Probability in the Engineering and Informational Sciences, **30**, 2016, 1–23.
doi:10.1017/S026996481500025X

INDEXABILITY AND OPTIMAL INDEX POLICIES FOR A CLASS OF REINITIALISING RESTLESS BANDITS

Sofía S. Villar
BCAM, Basque Center for Applied Mathematics and MRC Biostatistics Unit, IPH, University Forvie Site, Robinson Way, Cambridge CB2 0SR, UK and Lancaster University
E-mail: sofia.villar@mrc-bsu.cam.ac.uk

Motivated by a class of Partially Observable Markov Decision Processes with application in surveillance systems in which a set of imperfectly observed state processes is to be inferred from a subset of available observations through a Bayesian approach, we formulate and analyze a special family of multi-armed restless bandit problems. We consider the problem of finding an optimal policy for observing the processes that maximizes the total expected net rewards over an infinite time horizon subject to the resource availability. From the Lagrangian relaxation of the original problem, an index policy can be derived, as long as the existence of the Whittle index is ensured. We demonstrate that such a class of *reinitializing* bandits in which the projects' state deteriorates while active and resets to its initial state when passive until its completion possesses the structural property of indexability and we further show how to compute the index in closed form. In general, the Whittle index rule for restless bandit problems does not achieve optimality. However, we show that the proposed Whittle index rule is optimal for the problem under study in the case of stochastically heterogenous arms under the expected total criterion, and it is further recovered by a simple tractable rule referred to as the *1-limited Round Robin* rule. Moreover, we illustrate the significant suboptimality of other widely used heuristic: the Myopic index rule, by computing in closed form its suboptimality gap. We present numerical studies which illustrate for the more general instances the performance advantages of the Whittle index rule over other simple heuristics.

1. INTRODUCTION

Modern sensing technologies offer the possibility of efficiently performing tasks by adaptively deploying its sensing resources based on the information extracted from past measurements. Yet, realizing such system's overall performance gains requires appropriate on-line sensing rules. Thus, the general problem in sensor management is to design sensing algorithms that allow for the fruitful adoption of cutting edge technologies. A natural procedure to derive those rules is to represent the underlying resource allocation problem by some stochastic dynamic optimization model, whose optimal solution is traditionally characterized by a dynamic programming (DP) framework. However, those formulations, at least for realistic scenarios, typically have a prohibitively large size (possibly infinite), which dramatically hinders its practical application. Thus, fully exploiting the performance advantages offered by the new technologies by means of active dynamic sensing policies remains very challenging. For this reason, the design of both computationally feasible and nearly optimal sensing

 0269-9648/15 $25.00

strategies, as the ones proposed in this paper, continues to be a highly active applied research area.

An additional challenge to the design of adequate on-line active sensing schemes is to take into account specific situations that may affect the system's performance. For detection objectives, there have been significant efforts to deal with more general situations, for example with multiple objects, or with mobile objects, and even to include false targets. Yet, despite this abundant literature, the case in which targets react to sensing or may evade the searcher, remains understudied today. This paper addresses these two challenges by proposing a tractable scheduling rule for a multiple target detection problem in which targets react to sensing by remaining frozen in their current state and sensing is subject to misdetection errors. We formulate this detection problem, as a partially observable Markov decision processes (POMDP) with special structure, which further fits into the framework of the continuous state multi-armed restless bandit problem (MARBP).

The MARBP constitutes a theoretical framework under which resource allocation problems under uncertainty can be fruitfully analyzed. In its general version, the MARBP consists of choosing a subset of arms to activate at each period of time (out of a possibly larger set of arms), where the state of each arm evolves randomly over time, affecting their resulting flow of rewards (and/or costs). A natural goal for this problem is to choose the arms to activate so as to achieve the maximum expected total discounted or time-average rewards over an infinite time horizon.

In the so-called *classic* Bayesian Bernoulli version of the problem, whose origin dates back to the Second World War, arms' states evolve stochastically only when chosen, yielding a binary random reward. Such a variant, despite its simplified dynamics, was regarded unsolvable until Gittins and Jones [2,3] showed that its optimal solution admitted a simple expression in terms of an index function attached to each arm depending on its current state. The resulting optimal index rule activates at each time period the arm whose current index value is the maximum. More than a decade later, Whittle [21] proposed and studied the more general *restless* case in which non-chosen arms continue to evolve, and pointed out that neither the existence of the index function extending the classic case nor the optimality of the resulting index rule was guaranteed for such variant.

This *indexability* property, that is, the existence of an index function, introduced by Whittle for MARBP problems, cannot be taken for granted as it needs to be established for each specific model. Niño-Mora [14,15] provided the first general sufficient indexability conditions based on the achievable region approach to stochastic optimization which can be systematically deployed under certain conditions. Furthermore, the indexability of special classes of MARBP has been specifically addressed and thoroughly studied using various approaches. These include some families of *restless* bandits which arise in machine maintenance and stochastic scheduling problems with switching costs, as those in Glazebrook, Ruiz-Hernandez, and Kirkbride [4], the *bidirectional* bandits introduced in Glazebrook, Hodge, and Kirkbride [5], the *reinitializing* bandits in Jacko and Sanso [7], and restless models in telecommunication and opportunistic spectrum access as in Liu and Zhao [10], among others. These papers are part of the body of literature that has contributed to a significant advance in the understanding of this property, yet as Liu, Weber, and Zhao [11] put it "[...] *establishing indexability is still an open problem and often relies on numerical algorithms* ". Moreover, even when indexability is ensured, index computation usually poses further significant challenges [15].

In this paper, we establish the indexability of a class of MARBP that derives from a concrete family of POMDPs and it is motivated by a surveillance systems application. POMDPs admit a widespread range of applications, for example in navigation problems, artificial intelligence, sensor systems, machine maintenance, telecommunication networks,

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

among others but their optimal policy is often computationally intractable. Therefore, the most commonly used solution methods seek to find *good* approximate solutions based on some discretization or reduction of its infinite state space (see, e.g., [12,18]). Still, the high computational cost of solving POMDPs is the main cause limiting their practical implementation. The resulting bandit formulation is one in which arms (until reaching a terminal state), generate a decreasing stream of random rewards when chosen to be active and passive arms continue to change state (even if not chosen), although they do so according to a simple transition rule: returning to its initial state (i.e., to its state at time 0 when the controller starts operating). The class of problems is introduced in Section 2, and following Jacko [6] we shall refer to them as *reinitializing* bandits.

These *reinitializing* bandits have some common features with models previously addressed: it is similar to the reward depletion and replenishment model presented in [4], and it also shares with *bidirectional* bandits in [5], the property that the active and passive actions produce opposite movements on the state space. Another related application is found in [7], where a new type of congestion control scheduling method based on a MARBP is proposed, motivated by the Internet flows behaving according to the Transmission Control Protocol, and thus admitting a reinitializing feature. In [10], a similar problem with applications in opportunistic spectrum access is considered. The problem is formulated as a MARBP and, using a similar approach to the one deployed in this paper based on DP, indexability is established and the Whittle index in closed form is derived. Later in [11], following the same rationale, the authors studied the case in which the active action resets the state is considered and solved though a Whittle index policy. Both [10,11] share with the model presented in this paper the feature of having a continuous state space. In [11], as well as in this work a property of the problem is exploited to reduce the state space from a continuous one to a numerable one.

Despite these similarities, those models and the one addressed in this paper differ in two main aspects. First, the inclusion of an absorbing state in the model is a distinctive feature of the problem addressed in this paper that has not been considered in the previous works. Second, the introduction of imperfect observability of the state instead of the perfect observability (when sensing) assumption included in the models in [10,11], makes the resulting MARBP more realistic and challenging. Another novel contribution of this work is the introduction of a target's reacting to the sensing actions, along the lines of what was done in [9,19].

1.1. Main Contributions and Paper Structure

We start in Section 2 by describing the problem, stating the model's assumptions and formulating it as a MARBP. In Section 3, we demonstrate the existence of the index for this class of problems by establishing the monotonicity in an activity charge λ of activation policies using properties of the corresponding DP formulation. Once indexability is established, a closed-form formula for the Whittle index is derived which, despite the problem's simplified dynamics, it is far from being trivially deduced. Moreover, the importance of the indexability result is increased by the fact that the resulting Whittle index can be used as a well performing approximate solution method for a special family of POMDPs.

