm 1 by $\lambda^*_{g_k,y}$. Our key result for this section is $\mathsf{D}^{(g)}$ and $\mathsf{N}^{(g)}$ (and therefore $\lambda^*_{g_k,y}$) can be computed efficiently for models with controlled restarts.

For that matter, given any policy g, let τ_g denote the hitting time of the set $\Pi^{(g)} = \{x \in \mathcal{X} : g(x) = 1\}$. Let

$$\mathsf{L}^{(g)} := \mathbb{E}\Big[\sum_{t=0}^{\tau_g} \beta^t c(X_t, g(X_t)) \;\Big|\; X_0 \sim Q\Big] \quad \text{and} \quad \mathsf{M}^{(g)} := \mathbb{E}\Big[\sum_{t=0}^{\tau_g} \beta^t \;\Big|\; X_0 \sim Q\Big]$$

denote the expected discounted cost and expected discounted time for hitting $\Pi^{(g)}$ starting with an initial state distribution of Q. Then, using ideas from renewal theory, we can show the following.

Theorem 3 *For any policy g,*

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

Proof. The proof follows from standard ideas in renewal theory. By strong Markov property, we have

$$\begin{aligned}\mathsf{D}^{(g)} &= \mathbb{E}\Big[(1-\beta)\sum_{t=0}^{\tau_k} \beta^t c(X_t, g(X_t)) + \beta^{\tau_k+1}\mathsf{D}^{(g)} \;\Big|\; X_0 \sim Q\Big] \\ &= (1-\beta)\mathsf{L}^{(g)} + \mathbb{E}[\beta^{\tau_k+1} | X_0 \sim Q]\mathsf{D}^{(g)}.\end{aligned} \tag{18}$$

Using $\mathsf{M}^{(g)}$ definition, we have $\mathbb{E}[\beta^{\tau_k+1}|X_0 \sim Q] = 1 - (1-\beta)\mathsf{M}^{(g)}$. Substituting this in (18) and rearranging the terms we get $\mathsf{D}^{(g)} = \mathsf{L}^{(g)}/\mathsf{M}^{(g)}$.

For $\mathsf{N}^{(g)}$, by strong Markov property we have

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

Therefore, we get $\mathsf{N}^{(g)} = \big(1-(1-\beta)\mathsf{M}^{(g)}\big)/\beta\mathsf{M}^{(g)}$. □

Given any policy g, we can efficiently compute $\mathsf{L}^{(g)}$ and $\mathsf{M}^{(g)}$ using standard formulas for truncated Markov chains. For any vector v, let $v^{(g)}$ denote the vector with components indexed by the set $\{x \in \mathcal{X} : g(x) = 0\}$ and $\tilde{v}^{(g)}$ denote the remaining components. For example, if $\mathcal{X} = \{1, 2, 3, 4\}$,

$g = (1, 0, 1, 0)$, and $v = [1, 2, 3, 4]$, then $v^{(g)} = (2, 4)$ and $\tilde{v}^{(g)} = (1, 3)$. Similarly, for any square matrix Z, let $Z^{[g]}$ denote the square sub-matrix corresponding to elements $\{x \in \mathcal{X} : g(x) = 0\}$, and $\tilde{Z}^{[g]}$ denote the sub-matrix with rows $\{x \in \mathcal{X} : g(x) = 0\}$ and columns $\{x \in \mathcal{X} : g(x) = 1\}$. For example, in the above example,

$$\text{if } Z = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 8 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}, \text{ then } Z^{[g]} = \begin{bmatrix} 6 & 8 \\ 14 & 16 \end{bmatrix} \text{ and } \tilde{Z}^{[g]} = \begin{bmatrix} 5 & 8 \\ 13 & 15 \end{bmatrix}.$$