In Section 4, we proceed to study the properties of the proposed Whittle index rule, with a special emphasis on its relative performance when compared with other commonly used naive heuristics. Weber and Weiss [20] showed the asymptotical optimality of Whittle's heuristics under certain conditions and for a limiting case in which the number of arms in total and the number of arms to be activated go to infinity in a constant ratio. We prove the optimality of the Whittle Index rule for the proposed MARBP in the special case of a

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

finite number of heterogenous arms and when considering the Expected Total criterion (i.e., when the discount factor is 1). Moreover, in the case in which all arms can be activated at each time slot, we are able to give a closed-form expression for the suboptimality gap of other naive yet widely used index rules (e.g., the Myopic Index rule), a result that is in stark contrast to the equivalence between these rules reported for other similar models; see, for example [1,10,13,22].

To conclude the paper, in Section 5 we use numerical studies to illustrate how the theoretical results of the paper are deployed, we analyze also the performance of the alternative heuristics revised in Section 4 for this class of problems, and we show how well the Whittle index rule performs even in those cases in which the optimality of the rule is not analytically established.

2. PROBLEM DESCRIPTION AND MARBP FORMULATION

Consider the following problem in surveillance systems. There are N-independent locations (or sites), each containing one target (or object) hidden in it. There are M $(1 \leq M \leq N)$ sensors, each of which at every discrete period can search at most one of those locations. All sensors in the system are synchronized to operate over time slots $t = 0, 1, \ldots$, where a time slot corresponds to a Pulse Repetition Interval. Each target can be in one of two possible *visibility* states: a *hidden* (or bad) state, in which it is completely invisible to sensors but cannot perform its tasks, and an *exposed* (or good) state, in which it can perform its tasks but is prone to being detected by sensors. Targets are such that: (1) they perceive if they are being sensed; (2) they do not wish to be found, but they wish to perform their tasks; (3) while being sensed they do not change their status, that is, they stay frozen at their current state; and (4) while not being sensed they alternate between the two states randomly. Thus, if a target is in site n and is not sensed in period t, it becomes *exposed* in period $t+1$ with probability ϕ_n^0 and it becomes hidden with probability $(1-\phi_n^0)$, regardless of its initial state. The probability that a sensor searching for a target at site n finds it when it is *visible* is $0 < 1-\alpha_n < 1$, and hence the probability that an unfound target is visible at slot t changes by Bayes' theorem as the sensor's detection output is observed. The cost of a single search of a location n is $c_n \geq 0$ and yields a reward $r_n\beta^t$ when it succeeds at finding target n in slot t, where $0 \leq \beta \leq 1$ is a discount factor.

The goal is to design a tractable policy which addresses the following question: *How should the N locations be scheduled for being sensed so as to be close to maximizing the total expected discounted reward of finding all targets, using M or less sensors at each time slot?*

We therefore next formulate and investigate the following MARBP problem. The N targets are represented by N-independent projects (or arms) labeled by $n \in \mathcal{N} \triangleq \{1, \ldots, N\}$, each yielding a positive reward if completed. Sensing decisions are thus formulated by binary *action* processes $a_{n,t} \in \{0, 1\}$, where $a_{n,t} = 1$ represents that target n is sensed in slot t and $a_{n,t} = 0$, otherwise. Every unfound target is an incomplete project that can be in two states: (a "good/exposed" state: 1, a "bad/hidden" state: 0). It is assumed that projects can only be completed while being active on them and if they are in the "good" state. If a project is rested then its state changes randomly, while if a project is active its state can only evolve from the "good" state to a completed state T or it simply stays in its current state. Such a dependence of the transitions probabilities on the selected action is described in Figure 1.

The state of a project n defines a random process $S_{t,n}$ which is only partially observable to the decision maker in the following sense: only under the active action at time t in project n (i.e., $a_{n,t} = 1$) an imperfect measurement of its current state $o_{n,t}$ is available.

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

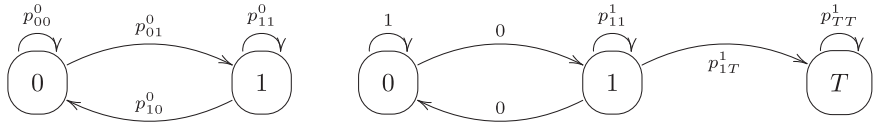

FIGURE **1.** The Markov chain associated to a generic project given each possible action, active: 1 and passive: 0. The arrows represent one-period transitions among the states 0 (bad) and 1 (good) with given probabilities under actions 0 (on the left) and 1 (on the right).

Furthermore, whenever project n is at state 0 and it is activated, the measurement of its state will coincide with its true state (i.e., $o_{n,t} = s_{n,t} = 0$); however, if process n is at state 1 at time t and it is activated, its measurement is correct (i.e., $o_{n,t} = s_{n,t} = 1$) only with probability $(1 - \alpha_n)$. Hence, the observation process is subject to *misdetection* errors with probability $\alpha_n \triangleq P(o_{n,t} = 0|s_{n,t} = 1)$, where $0 < \alpha_n < 1$. Notice that we exclude the extreme cases $\alpha = 0$ or $\alpha = 1$, respectively corresponding to *complete observability* of the process, because we are interested in the non-trivial problem of how to use the partial information given by the observable process $o_{n,t}$ to gain information on the partially observable process $S_{n,t}$.

The resulting model is thus a POMDP in which the N states of the projects are observable only through the active action, and they are imperfectly observed in the particular sense that the "good" state can be mistaken for the "bad" state with a given misdetection probability α_n. We shall further consider that project n is completed as soon as a "good" state is observed, that is, when $o_{n,t} = 1$. In terms of the transition probabilities in Figure 1, we have that $p^1_{11} = 1 - P(o_t = 1|s_{n,t} = 1) = \alpha_n$, $p^1_{1T} = P(o_t = 1|s_{n,t} = 1) = (1 - \alpha_n)$. And to introduce the reinitializing feature, we assume that under the passive action the project's state resets the probability of being in the "good" state to its value at time 0, which we denote $0 < \phi_0 < 1$, regardless its current state, that is, $p^0_{11} + p^0_{01} = \phi^0_n$ and $p^0_{10} + p^0_{00} = (1 - \phi^0_n)$.

We further consider that at every discrete time slot over an infinite horizon $t = 0, 1, \ldots$, at most M processes can be selected for activation/observation, with $1 \leq M \leq N$, incurring in an observation cost per activated process denoted by $c_n \geq 0$ and yielding a final reward per job completion $r_n > 0$ if a process n is observed to be at state 1, that is, $o_{n,t} = 1$, yielding no more rewards thereafter. Hence, at each time slot we must decide on which processes to observe so as to maximize the total expected rewards.

Observation decisions are thus formulated by binary *action* processes $a_{n,t} \in \{0, 1\}$, where $a_{n,t} = 1$ represents that process n is observed in slot t and $a_{n,t} = 0$, otherwise. Any feasible *observation scheduling rule* which prescribes how to sequentially observe processes over time, will be denoted as $\boldsymbol{\pi}$, and belongs to the class $\boldsymbol{\Pi}(M)$ of *admissible policies* composed by the *non-anticipative* scheduling polices (i.e., those based on the history of states and actions) which observe M or less processes per slot, that is, $\sum_{n \in \mathsf{N}} a_{n,t} \leq M$, $t = 0, 1, 2, \ldots$.

At $t = 0$ process n has probability ϕ^0_n of being in state $s_{n,0} = 1$, since we assume that at time $t = 0$ no process was previously activated. Thereafter the posterior probability that some process n is in state 1, denoted as $p_{n,t} \in \mathcal{P} \triangleq [0, \phi^0_n]$ (henceforth referred to as its *belief state*), must be computed conditioning on past observations and also on the selected actions via Bayes' rule. Notice that this posterior probability is a sufficient statistic of each project's state. Even though each project has three possible states: exposed (1), hidden (0) or completed (T), the last two states yield no reward and the state of completion T is perfectly observed (as the reward is then collected). Therefore, the only information gained from the measurements is about the exposed (1) and hidden states (2).

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

Finally, given all the elements of the model, we denote by $R_n(p_{n,t}, a_{n,t}) \triangleq (r_n\,(1-\alpha_n)\,p_{n,t} - c_n)\,a_{n,t}$ the one-slot expected net reward function earned when taking action $a_{n,t}$ at time slot t on process n when its probability state is $p_{n,t}$.

We shall consider that the objective of the controller is to design a policy that sequentially selects at most M out of N processes to observe at each time slot so as to maximize the total expected discounted reward over an infinite horizon, given a discount factor $0 \leq \beta \leq 1$. Such an objective can be addressed by considering the following dynamic optimization problem: find an expected β-discounted optimal policy such that:

$$V_\beta^*(\boldsymbol{\phi}^0) \triangleq \max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}(\boldsymbol{M})} \mathsf{E}_{\boldsymbol{\phi}^0}^{\boldsymbol{\pi}} \left[\sum_{t=0}^{\infty}\sum_{n\in\mathsf{N}} \beta^t R_n(p_{n,t}, a_{n,t})\right], \tag{1}$$

where $\boldsymbol{\phi}^0 = (\phi_n^0) = \mathbf{p}_0$ is the initial joint belief state, and $\mathsf{E}_{\boldsymbol{\phi}^0}^{\boldsymbol{\pi}}[\cdot]$ denotes expectation under policy $\boldsymbol{\pi}$ conditional on $\mathbf{p}_0 = \boldsymbol{\phi}^0$. Further, as it will be discussed later, (1) is bounded by a finite constant, thus the problem of finding an expected total-optimal policy is well defined for this model and thus it will be considered by letting $\beta = 1$.

The optimal scheduling problem posed by (1) describes a constrained POMDP consisting of optimally deciding which processes to observe to maximize rewards given the resource constraint and based on the current estimate of the belief state of all processes. The approach followed in this paper to address the high computational cost of optimally solving POMDPs is to exploit the fact that the POMDP in (1) can be analyzed as a MARBP with a continuous state variable $\mathbf{p} = (p_n)$ and a reinitializing feature. Thus, each process n constitutes an independent single-bandit model, with two possible actions: "observe" and "not observe" and whose state is given by its belief state, that is, the probability $p_{n,t}$ of being in state 1 at time t. The optimal solution for MARBPs is also generally intractable (see, e.g., [17]), yet we shall follow the solution approach for MARBPs based on a Lagrangian approach, first proposed by Whittle [21], which often results in nearly optimal and tractable solution.

Since for the model at hand, processes' state transitions are independent, the stochastic evolution of each arm depends only on the decisions taken for it and on its own specific parameters. Hence, each arm is a single-bandit problem which can be individually considered for establishing index existence and also for index computation. Therefore, in what follows we first describe the elements of the single-bandit problem modeling the optimal observation decisions of an individual process, and next, based on these elements, we define the indexability property that is required to hold in order to derive the Whittle Index rule for the original problem (1). Thus, the following discussion focuses on a generic single-bandit problem, and henceforth its label n is dropped to simplify the notation.

2.1. Single-Bandit POMDP: Definition

Each of the N processes is represented by a single-bandit problem, which can be defined by its composing elements as follows.

- **Action Space**: a binary action set $a_t \in \{0, 1\}$, where $a_t = 1$ represents that the process is observed in slot t and 0, otherwise;
- **State Space**: a continuous state space, denoted by $\mathcal{P} \triangleq [0, \phi^0]$ containing all the possible belief states of the random process. The final state of the process T, which is an absorbing state that after being reached is never abandoned, causes the process to yield no rewards nor costs thereafter. Given that the only case in which we are certain that a project will not yield a reward (by being in the exposed state) is when it is completed (i.e., it is in state T), we shall assume that projects only have a zero

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

probability of being in the good state if they have been successfully completed. Thus the terminal state can be conveniently represented by the belief state 0.

Since at the state T there is no actual decision to take, we therefore adopt the convention that at $p = 0$ the selected action is $a = 0$, yielding no rewards and producing no transitions. For the rest of the belief states, there is a decision to be taken by the controller, and thus we shall refer to that set of states as the *controllable* set of states, that is, $p \in \mathcal{P} \setminus \{0\}$.

- **State Dynamics**: a transition rule that specifies how the state evolves stochastically over $\mathcal{P}$ in time depending on the selected action a and on the current state p. If $p = 0$, it stays at this state thereafter. For any $p \in \mathcal{P} \setminus \{0\}$, the belief state process p_t evolves according to the dynamics below:

$$\begin{array}{rlll} \text{For } p_t \in \mathcal{P} \setminus \{0\} & & & \\ & \text{if } a_t = 0 \quad p_{t+1} = \phi^0 & \text{w.p.} & 1; \\ & \text{if } a_t = 1 \quad p_{t+1} = \frac{\alpha p_t}{1-(1-\alpha)p_t} & \text{w.p.} & 1-(1-\alpha)p_t; \\ & \qquad\qquad\quad p_{t+1} = 0 & \text{w.p.} & (1-\alpha)p_t. \\ \text{For } p_t = 0 & & & \\ & \text{if } a_t = 0 \quad p_{t+1} = 0 & \text{w.p.} & 1 \end{array} \tag{2}$$

- **Rewards**: a one-period reward function given by $R(p_t, a_t) \triangleq (r(1-\alpha)p_t - c)\, a_t$ for $p_t \in \mathcal{P} \setminus \{0\}$ and $R(0,0) \triangleq 0$;
- **Costs**: a fixed parameter, $\lambda \in \mathbb{R}$, introduced to represent an extra *activity-charge* that must be paid by the controller whenever $a_t = 1$.

Thus, the infinite horizon single-bandit problem is formulated as

$$V^*_\beta(\phi^0, \lambda) \triangleq \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} \mathsf{E}^{\boldsymbol{\pi}}_{\phi^0} \sum_{t=0}^{\infty} \beta^t \left[R(p_t, a_t) - \lambda a_t \right], \tag{3}$$

We shall further denote the optimal *active set* for (3) as a function of the parameter λ by $A^*(\lambda)$. Hence, for some $p \in \mathcal{P} \setminus \{0\}$, $p \in A^*(\lambda)$ if and only if $p_t = p$, then the optimal action $a^*_t = 1$.

Let us now introduce the definition of indexability.

DEFINITION 2.1: For any value of parameter $\lambda \in \mathbb{R}$, subproblem (3) *is indexable if its optimal active set* $A^*(\lambda)$ *decreases monotonically from* $\mathcal{P} \setminus \{0\}$ *to* $\emptyset$ *as* λ *goes from* $-\infty$ *to* ∞.

Whittle's original indexability definition was formulated in terms of optimal passive sets, letting the multiplier λ be a subsidy for passivity. In Definition 2.1, the parameter λ as can be interpreted as a tax for activity. Such a definition ensures the existence of critical values of the multiplier λ which induce a *nesting* ordering of the optimal active sets as a function of λ.

3. WHITTLE INDEX EXISTENCE AND INDEX CHARACTERIZATION

In this section, we shall establish the validity of the following theorem, which ensures that indexability holds for the model at hand.

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

THEOREM 3.1: *The single-bandit problem* (3) *is indexable according to Definition 2.1.*

DP Analysis and Proof of Theorem 3.1

As in [10,11], we shall prove 3.1 following a DP approach.

First, we define $b(p) \triangleq 1-(1-\alpha)p$, which denotes the probability that the process in belief state p has not reached its final state after one active time period. Further, we shall use the notation $\phi^1(p) \triangleq \alpha p/1-(1-\alpha)p$ to represent the posterior probability of a misdetection when the process is observed in a belief state p.

Hence, for some fixed parameter λ, the DP equation for the β-discounted problem (3) is written as

$$V_\beta^*(\phi^0,\lambda) = \max\{R(\phi^0,1)-\lambda+b(\phi^0)\beta V_\beta^*(\phi^1(\phi^0),\lambda)\,;\,\beta V_\beta^*(\phi^0,\lambda)\} \tag{4}$$

where we have used the fact that when the final state of the process (represented by $p=0$) is reached, the process yields no rewards nor costs and the selected action by default is $a_t=0$, by setting $V_\beta^*(0,\lambda)=0$ for all possible β and λ.

Next, we write the DP equations for each possible partition of the set $\mathbb{P}\setminus\{0\}$, that is, in terms of the optimal active set $A^*(\lambda)$ for a fixed λ, as follows:

$$V_\beta^*(p,\lambda) = R(p,1)-\lambda+b(p)\beta V_\beta^*(\phi^1(p),\lambda) \qquad \text{for } p\in A^*(\lambda), \tag{5}$$

$$V_\beta^*(p,\lambda) = \beta V_\beta^*(\phi^0,\lambda) \qquad \text{for } p\notin A^*(\lambda). \tag{6}$$

The proof of Theorem 3.1 is based on the following property of the optimal value function.