Then, from standard formulas for truncated Markov chains, we have

Proposition 4 *For any policy g, let c_0 and c_1 denote column vectors corresponding to $c(\cdot, 0)$ and $c(\cdot, 1)$. Then,*

$$\mathsf{L}^{(g)} = Q^{(g)}(I - \beta P^{[g]})^{-1}(c_0^{(g)} + \beta \tilde{P}^{[g]} \tilde{c}_1^{(g)}) + \tilde{Q}^{(g)} \tilde{c}_1^{(g)},$$
$$\mathsf{M}^{(g)} = Q^{(g)}(I - \beta P^{[g]})^{-1}(\mathbf{1}^{(g)} + \beta \tilde{P}^{[g]} \tilde{\mathbf{1}}^{(g)}) + \tilde{Q}^{(g)} \tilde{\mathbf{1}}^{(g)}.$$

This gives us an efficient method to compute $\mathsf{L}^{(g)}$ and $\mathsf{M}^{(g)}$, which can in turn be used to compute $\mathsf{D}^{(g)}$ and $\mathsf{N}^{(g)}$ and used in a modified version of Algorithm 1 as explained above.

6 Numerical experiments

In this section, we 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) (shown in Algorithm 2). The code is available at [2].

Algorithm 2 Myopic Heuristic.

```
1: Input:  Set 𝒩 of arms; arms m to be activated, t = 1.
2: while t ≥ 1 do
3:    let ℓ = 0, ℳ = ∅, and 𝒵 = 𝒩.
4:    while ℓ ≤ m do
5:       Let i*_ℓ ∈ arg min_{i∈𝒵} Σ_{j∈𝒵\{i}} c^j(X^j_t, 0) + c^i(X^i_t, 1).
6:       let ℳ = ℳ ∪ {i*_ℓ}, 𝒵 = 𝒵 \ {i*_ℓ}.
7:       ℓ = ℓ + 1.
8:    end while
9:    Activate arms in 𝒵.
10:   t = t + 1.
11: end while
```

6.1 Experimental setup

In our experiments, we consider restart bandits with $P(1) = [\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0}]$. There are two other components of the model: The transition matrix $P(0)$ and the cost function c. We choose these components as follows.

6.1.1 The choice of transition matrices.

We have three setups for choosing $P(0)$. The first setup is a family of 4 types of structured stochastic monotone matrices, which we denote by $\mathcal{P}_\ell(p)$, $\ell \in \{1, \ldots, 4\}$, 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. In the third setup, we generate random stochastic matrices using Levy distribution. The details of these models are presented in the supplementary material.

6.1.2 The choice of the cost function.

For all our experiments we choose $c(x, 0) = (x - 1)^2$ and $c(x, 1) = 0.5(|\mathcal{X}| - 1)^2$.

6.2 Experimental details and result

We conduct different experiments to compare the performance of Whittle index with the optimal policy and the myopic policy for different setups (described in Section 6.1) and for different sizes $|\mathcal{X}|$ 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 $\beta = 0.95$.

We evaluate the performance of a policy via Monte Carlo simulations over S trajectories, where each trajectory is of length T. In all our experiments, we choose $S = 2500$ and $T = 250$.

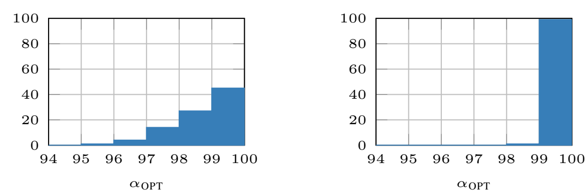
Experiment 1) Comparison of Whittle index with the optimal policy for structured models.

The optimal policy is computed by solving the MDP for Problem 1. The state for this MDP is $|\mathcal{X}|^n$. So, we can obtain the optimal policy only for small values of $|\mathcal{X}|$ and n. We choose $|\mathcal{X}| = 5$ 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 $\alpha_{\text{OPT}} = J(\text{OPT})/J(\text{WIP})$ denote the relative performance (in percentage) of WIP compared to OPT. In our experiments, α_{OPT} was in the range of 99.95%–100% for all choices of the problem parameters.

Experiment 2) Comparison of Whittle index with the optimal policy for randomly sampled models.

As before, we pick $|\mathcal{X}| = 5$ and $n = 5$ so that it is feasible to calculate the optimal policy. For each arm, we sample the transition matrix from $\mathcal{R}(5/|\mathcal{X}|)$. We repeat the experiment 250 times. The histogram of α_{OPT} over experiments for $m \in \{1, 2\}$ is plotted in Figure 1. Similar to the result of Experiment 1, WIP has a reasonable relative performance with respect to OPT.



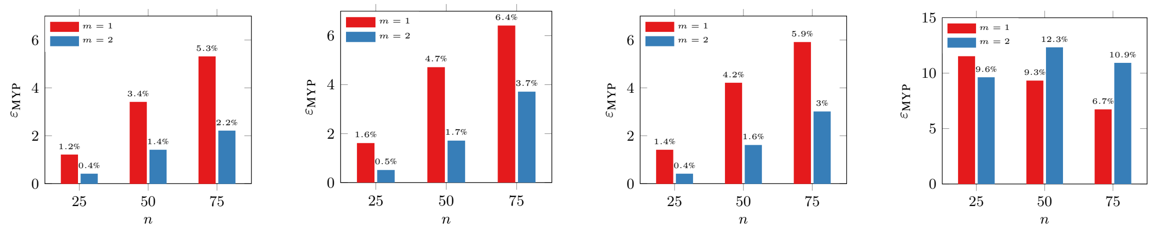
(a) $m = 1$. **(b)** $m = 2$.

Figure 1: 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 $|\mathcal{X}| = 25$, $n \in \{25, 50, 75\}$, and $m \in \{1, 2, 5\}$. In this case, let $\varepsilon_{\text{MYP}} = (J(\text{MYP}) - J(\text{WIP}))/J(\text{MYP})$ denote the relative improvement of WIP compared to MYP. The results of ε_{MYP} for different choice of the parameters are shown in Figure 2.

In Figure 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 Figure 2d with respect to n varies differently from the rest of the models.

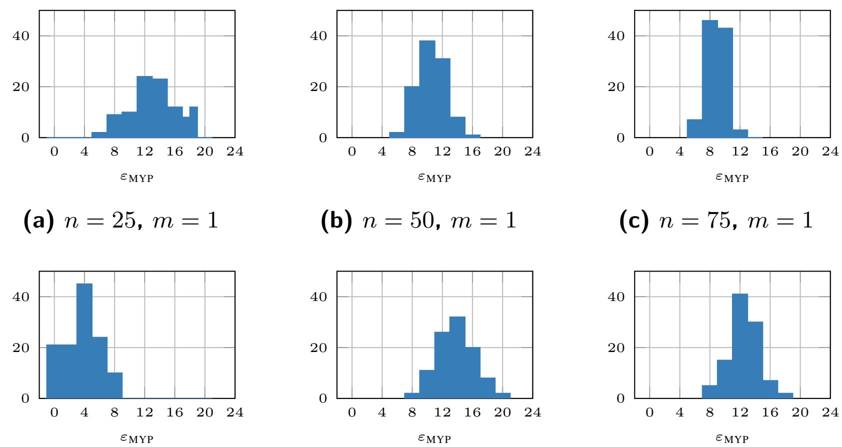


(a) $\ell = 1$ **(b)** $\ell = 2$ **(c)** $\ell = 3$ **(d)** $\ell = 4$

Figure 2: Relative improvement ε_{myp} of wip vs. myp for $|\mathcal{X}| = 25$ when $\ell \in \{1, \dots, 4\}$, $n \in \{25, 50, 75\}$, and $m \in \{1, 2\}$ for Experiment 3.

Experiment 4) Comparison of Whittle index with the myopic policy for randomly sampled models.

We generate 250 random models as described in Experiment 2 but for $|\mathcal{X}| = 25$ 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 Figure 3.



(a) $n = 25$, $m = 1$ **(b)** $n = 50$, $m = 1$ **(c)** $n = 75$, $m = 1$

(d) $n = 25$, $m = 2$ **(e)** $n = 50$, $m = 2$ **(f)** $n = 75$, $m = 2$

Figure 3: Relative improvement ε_{myp} of wip vs. myp for $|\mathcal{X}| = 25$ when $n \in \{25, 50, 75\}$, and $m \in \{1, 2\}$ for Experiment 4.