LEMMA 3.2: *Function* $V_\beta^*(\phi^0,\lambda)$ *is nonnegative, piecewise-linear in* λ *and non-increasing in* λ.

PROOF: We shall use the fact that the evolution of the state variable after t (unsuccessful) active slots starting from an initial belief state p generates an iterated mapping $p\mapsto\phi^1(p)$, that is, $\phi_0^1(p)\triangleq p$ and $\phi_t^1(p)\triangleq\phi^1(\phi_{t-1}^1(p))$ for $t\geq 1$. Such a mapping represents the Bayesian update of the belief state and it is decreasing in t, since for all p and α it holds that $\phi^1(p)<p$, and further, it defines a *Möbius Transformation*. Using the matrix form of such non-linear functions, we can derive by induction a closed-form expression for the tth iterate of $\phi^1(p)$ to be

$$\phi_t^1(p) = \left[1-\left(1-\frac{1}{p}\right)\left(\frac{1}{\alpha}\right)^t\right]^{-1}. \tag{7}$$

Note that from (7) it can also be shown that $\phi_t^1(p)$ is a decreasing function in t. (See [23] for a detailed description on how to derive such closed-form expressions.)

Thus, once the process leaves state ϕ^0, as long as it does not reach its final state, it only returns to ϕ^0 after reinitializing the process, that is, after being passive for a time slot. Hence, for any $p\in A^*(\lambda)$, we denote by $t^*(p,\lambda)$ the number of (unsuccessful) active time slots that, starting from a belief state p, may elapse until it is optimal to be passive, and

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

we rewrite (5) as a function of $t^*(p,\lambda)$ as follows:

$$V_\beta^*(p,\lambda) = \sum_{t=0}^{t^*(p,\lambda)-1} \beta^t \left(R(\phi_t^1(p),1)-\lambda\right) \prod_{s=0}^{t-1} b(\phi_s^1(p)) + \beta^{t^*(p,\lambda)+1} \prod_{s=0}^{t^*(p,\lambda)-1} b\left(\phi_s^1(p)\right) V_\beta^*(\phi^0,\lambda). \tag{8}$$

The first term in expression (8) is the optimal expected discounted reward generated during the $t^*(p,\lambda)$ active time slots and the second term represents the expected discounted optimal value function starting at the reinitializing state in $t^*(p,\lambda)+1$, that is, after allowing for one passive slot.

The decreasing feature of the active dynamics (7) implies that starting from some state p in the optimal active set, the original optimization problem can be analyzed as an optimal stopping problem. From some p in the active set, the system may visit states $\phi_t^1(p)$ for $t=1,2\ldots$ (by repeatedly selecting the active action starting from p) or state 0 (if it reaches its final state), but the states different from ϕ^0 can only be visited in a predetermined (decreasing) order, as $\phi_t^1(p)$. Exploiting this *deteriorating* feature, the solution to problem (3) can be equivalently described in terms of optimal active sets or in terms of the optimal number of active time slots starting from a given state.

Also, notice that if $\phi^0 \notin A^*(\lambda)$, it follows from (6) that $V_\beta^*(\phi^0,\lambda)=0$, and further given the problem's dynamics, the optimal belief state trajectory remains constant at ϕ^0, thus never activating the process, that is, $A^*(\lambda)=\emptyset$ and hence, $t^*(\phi^0,\lambda)=0$. Thus, the nontrivial case to consider corresponds to all possible λ such that $\phi^0 \in A^*(\lambda)$, that is, where $t^*(\phi^0,\lambda)>0$.

Next, we will invoke the auxiliary results in Lemma 3.3 (whose proof is deferred to the Appendix) to simplify expression (8).

LEMMA 3.3: *We have*

$$(a)\quad \prod_{s=0}^{t-1}\left[1-(1-\alpha)\phi_s^1(p)\right] = 1-(1-\alpha^t)p; \quad (b)\quad \phi_t^1(p)\prod_{s=0}^{t-1}\left[1-(1-\alpha)\phi_s^1(p)\right] = \alpha^t p.$$

For any λ such that $\phi^0 \in A^*(\lambda)$, setting $p=\phi^0$ in (8) and using Lemma 3.3, $V_\beta^*(\phi^0,\lambda)$ can be computed in closed-form as a function of $t^*(\phi^0,\lambda)$ with the expression below:

$$\frac{\left(\frac{1-(\alpha\beta)^{t^*(\phi^0,\lambda)}}{1-(\alpha\beta)}\right)\left[r(1-\alpha)\phi^0\right]-(\lambda+c)\left(\phi^0\frac{1-(\alpha\beta)^{t^*(\phi^0,\lambda)}}{1-(\alpha\beta)}+\frac{1-\beta^{t^*(\phi^0,\lambda)}}{1-\beta}(1-\phi^0)\right)}{1-\beta^{t^*(\phi^0,\lambda)+1}(1-(1-\alpha^{t^*(\phi^0,\lambda)})\phi^0)}. \tag{9}$$

Denote by $V_\beta(\phi^0,\lambda,i)$ the expression (9) evaluated by setting $t^*(\phi^0,\lambda)=i$. Notice that, solving problem (4), that is, finding the states that belong to $A^*(\lambda)$, is therefore equivalent to finding the maximum positive integer i such that it holds: $\phi_i^1(\phi^0)\in A^*(\lambda)$. Thus, if $t^*(\phi^0,\lambda)=i$ it must be the case that $\phi_t^1(\phi^0)\in A^*(\lambda)$ only if $t\le i$.

Thus, substituting $\phi_i^1(\phi^0)$ for p in Eqs (5) and (6), and given that $V_\beta^*(\phi^0,\lambda)=V_\beta(\phi^0,\lambda,i)$, using (4) we have that:

$$R(\phi_i^1(\phi^0),1)-\lambda+\beta^2 b(\phi_i^1(\phi^0))V_\beta(\phi^0,\lambda,i) > \beta V_\beta(\phi^0,\lambda,i), \tag{10}$$

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

where we have also assumed that in the case of the maximum being achieved by both actions, the selected action by default is the passive action. Further, $V_\beta(\phi^0, \lambda, i)$ is computed for $t^*(\phi^0, \lambda) = i$ using expression (9). Thus, rearranging (10) we have that

$$V_\beta(\phi^0, \lambda, i) - \frac{R(\phi_i^1(\phi^0), 1) - \lambda}{\beta[1 - \beta^2 b(\phi_i^1(\phi^0))]} < 0. \tag{11}$$

Therefore, for some given λ such that $\phi^0 \in A^*(\lambda)$, $t^*(\phi^0, \lambda)$ is determined as the maximum non-negative integer i such that (11) holds. Furthermore, for $t^*(\phi^0, \lambda) = i$, where i is some positive integer, it must be the case that:

$$V_\beta(\phi^0, \lambda, i) > V_\beta(\phi^0, \lambda, j) > 0, \quad j = 1, \ldots, (i-1). \tag{12}$$

The first relation in (12) is a consequence of the fact that $t^*(\phi^0, \lambda) = i$, while the second relation in (12), that is, $V_\beta^*(\phi^0, \lambda, j) > 0$ follows from the fact that $V_\beta^*(\phi^0, \lambda, 0) = 0$. Notice also that if starting at ϕ^0 it is optimal to be active for i time slots, then it must be optimal to be active at any time slot between 0 and $i - 1$.

Consider now some $\lambda' < \lambda$. It follows from (9) that, for a fixed i, $V_\beta(\phi^0, \lambda, i)$ is a linear decreasing function of λ. Thus, when (12) holds for a given λ, it will hold also for $\lambda' < \lambda$. Therefore, the set of integers for which the optimality (12) holds is a non-increasing set with respect to λ. Therefore, $t^*(\phi^0, \lambda)$ is a non-increasing (piece-wise constant) function of λ and it further follows that $V_\beta^*(\phi^0, \lambda)$ as in (9) is a nonnegative and non-increasing piece-wise linear function in λ. ■

Next, we announce the following corollary, which is a direct consequence of Lemma 3.2.