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}.

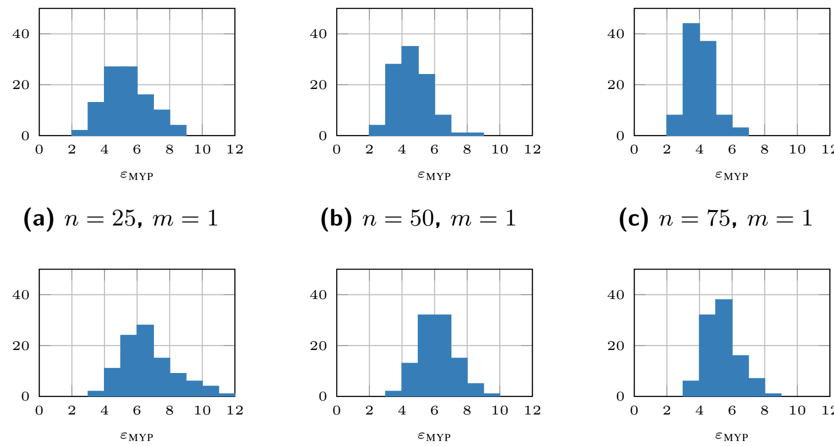
Experiment 5) Comparison of Whittle index with the myopic policy for restart models.

We generate 250 random stochastic matrices for $P(0)$.[3] We set $|\mathcal{X}| = 25$ and $n \in \{25, 50, 75\}$ and $m \in \{1, 2\}$. For each case, ε_{MYP} is computed and the histogram of ε_{MYP} for different choices of the parameters is shown in Figure 4.

7 Conclusion

We present two general sufficient conditions for restless bandit processes to be indexable. The first condition depends only on the transition matrix $P(1)$ while the second condition depends on both $P(0)$ and $P(1)$. These sufficient conditions are based on alternative characterizations of the passive set, which might be useful in general as well. We also present refinements of these sufficient conditions that are simpler to verify. Two of these simpler conditions are worth highlighting: models where the

[3] Each row of the matrix is generate according to Section 1.3 of the supplementary material.



(a) $n = 25$, $m = 1$ **(b)** $n = 50$, $m = 1$ **(c)** $n = 75$, $m = 1$

(d) $n = 25$, $m = 2$ **(e)** $n = 50$, $m = 2$ **(f)** $n = 75$, $m = 2$

Figure 4: Relative improvement $\varepsilon_{\mathsf{myp}}$ of wip vs. myp for $|\mathcal{X}| = 25$ when $n \in \{25, 50, 75\}$, and $m \in \{1, 2\}$ for Experiment 5.

active action resets the state according to a known distribution and models where the discount factor is less than 0.5.

We present a general algorithm to compute Whittle index for indexable RBP. The main idea of the algorithm is to identify a permutation $(x_1, \dots, x_K)$ of the states such that $\{w(x_k)\}_{k \in \mathcal{K}}$ forms a (weakly) increasing sequence.

Finally, we show how to refine the results for two classes for restless bandits: stochastic monotone bandits and restless bandits with controlled restarts. We also present a detailed numerical study which shows that Whittle index policy performs close to the optimal policy and considerably better than a myopic policy.

A Proof of Proposition 1

We first present a preliminary result.

Lemma 4 *For $\tau = 0$, the policy h_0 satisfies $J_\lambda^{(h_0)}(x) = H_\lambda(x, 1) = (1 - \beta)c(x, 1) + W_\lambda$.*

Proof. Consider the stopping time $\tau = 0$. The policy h_0, takes the active action at time 0 and follows the optimal policy afterwards. Thus, for any $x \in \mathcal{X}$, $J^{(h_0)}(x) = (1 - \beta)(c(x, 1) + \lambda) + \beta \sum_{y \in \mathcal{X}} P_{xy}(1) V_\lambda(y) = H_\lambda(x, 1)$. By (4) and (6) we have $H_\lambda(x, 1) = (1 - \beta)c(x, 1) + W_\lambda(x)$. □

We now proceed with the proof of Proposition 1. By definition, $\Pi^{(a)} = \Pi_\lambda$. We establish the equality of other characterizations.