COROLLARY 3.4: *If $p \in A^*(\lambda)$ for some $p \in \mathcal{P} \setminus \{0\}$, then it must be that $p \in A^*(\lambda')$ for $\lambda' < \lambda$.*

PROOF: The proof follows from the relation between $t^*(\phi^0, \lambda)$ and the optimal active set $A^*(\lambda)$. Suppose it is known that when the process is at state ϕ^0 it is optimal to take the active action (as long as the process does not yield its final reward) for the next i steps, that is, $t^*(\phi^0, \lambda) = i$. Then, it must be the case that the belief states in the sequence $\phi_t^1(\phi^0)$ for $t = 0, \ldots, (i-1)$ belong to the active set $A^*(\lambda)$ Further, given the non-decreasing property of $t^*(\phi^0, \lambda)$, the set $A^*(\lambda')$ for $\lambda' < \lambda$ must (at least) include the set composed by the sequence $\phi_t^1(\phi^0)$ for $t = 0, \ldots, (i-1)$.

Alternatively, after some algebraic manipulations, expression (12) can be written for fixed i and $s = 1, \ldots, (i-1)$ as a function of λ which is also linear and decreasing in λ. ■

PROOF: Finally, indexability of the single-process problem (3), as defined in Definition 2.1, follows from Corollary 3.4. ■

3.1. The Whittle Index

Based on the indexability result established by 3.1, we announce in (13) the Whittle index closed-form expression for the single-process problem (3).

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

THEOREM 3.5: *The Whittle index, denoted by* $\lambda^W(p)$*, for the single-bandit problem* (3) *and for* $p \in \mathcal{P} \setminus \{0\}$*, is computed as follows:*

$$\lambda^W(p) = \frac{r(1-\alpha)\left[(1-\widehat{\beta})p - \beta(1-\beta b(p))\frac{1-(\alpha\beta)^{d(\phi^0,p)}}{1-(\alpha\beta)}\phi^0\right]}{(1-\widehat{\beta}) - \beta(1-\beta b(p))\left[\frac{1-(\alpha\beta)^{d(\phi^0,p)}}{1-(\alpha\beta)}\phi^0 + \frac{1-\beta^{d(\phi^0,p)}}{1-\beta}(1-\phi^0)\right]} - c, \tag{13}$$

where $\widehat{\beta} \triangleq \beta^{d(\phi^0,p)+1}(1-(1-\alpha^{d(\phi^0,p)})\phi^0))$*, and* $d(\phi^0,p) \triangleq \left\lceil \frac{\log\left(\frac{(1-p)\,\phi^0}{(1-\phi^0)\,p}\right)}{\log\left(\frac{1}{\alpha}\right)} \right\rceil$. (14)

PROOF: Once indexability has been established, the Whittle index $\lambda^W(p)$ for some belief state $p \in \mathcal{P} \setminus \{0\}$ is computed as the value of the multiplier λ such that the active and passive action are indifferent when the process is at state p, that is, the value of $V^*_\beta(p, \lambda^W(p))$ as computed by Eqs (5) (i.e., with $a = 1$) or (6) (i.e., with $a = 0$) is the same. Further, given the properties of the active dynamics previously derived, it follows that if for $\lambda = \lambda^W(p)$ at p both actions are indifferent, then it must be true that for belief states larger than p it is optimal to be active (until either the process reaches it final state or it reaches a belief state below p) while for belief states smaller than p it is optimal to *reinitialize* the process. Moreover, for some $p \in \mathcal{P} \setminus \{0\}$ there will exist a strictly positive integer t such that $\phi_t(\phi^0) \leq p < \phi_{t-1}(\phi^0)$ and $V_\beta(\phi_{t-1}(\phi^0), \lambda^W(p))$ takes its maximum value setting $a = 1$ (i.e., $V^*_\beta(\phi_{t-1}(\phi^0), \lambda^W(p))$ and it is computed using (5)) while $V_\beta(\phi_t(\phi^0), \lambda^W(p))$ takes its maximum vale when $a = 0$ (i.e., $V^*_\beta(\phi_t(\phi^0), \lambda^W(p))$ as in (6)).

The above reasoning allows us to conclude that $t^*(\phi^0, \lambda^W(p))$ is exactly the number of active slots required to make the state go from the initial state ϕ^0 to a state at most equal to p, that is, $t^*(\phi^0, \lambda^W(p)) = d(\phi^0, p) \triangleq \{t \geq 1 : \phi_t(\phi^0) \leq p < \phi_{t-1}(\phi^0)\}$. Notice that $\lambda^W(p)$ can be interpreted as the value of λ such that the optimal maximum number of (unsuccessful) periods that a process starting at the initial state must be activated before advising to reinitialize the process is exactly $t^*(\phi^0, \lambda^W(p))$. From expression (7) and given its definition, $d(\phi^0, p)$ can be computed in closed-form using the corresponding expression (14).

Therefore from Eqs (5) and (6), using the fact that for $\lambda = \lambda^W(p)$ it holds that $t^*(\phi^0, \lambda^W(p)) = d(\phi^0, p)$, we write the DP equations for some $\phi^1_{d(\phi^0,p)}(\phi^0) \leq p < \phi^1_{d(\phi^0,p)-1}(\phi^0)$ as:

$$V^*_\beta(p, \lambda^W(p)) = R(p,1) - \lambda^W(p) + b(p)\beta^2 V^*_\beta(\phi^0, \lambda^W(p)), \qquad \text{if } p \in A^*(\lambda^W(p)), \tag{15}$$

$$V^*_\beta(p, \lambda^W(p)) = \qquad \beta V^*_\beta(\phi^0, \lambda^W(p)), \qquad \text{if } p \notin A^*(\lambda^W(p)). \tag{16}$$

The critical value $\lambda^W(p)$ is such that (15) equals (16). Thus,

$$\lambda^W(p) = R(p,1) - \beta[1-\beta b(p)]V^*_\beta(\phi^0, \lambda^W(p)). \tag{17}$$

Next, we compute $V^*_\beta(\phi^0, \lambda^W(p))$ using expression (9) setting $t^*(\phi^0, \lambda^W(p)) = d(\phi^0, p)$ as in (14) and substitute it in (17). After tedious yet straightforward algebraic manipulations expression (13) is obtained. ■

COROLLARY 3.6: *The Whittle Index defined in* (13) *for the single-bandit problem* (3) *is a continuous and monotone increasing function in* p*, for any* $p \in \mathcal{P} \setminus \{0\}$*.*

PROOF: Both properties can be shown through algebraic manipulations of expression (13). The function $d(\phi_0, p)$ is a piecewise constant(left continuous)function. In particular, it

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

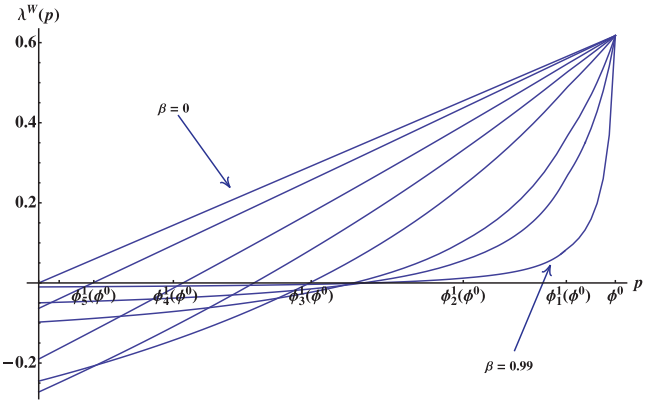

FIGURE **2.** (Color online) The Whittle index of a process with parameters: $\phi^0 = 0.95$, $R = 1$, $\alpha = 0.35$ and $c = 0$, and computed for instances with discount factors $\beta \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99\}$. The optimal active set in this example is the set of p for which $\lambda^W(p) > 0$.

remains constant for all $p \in \mathcal{P} \setminus \{0\}$ such that $p \neq \phi_t(\phi_0)$ for $t = 0, 1, \ldots$, while it has decreasing jump discontinuities in the set $p = \phi_t(\phi_0)$ as $t = 0, 1, \ldots$.

Thus, within the belief state intervals $\big[\, \phi_t(\phi^0)\,;\, \phi_{t-1}(\phi^0)\,\big)$, with $\phi_t(\phi^0)$ given by expression (7), for all natural $t \geq 1$ it therefore holds that $d(\phi_0, p)$ remains constant and $\lambda^*(p)$ is a linear, continuous and increasing function in p.