(i) $\Pi_\lambda^{(a)} = \Pi_\lambda^{(b)}$. We have $x \in \Pi_\lambda \overset{(a)}{\iff} g_\lambda(x) = 0 \overset{(b)}{\iff} H_\lambda(x, 0) < H_\lambda(x, 1)$ where (a) follows from (5) and (b) follows from the dynamic program (3).

(ii) $\Pi_\lambda^{(b)} \subseteq \Pi_\lambda^{(c)}$. Let σ denote the hitting time of $\mathcal{X} \setminus \Pi_\lambda$. If we start in state $x \in \Pi_\lambda^{(b)} = \Pi_\lambda$, then the policy h_σ is same as the optimal policy. Hence, $J_\lambda^{(h_\sigma)}(x) = H_\lambda(x, 0)$. Thus, for any $x \in \Pi_\lambda^{(b)} = \Pi_\lambda$, $J_\lambda^{(h_\sigma)}(x) = H_\lambda(x, 0) \overset{(a)}{<} H_\lambda(x, 1) \overset{(b)}{=} J_\lambda^{(h_0)}(x)$ where (a) follows from fact that $x \in \Pi^{(b)}$ and (b) from Lemma 4.

(iii) $\Pi_\lambda^{(c)} \subseteq \Pi_\lambda^{(b)}$. Let $x \in \Pi_\lambda^{(c)}$ and $\sigma \in \Sigma$ denote a stopping time such that $J_\lambda^{(h_\sigma)}(x) < J_\lambda^{(h_0)}(x)$. Now, the optimal policy performs at least as well as policy h_σ. Therefore, $V_\lambda(x) \leq J_\lambda^{(h_\sigma)}(x)$. Combining this result with Lemma 4 we have $V_\lambda(x) < H_\lambda(x, 1)$. Thus, we must have $V_\lambda(x) = H_\lambda(x, 0)$ which results in $H_\lambda(x, 0) < H_\lambda(x, 1)$ which implies $x \in \Pi_\lambda^{(b)}$.

(iv) $\Pi_\lambda^{(c)} = \Pi_\lambda^{(d)}$. According to the definitions of $L(x,\tau)$ and $W_\lambda(x)$ we have

$$J_\lambda^{(h_\tau)}(x) = (1-\beta)L(x,\tau) + \mathbb{E}[\beta^\tau W_\lambda(X_\tau)|X_0 = x]. \tag{19}$$

Thus, $J_\lambda^{(h_\sigma)}(x) < J_\lambda^{(h_0)}(x)$ if and only if

$$(1-\beta)L(x,\sigma) + \mathbb{E}[\beta^\sigma W_\lambda(X_\sigma)|X_0 = x] < (1-\beta)c(x,1) + W_\lambda(x) \tag{20}$$

where we have used (19) for $J_\lambda^{(h_\tau)}(x)$ and Lemma 4 for $J_\lambda^{(h_0)}(x)$. Rearranging the terms of (20) we get the expression in $\Pi_\lambda^{(d)}$. Hence, $\Pi_\lambda^{(c)} = \Pi_\lambda^{(d)}$.

B Proof of Theorem 1

B.1 Proof of Theorem 1.a

We first present a preliminary result.

Lemma 5 *Under (13), for any $\lambda_1 < \lambda_2$ and $\sigma \in \Sigma$, $\sigma \neq 0$, we have that for any $x \in \mathcal{X}$,*

$$W_{\lambda_1}(x) - \mathbb{E}[\beta^\sigma W_{\lambda_1}(X_\sigma)|X_0 = x] \leq W_{\lambda_2}(x) - \mathbb{E}[\beta^\sigma W_{\lambda_2}(X_\sigma)|X_0 = x],$$