In order to show continuity of the index function $\lambda^*(p)$, it remains to establish the continuity for the set of points in which the function $d(\phi_0, p)$ has jump discontinuities. For such a purpose, using the fact that at those critical values of p the active and passive actions are indifferent to the decision maker, it can be shown that

$$\lim_{p \rightarrow : \phi_t(\phi^0)^-} \lambda^*(p) = \lim_{p \rightarrow : \phi_t(\phi^0)^+} \lambda^*(p) = \lambda^*(\phi_t(\phi^0)), \quad \forall\, t = 1, 2, \ldots,$$

which completes the proof of Corollary 3.6. ■

COROLLARY 3.7: *The single-bandit problem* (3) *is optimally solvable by threshold policies, that is, for every* $\lambda \in \mathbb{R}$ *there exists a threshold* $p^*(\lambda)$ *such that for any* $p \in \mathcal{P} \setminus \{0\}$ *it is optimal to activate the process if and only if* $p > p^*(\lambda)$.

PROOF: It follows from the analysis to derive the Whittle index in Theorem 13, that the optimal policy for problem (3) can be expressed as follows: $a^* = 1$ for all $p \in \mathcal{P} \setminus \{0\}$: $p \geq \phi_{n-1}^1(\phi^0)$ and $a^* = 0$, otherwise. ■

3.2. Example

To illustrate the previous analysis with an example, consider a process with the following parameters: $\phi^0 = 0.95$, $r = 1$, $\alpha = 0.35$ and $c = 0$. Figure 2 plots the corresponding Whittle index function, given by expression (13), for instances with discount factors $\beta \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99\}$.

Notice that the Whittle index for the instance $\beta = 0$ reduces to the index $\lambda^W(p) = R(p, 1)$, commonly known as the *Myopic* index, and thus henceforth denoted as $\lambda^M(p) \triangleq R(p, 1)$. Further, when $\beta = 0$ the optimal policy advices to observe the process regardless

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

of its current state p, that is, the optimal active set is $\mathcal{P} \setminus \{0\}$ or the optimal stopping time for any $p \in \mathcal{P} \setminus \{0\}$ is infinite. However, as β increases, the optimal active set becomes a subset of the controllable states, for example, when $\beta = 0.3$ it holds that, starting from ϕ^0 it is optimal to be active for 4 periods (until state $\phi_4^1(\phi^0)$ is reached), while for $\beta = 0.99$ it is optimal to be active for 3 periods (until state $\phi_3^1(\phi^0)$ is reached). In the limit, for $\beta = 1$ the Whittle index converges to an index that takes value 0 for all $\mathcal{P} \setminus \{0, \phi^0\}$, while in the reinitializing state ϕ^0 it takes the value $R(\phi^0, 1)$.

It follows from the previously mentioned equivalence between the optimal active sets and the optimal stopping times, that for every β, the Whittle index policy can be equivalently expressed in terms of the optimal maximum possible number of observations starting from the initial state ϕ^0 until the process yields its final reward, which we shall denoted by t^*_β, therefore admitting a simpler and tractable expression alternative to the computation of expression (13). Thus, for $\beta = 1$ the Whittle Index rule is equivalent to a 1-limited Round Robin observation rule, for $\beta = 0$ it is equivalent to an ∞-limited Round Robin rule and for some general $0 < \beta < 1$ it is equivalent to a t^*_β-limited Round Robin rule.

4. PROPERTIES OF THE WHITTLE INDEX RULE

We now explicitly define a Whittle Index rule for the multi-armed problem (1) based on the index expression (13), and two alternative naive heuristics index rules. We shall further establish the optimality of the Whittle Index rule for solving problem (1) for the special case of N *stochastically heterogeneous* processes (i.e., having distinct parameter specifications) under the Expected Total criterion, that is, the case corresponding to letting $\beta = 1$. We shall also give a closed-form expression for the suboptimality gap of the alternative index rules for the special case in which there is no constraint in the number of processes that can be simultaneously observed, that is, for $M = N$.

Definition 4.1: The Whittle Index rule for the multi-armed problem (1) is implemented as follows: at time t, the index is computed using expression (13) for each of the N processes independently, and the M processes yielding the highest index values, as long as they are positive (i.e., $\lambda^W_{n,t}(p) > 0$), are observed at time t. Further, in the case of a tie among two or more (positive) index values, we shall choose to observe the process that has been least (unsuccessfully) observed up to time t. If processes have been previously observed the same number of periods, ties are broken arbitrarily.

The use of such a problem specific tie-breaking rule is a novel feature proposed in this paper. Any identical processes, when in the same state at time t, will have the same index value at that time, although they may have been (unsuccessfully) observed a different number of times. For instance, if at time t the controller must choose between two processes, both at a common state ϕ^0, but process 1 has been observed before while process 2 has never been observed, the expected net reward of observing each of them respectively is $R(\phi^0, 1)b(\phi^0)$ and $R(\phi^0, 1)$. Naturally, if the least observed process has a higher priority despite the fact that their Whittle Index value is the same, then we expect to obtain a higher immediate expected reward by observing it. This difference will be particularly important for the discounted case, in which the time of job completion affects the rewards obtained from them. Moreover, the inclusion of this additional tie-breaking rule can be used for simplifying some of the optimality results' proofs presented in this section.

We shall further define two alternative well-known index-based heuristics for the multi-armed problem (1): the *Myopic* and the *Belief* Index rules, respectively taking index

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

$\lambda^M(p) \triangleq R(p,1)$ and $\lambda^B(p) \triangleq p$. For the sake of the fairness in the comparison, they will be implemented in an analogous way as the Whittle index (i.e., using the same tie-breaking rule). Usually for the cases in which not only the optimal policy for the MARBP but also the Whittle Index rule for a single-bandit subproblem is not easily derived, this type of simpler rules are the most commonly implemented. For an example of the application of these two alternative index-based policies and a comparison of their performance against the Whittle Index rule see, for example [16] or [8].

Furthermore, we propose the following tractable heuristic rule:

The $\bar{n}$-limited Round Robin rule, observes the M least observed processes whose state is greater than $\phi^1_{\bar{n}}(\phi^0)$, as long as $R(\phi^1_{\bar{n}-1}(\phi^0),1) > 0$.

THEOREM 4.2: *The Whittle index rule in Definition 4.1 is optimal for problem* (1) *for the case of* N *stochastically heterogeneous processes for any* $1 \leq M \leq N$ *when* $\beta = 1$. *Further, it is equivalent to the following simple* 1*-limited Round Robin rule:* at each time slot t, observe (at most) M processes only if they are in their reset state ϕ^0_n, as long as $R(\phi^0_n,1) > 0$.

PROOF: Following an approach similar to the one in [10], we shall show optimality by deriving and comparing relevant bounds on the resulting value functions under different rules. Consider first the case in which $c = 0$. A natural upper bound for the objective function (1) when $\beta = 1$ and for any $1 \leq M \leq N$ is such that $V_1^*(\phi^0) \leq \sum_{n=1}^N r_n$. Given that each one of the N processes generates a reward r_n when observed at state $s = 1$, the best that any observation scheduling rule could do is to succeed with all of them, hence $\sum_{n=1}^N r_n$ is the (obvious) maximum attainable value for the total expected objective function.

Next, we compute the expected value of the objective function under the Whittle Index rule (as in Definition 4.1).

For $\beta = 1$, the Whittle rule induces a 1-*limited Round Robin* scheme in which every process is observed once every two slots, as long as it is in the state ϕ^0_n and until it yields its final reward. Under such a rule, every process will yield its final reward in finite time with probability 1, given that the probability of completing a job by time t is $1 - b(\phi^0_n)^t$. So for any possible $1 \leq M \leq N$, all the N processes will be eventually operated under this rule, though at different moments of time, and all the possible rewards will be achieved in a finite time.

Using the above reasoning, for $\beta = 1$ and operated under the Whittle Index rule the expected flow of rewards yields the following value for the objective function, denoted by $V_1^W(\phi^0)$:

$$\begin{aligned} V_1^W(\boldsymbol{\phi}^0) &= \sum_{n=1}^N r_n(1-\alpha_n)\phi^0_n(1 + b(\phi^0_n) + b(\phi^0_n)^2 + \cdots + b(\phi^0_n)^t + \cdots) \\ &= \sum_{n=1}^N r_n(1-\alpha_n)\phi^0_n \sum_{t=0}^{\infty} b(\phi^0)^t. \end{aligned} \tag{18}$$

By Lemma 3.3, (18) is reduced to rN, which coincides with the upper bound of the objective function (1). Hence, given that $V_1^W(\phi^0) = \sum_{n=1}^N r_n$, the Whittle Index rule is optimal $V_1^W(\phi^0) = V_1^*(\phi^0)$. Further, notice that because the 1-*limited Round Robin* scheme can be implemented for every process in finite time, such an optimality result is true regardless of the value of M.

Regarding the case in which problem (1) is considered for $c > 0$ or under the β-discounted criterion, we cannot show optimality using the rough bound on the value function

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

for the expected total case, yet an upper bound can be derived solving the Whittle relaxation of problem (1) and using a Lagrangian approach to solve it, that is,

$$V_\beta^* \leq V_\beta^L \triangleq \inf_{\lambda \geq 0} \left[\sum_{n=1}^{N} V_\beta^n(\phi_n^0, \lambda) + \lambda \frac{M}{(1-\beta)} \right], \tag{19}$$

where $V_\beta^n(\phi_n^0, \lambda)$ is defined as in (3) for every $n = 1, \ldots, N$.

By solving the convex optimization problem posed by (19), it is derived that V_β^L for the case $\beta \to 1$ results in $\lambda^* = 0$ and $V_1^L = \sum_{j \in J}[r_j - (c/[(1-\alpha_j)\phi_j^0])]$ where J is the set of processes for which it holds that $c_n < R(\phi_n^0, 1)$. By an analogous reasoning to the one used for computing the total expected value function under the Whittle index rule in (20), it follows that the 1-*limited Round Robin* scheme induced by this rule will yield all the final rewards incurring in an expected cost equal to the expected cost of the induced cycles, that is, $V_1^W = \sum_{j \in J}[r_j - (c/[(1-\alpha_j)\phi_j^0])]$. Thus, again $V_1^W = V_1^L = V_1^*$. ■

From the above proof it follows as well that when $c = 0$ the optimality result will hold for any $\bar{n}$-limited Round Robin rule such that the length of its active cycle $\bar{n}$ is finite. Given that the Myopic and Belief rule are equivalent to a ∞-limited Round Robin rule, we can expect them to be suboptimal in that case. Next, we introduce a theorem stating its suboptimality gap in closed-form for that case and under certain assumptions.

THEOREM 4.3: *The suboptimality gap* $(V_1^M(\boldsymbol{\phi}^0) - V_1^*(\boldsymbol{\phi}^0))$ *for the total expected performance achieved under the Myopic index rule or the Belief index rule for the special case of stochastically heterogeneous processes with* $c = 0$, $\beta = 1$ *and*

(a) $M = N$, *is* $\sum_{n=1}^{N} r_n(1 - \phi_n^0)$;
(b) $M = 1$ *and processes such that:* $\phi_1^0 \geq \phi_2^0 \geq \cdots \geq \phi_N^0$ *and* $\phi_1^1(\phi_1^0) \leq \phi_N^0$, *is* $r_l(1 - \phi_n^0)$, *where* $r_{min} \leq r_l \leq r_{max}$.

PROOF: First, we compute the value of the total expected performance under the Myopic index rule or the Belief index rule under the assumptions in (a). Notice that both rules are equivalent in the case of heterogeneous processes with $c = 0$ and $M = N$. Both index functions are strictly increasing in the belief state, and both are strictly positive when considering a given state of the set of controllable states, therefore inducing identical decisions over time. However, notice that both policies are not equivalent to the Whittle index rule, basically because the Myopic and Belief index rules are equivalent to an ∞-limited Round Robin rule while the Whittle index rule is equivalent to a 1-limited Round Robin rule.

For $\beta = 1$, $c = 0$ and $M = N$, operated under the Myopic (or the Belief) index rules, the expected flow of rewards yields the following expected value $V_1^M(\phi^0)$ (or $V_1^B(\phi^0)$) for the objective function:

$$\begin{aligned} V_1^M(\boldsymbol{\phi}^0) &= \sum_{n=1}^{N} r_n(1-\alpha_n) \left[\phi_n^0 + \phi_1^1(\phi_n^0) b(\phi_n^0) + \phi_2^1(\phi_n^0) b(\phi_n^0) b(\phi_1^1(\phi_n^0)) + \cdots \right. \\ &\qquad \left. + \phi_t^1(\phi_n^0) \prod_{s=0}^{t-1} b(\phi_s^1(\phi_n^0)) + \cdots \right] \\ &= \sum_{n=1}^{N} r_n(1-\alpha_n) \sum_{t=0}^{\infty} \phi_t^1(\phi_n^0) \prod_{s=0}^{t-1} b(\phi_s^1(\phi_n^0)). \end{aligned} \tag{20}$$

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

As a consequence of Lemma 3.3, (20) reduces to $V_1^M(\boldsymbol{\phi}^0) = \sum_{n=1}^N r_n \phi_n^0 < V_1^*(\phi^0)$. Hence, the suboptimality gap for these two simple index rules is computed in closed-form for the case $M = N$ and $c = 0$ to be $\sum_{n=1}^N r_n(1 - \phi_n^0) > 0$. Thus, as it will be illustrated through the computational experiments, the gap decreases with ϕ_n^0.

Next, we compute the value of the total expected performance under the Myopic index rule or the Belief index rule under the assumptions in (b). The relation between the restarting states ϕ_n^0 (together with the tie-breaking rule) ensures that under the Myopic or Belief rule no process will be activated two consecutive times until there is only one process that has not yet yield its final reward. Once there is only one process, which we denote by the subscript l, with the possibility of yielding a reward, the Myopic (and the Belief) policy would activate that process over an infinite period of time. As a consequence of Lemma 3.3, (20) reduces to $V_1^M(\phi^0) = \sum_{n \in N \setminus l} r_n + r_l \phi_l^0 < V_1^*(\phi^0)$.

For the case of stochastically identical process, (20) for case (b) reduces to $V_1^M(\phi^0) = (N-1)r + r\phi^0$, and the suboptimality gap to $(1 - \phi^0)r$. ■

Regarding the case in which problem (1) is considered for $\beta = 1$ and $c > 0$, following an analogous reasoning, it can be shown that the simpler index rules will also be suboptimal in this case. Let $\bar{t}_n(c)$ be the maximum prescribed number of consecutive active slots for process n on excess of 1 that are prescribed under the Myopic or Belief index rules when the cost is c. The simpler index rules can be shown to attain the following expected performance values: $V_1^M = V_1^B = \sum_{j \in J}[\phi_n^0 r_n - (c/(1-\alpha_n))] - (c \sum_{j \in J} \bar{t}_n(c)(1-\phi_n^0)/(1-\alpha_n^{\bar{t}_n(c)})\phi_n^0)$. Therefore, it follows that the Whittle index rule is also not only optimal for any c when $\beta = 1$ but also that the other index rules are suboptimal, since it holds that $V_1^* - V_1^M = \sum_{j \in J}(1-\phi_j^0)[r_j - (c/(1-\alpha_j)\phi_j^0)] + (c \sum_{j \in J} \bar{t}_n(c)(1-\phi_n^0)/(1-\alpha_n^{\bar{t}_n(c)})\phi_n^0) > 0$ for $c > 0$ and $\bar{t}(c) > 0$.

For the more general case, in which $M < N$ and $\beta = 1$, the computation in closed-form of the total expected value function under the Myopic or Belief index rule is less straightforward. Yet, it can be intuitively argued that both of them will also be suboptimal. Under any of these heuristics the first M terms of the sum defining $V_1^M(\phi^0)$ coincide with the ones achieved by the Whittle index rule. From the $M + 1^{th}$ term onwards both the Myopic or Belief index rule, with a strictly positive probability, take an action that is different from the optimal action (i.e., the action prescribed by the Whittle index rule), thus generating a total expected reward strictly less than rN.