Proof. By (6), we have for any $x \in \mathcal{X}$,

$$\begin{aligned}&(W_{\lambda_2}(x) - \mathbb{E}[\beta^\sigma W_{\lambda_2}(X_\sigma)|X_0 = x]) - (W_{\lambda_1}(x) - \mathbb{E}[\beta^\sigma W_{\lambda_1}(X_\sigma)|X_0 = x])\\ &\quad = (1-\beta)\Delta_\lambda\big(1 - M(x,\sigma)\big) + \beta\mathbb{E}\Big[\sum_{y\in\mathcal{X}}\big(P_{xy}(1) - \beta^\sigma P_{X_\sigma y}(1)\big)\big(V_{\lambda_2}(y) - V_{\lambda_1}(y)\big)\;\Big|\; X_0 = x\Big]\end{aligned} \tag{21}$$

Now since $\sigma \geq 1$, $M(x,\sigma) \leq \beta$ and,

$$(1-\beta)\Delta_\lambda(1 - M(x,\sigma)) \geq \Delta_\lambda(1-\beta)^2 \tag{22}$$

Now consider,

$$\begin{aligned}&\beta\mathbb{E}\Big[\sum_{y\in\mathcal{X}}\big(P_{xy}(1) - \beta^\sigma P_{X_\sigma y}(1)\big)\big(V_{\lambda_2}(y) - V_{\lambda_1}(y)\big)\;\Big|\; X_0 = x\Big]\\ &\overset{(a)}{\geq} \beta\mathbb{E}\Big[\sum_{y\in\mathcal{X}}\big(P_{xy}(1) - \beta P_{X_\sigma y}(1)\big)\big(V_{\lambda_2}(y) - V_{\lambda_1}(y)\big)\;\Big|\; X_0 = x\Big]\\ &\overset{(b)}{\geq} \beta\Delta_\lambda\mathbb{E}\Big[\sum_{y\in\mathcal{X}}\Big\{\big[P_{xy}(1) - \beta P_{X_\sigma y}(1)\big]^+ N^{(g_{\lambda_2})}(y)\\ &\qquad\qquad + \big[P_{xy}(1) - \beta P_{X_\sigma y}(1)\big]^- N^{(g_{\lambda_1})}(y)\Big\}\Big| X_0 = x\Big]\\ &\overset{(c)}{\geq} -\Delta_\lambda(1-\beta)^2,\end{aligned} \tag{23}$$

where (a) holds due to $\sigma \geq 1$ and (b) holds by Lemma 2 and (c) follows from (13). Substituting (22) and (23) in (21), we get the result of the Lemma. □

We now proceed with the proof of Theorem 1a. Consider $\lambda_1 < \lambda_2$. Suppose $x \in \Pi_{\lambda_1}$. By Proposition 1.d, there exists a $\sigma \neq 0$ such that $(1-\beta)\left(L(x,\sigma) - c(x,1)\right) < W_{\lambda_1}(x) - \mathbb{E}[\beta^\sigma W_{\lambda_1}(X_\sigma)|X_0 = x]$. Combining this result with the result of Lemma 5, we infer

$$(1-\beta)\left(L(x,\sigma) - c(x,1)\right) < W_{\lambda_2}(x) - \mathbb{E}[\beta^\sigma W_{\lambda_2}(X_\sigma)|X_0 = x].$$

Thus, $x \in \Pi_{\lambda_2}$. Hence, $\Pi_{\lambda_1} \subseteq \Pi_{\lambda_2}$ and the RBP is indexable.

B.2 Proof of Theorem 1.b

Consider $\lambda_1 < \lambda_2$. A RBP is indexable if $\Pi_{\lambda_1} \subseteq \Pi_{\lambda_2}$ or equivalently, for any x such that $H_{\lambda_1}(x,0) < H_{\lambda_1}(x,1)$ then $H_{\lambda_2}(x,0) < H_{\lambda_2}(x,1)$. A sufficient condition for that is to show that $H_{\lambda_1}(x,1) - H_{\lambda_1}(x,0) \le H_{\lambda_2}(x,1) - H_{\lambda_2}(x,0)$, or equivalently, show that $H_{\lambda_2}(x,0) - H_{\lambda_1}(x,0) \le H_{\lambda_2}(x,1) - H_{\lambda_1}(x,1)$. We prove this inequality as follows.

Let $\Delta_\lambda = \lambda_2 - \lambda_1$. By (4), we have for any $x \in \mathcal{X}$,

$$\begin{aligned}
&(H_{\lambda_2}(x,1) - H_{\lambda_1}(x,1)) - (H_{\lambda_2}(x,0) - H_{\lambda_1}(x,0)) \\
&\quad = \Delta_\lambda(1-\beta) + \beta \sum_{y\in\mathcal{X}} (P_{xy}(1) - P_{xy}(0))(V_{\lambda_2}(y) - V_{\lambda_1}(y)) \\
&\quad \overset{(a)}{\ge} \Delta_\lambda \Big(1 - \beta + \beta \sum_{y\in\mathcal{X}} [P_{xy}(1) - P_{xy}(0)]^+ N^{(g_{\lambda_2})}(y) + [P_{xy}(1) - P_{xy}(0)]^- N^{(g_{\lambda_1})}(y)\Big) \overset{(b)}{\ge} 0
\end{aligned}$$

where (a) follows from Lemma 2 and (b) holds by (14). Therefore the RBP is indexable.

C Proof of Proposition 2

We prove the result of each part separately.

a. This follows from observing that

$$\begin{aligned}
&\sum_{y\in\mathcal{X}} \Big\{ \big[\beta P_{zy}(1) - P_{xy}(1)\big]^+ N^{(g)}(y) - \big[P_{xy}(1) - \beta P_{zy}(1)\big]^+ N^{(h)}(y) \Big\} \\
&\quad \overset{(a)}{\le} \sum_{y\in\mathcal{X}} \big[\beta P_{zy}(1) - P_{xy}(1)\big]^+ N^{(g)}(y) \\
&\quad \overset{(b)}{\le} \sum_{y\in\mathcal{X}} \big[\beta P_{zy}(1) - P_{xy}(1)\big]^+ \le \max_{x,z\in\mathcal{X}} \sum_{y\in\mathcal{X}} \big[\beta P_{zy}(1) - P_{xy}(1)\big]^+
\end{aligned}$$

where we are ignoring negative terms in (a) and using $N^{(g)}(x) \le 1$ in (b).

b. For any $x, y, z \in \mathcal{X}$, $P_{xy}(1) - \beta P_{zy}(1) = (1-\beta) P_{xy}(1)$. Thus,

$$\begin{aligned}
&\sum_{y\in\mathcal{X}} \Big\{ \big[\beta P_{zy}(1) - P_{xy}(1)\big]^+ N^{(g)}(y) - \big[P_{xy}(1) - \beta P_{zy}(1)\big]^+ N^{(h)}(y) \Big\} \\
&\quad = - \sum_{y\in\mathcal{X}} (1-\beta) P_{xy}(1) N^{(h)}(y) \le 0 < \frac{(1-\beta)^2}{\beta}.
\end{aligned}$$

c. This follows from observing that

$$\begin{aligned}
&\sum_{y\in\mathcal{X}} \Big\{ \big[P_{xy}(0) - P_{xy}(1)\big]^+ N^{(g)}(y) - \big[P_{xy}(1) - P_{xy}(0)\big]^+ N^{(h)}(y) \Big\} \\
&\quad \overset{(a)}{\le} \sum_{y\in\mathcal{X}} \big[P_{xy}(0) - P_{xy}(1)\big]^+ N^{(g)}(y) \\
&\quad \overset{(b)}{\le} \sum_{y\in\mathcal{X}} \big[P_{xy}(0) - P_{xy}(1)\big]^+ \le \max_{x\in\mathcal{X}} \sum_{y\in\mathcal{X}} \big[P_{xy}(0) - P_{xy}(1)\big]^+
\end{aligned}$$

where we are ignoring negative terms in (a) and using $N^{(g)}(x) \le 1$ in (b).

d. $\beta \le 0.5$ implies that

$$\frac{1-\beta}{\beta} \ge 1 \ge \max_{x\in\mathcal{X}} \big[P_{xy}(0) - P_{xy}(1)\big]^+$$

which is the same as sufficient condition (c) established above.

References

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

[2] Nima Akbarzadeh and Aditya Mahajan. Restless bandits with controlled restarts. https://codeocean.com/capsule/8680851/tree/v1.

[3] 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.

[4] Nima Akbarzadeh and Aditya Mahajan. Restless bandits with controlled restarts: Indexability and computation of whittle index. In Conf. Decision Control, pages 7294–7300, 2019.

[5] 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. Math. Operat. Res., 57(1):21–39, 2003.

[6] Thomas W. Archibald, D. P. Black, and Kevin D. Glazebrook. Indexability and index heuristics for a simple class of inventory routing problems. Operat. Res., 57(2):314–326, 2009.

[7] Konstantin Avrachenkov, Urtzi Ayesta, Josu Doncel, and Peter Jacko. Congestion control of TCP flows in internet routers by means of index policy. Computer Networks, 57(17):3463–3478, 2013.

[8] Urtzi Ayesta, Martin Erausquin, and Peter Jacko. A modeling framework for optimizing the flow-level scheduling with time-varying channels. Performance Evaluation, 67(11):1014–1029, 2010.

[9] Biswarup Bhattacharya. Restless bandits visiting villages: A preliminary study on distributing public health services. In ACM SIGCAS Conf. on Computing and Sustainable Societies, pages 1–8. ACM, 2018.

[10] Jhelum Chakravorty and Aditya Mahajan. Multi-armed bandits, gittins index, and its calculation. Methods and applications of statistics in clinical trials: Planning, analysis, and inferential methods, 2:416–435, 2014.

[11] 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. Operat. Res., 61(6):1277–1294, 2013.

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

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

[14] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride. Some indexable families of restless bandit problems. Adv. Appl. Prob., 38(3):643–672, 2006.

[15] KD Glazebrook, DJ Hodge, and Christopher Kirkbride. Monotone policies and indexability for bidirectional restless bandits. Adv. Appl. Prob., 45(1):51–85, 2013.

[16] KD Glazebrook and HM Mitchell. An index policy for a stochastic scheduling model with improving/deteriorating jobs. Naval Research Logistics, 49(7):706–721, 2002.

[17] Kevin D Glazebrook, Christopher Kirkbride, and Jamal Ouenniche. Index policies for the admission control and routing of impatient customers to heterogeneous service stations. Operat. Res., 57(4):975–989, 2009.

[18] 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. Euro. J. Operat. Res., 165(1):267–284, 2005.

[19] Peter Jacko. Optimal index rules for single resource allocation to stochastic dynamic competitors. In Conf. Perf. Eval. Method. Tools. ACM, 2012.

[20] Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. 56(11):5547–5567, 2010.

[21] Jose Niño-Mora. Restless bandits, partial conservation laws and indexability. Adv. Appl. Prob., 33(1):76–98, 2001.

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

[23] José Niño-Mora. Restless bandit marginal productivity indices, diminishing returns, and optimal control of make-to-order/make-to-stock M/G/1 queues. Mathematics of Operations Research, 31(1):50–84, 2006.

[24] Christos H Papadimitriou and John N Tsitsiklis. The complexity of optimal queuing network control. Math. Operat. Res., 24(2):293–305, 1999.

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

[26] 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.

[27] J. Wang, X. Ren, Y. Mo, and L. Shi. Whittle index policy for dynamic multichannel allocation in remote state estimation. IEEE Transactions on Automatic Control, 65(2):591–603, 2020.

[28] Richard R Weber and Gideon Weiss. On an index policy for restless bandits. J. Appl. Prob., 27(3):637–648, 1990.

[29] Peter Whittle. Restless bandits: Activity allocation in a changing world. J. Appl. Prob., 25(A):287–298, 1988.

[30] Zhe Yu, Yunjian Xu, and Lang Tong. Deadline scheduling as restless bandits. 63(8):2343–2358, 2018.