In fact, denoting by $\hat{N}_t(M)$ the expected number of processes which have not yet yielded their final reward at time t for some $1 \leq M \leq N$, the probability that at time t the naive index rules *diverge* from the optimal action (i.e., the action prescribed by the Whittle index rule) is at least equal to: $P(0 < \hat{N}_t(M) \leq M)$. When the event $(0 < \hat{N}_t(M) \leq M)$ occurs, the simpler index rules prescribe to activate all of the $\hat{N}_t$ processes, even if they are not in their initial state (i.e., they never prescribe to reinitialize those processes). Furthermore, from the time at which the $M + 1^{th}$ term of the sum $V_1^M(\phi^0)$ is added onwards, the event $(0 < \hat{N}_t(M) \leq M)$ may occur with a strictly positive probability for any $1 \leq M \leq N$, that is, with the probability that in the previous M observations, at least $(N - M)$ and at most $(N - 1)$ processes have reached its final state.

The number of processes reaching its final state at a given time defines a binomial random variable, hence: $P(0 < \hat{N}_t(M) \leq M) = \sum_{i=N-M}^{N-1} \binom{N-1}{N-i}[1 - (1-\alpha)\phi_n^0]^{N-i}[(1-\alpha)\phi_n^0]^i > 0$. Notice further that for $M = N$, that probability is equal to 1, which yields a value function as in (20). Also, as M diminishes, the minimum for the probability of divergence from the optimal action decreases, and since this probability tends to zero as $t \to \infty$, we can expect that the best performance of the simpler index rules will be

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

achieved when the number of possible observations per period is the lowest. We would like to emphasize that this suboptimality result is also noteworthy since it implies that the Myopic Index and the Belief Index rules perform worse as the resource constraint is less binding.

For the β-discounted criterion, for each β the Whittle index rule is equivalent to a simple t^*_β-*limited round robin* heuristic rule. In general, for every β the Whittle index rule will be equivalent to a heuristic that advices to observe processes starting at the initial state at most for t^*_β slots (unless the process reaches its final state) and resting for one slot after that. Given the fact that the Whittle index rule in this case also induces *cycles* equivalent to a t^*_β-limited Round Robin policy, we expect that this simple rule will also outperform the other rules for the β-discounted criterion, especially as $\beta \to 1$ and M increases. Also, following this reasoning it can be argued that this will be the case in any other situation in which the observation rule selected tends to reinitialize processes less often than the optimal rule would do.

5. NUMERICAL STUDY

We present the results of some of the simulation studies performed with the goal of illustrating the ideas presented in the previous sections. The experiments are based on MATLAB implementation codes developed by the author, where the relative performance of the proposed Whittle index rule is compared against the other previously described naive index policies and to the corresponding upper bound. For each instance, 10^4 independent simulation runs were performed on a horizon of $T = 10^4$ time slots. In each experiment, the resulting mean total reward function under different rules is reported together with 95% confidence intervals around that mean to evaluate the statistical significance of the results. All the results reported in this section are statistically significant at a 95% confidence level.

In all instances, the rules considered are: the Whittle index rule, the *myopic* index rule, the *belief state* index rule (both as defined in Section 4), the t^*-*limited round robin* rule, based on the simple rule equivalent of the Whittle index rule proposed in Section 4, the $\bar{n}$-*limited* round robin rule with $\bar{n} \neq t^*$, and the *random* selection policy which picks a process to observe at random, with each process having the same probability of being selected.

5.1. Experiment #1

In this experiment we illustrate the optimality result announced in 4.2 and the suboptimality gap of the naive index rules presented in 4.3 while studying the effect of varying the hard sample-path resource constraint in an instance with non-identical processes for the case $\beta = 1$. We considered a total of $N = 100$ processes, where 75 have a reinitializing state equal to $\phi^0 = 0.5$ while for the 25 remaining it is $\phi^0 = 0.8$. For the N processes we have considered a common misdetection error $\alpha = 1/3$, $r = 1$ and $c = 0$. This base instance was modified by letting M increase in 1 unit from 1 to N. The upper bound on the objective function is in this case $rN = 100$.

Following the arguments introduced in Section 4, it can be shown that for the case $\beta = 1$ and $c = 0$ any $\bar{n}$-limited round robin observation scheduling rule with $\bar{n} < \infty$ will be optimal. However, the lifetime of the system (i.e., the time until all processes yield its final reward) will grow significantly the more the cycle diverges from the optimal one. By the same reasoning, the optimality of other finite limited Round Robin rules different from the t^*-limited rule will not hold for $c \neq 0$ or $\beta < 1$; however, we expect that these rules,

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

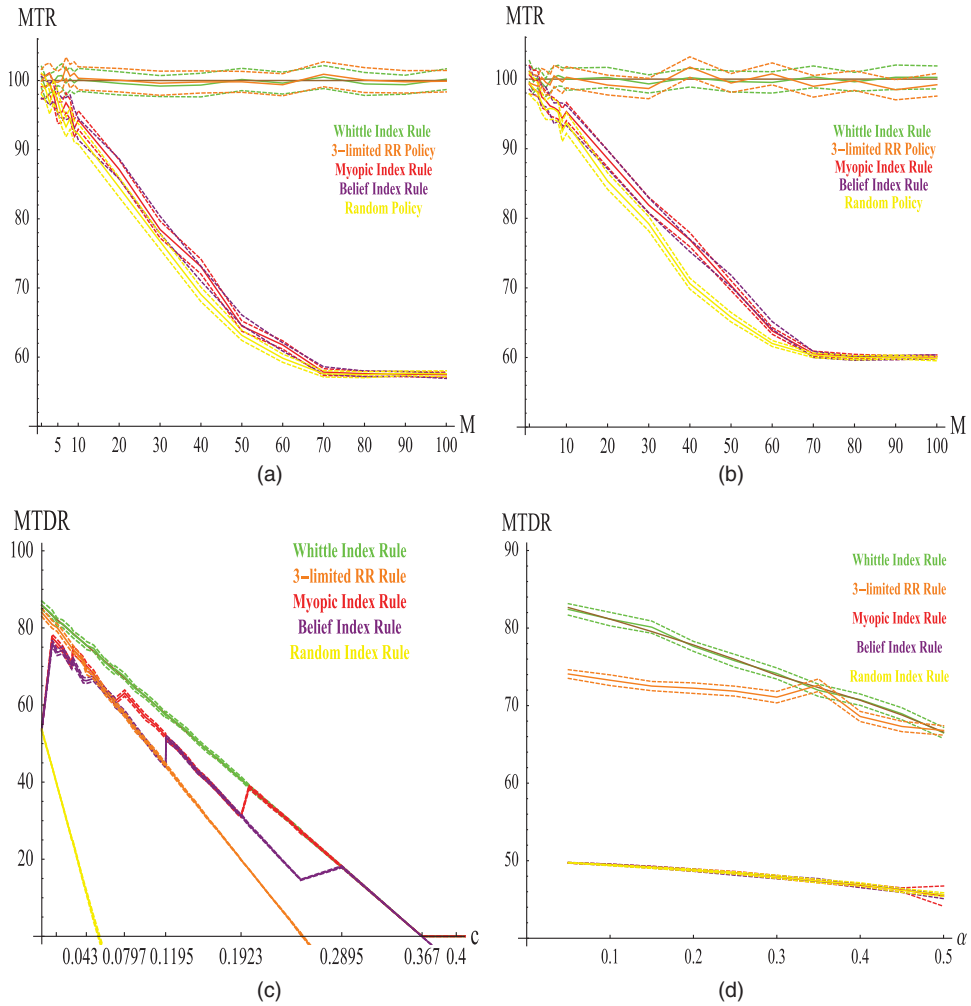

FIGURE **3.** (Color online) Computational experiments: performance results (a) Mean Total Reward vs. M (Non-Identical Processes). (b) Mean Total Reward vs. M (Identical Processes). (c) Mean Total Discounted Reward versus c. (d) Mean Total Discounted Reward versus α.

though suboptimal, may outperform the Myopic and Belief rules only in some instances. These issues shall be explored in Experiment 5.3.

In this experiment, we illustrate the optimality of other $\bar{n}$-limited round robin rules by implementing a rule with a finite cycle of (at most) $\bar{n} = 3$ consecutive active time slots, that is, a 3-limited Round Robin rule. Results displayed in Figure 3(a) show that the Whittle index rule is statistically identical to the upper bound on the total expected rewards. Furthermore, the $\bar{n}$-limited round robin policy, which was computed in this experiment both for $\bar{n} = t^* = 1$ and $\bar{n} = 3$ (though the figure displays only the 3-limited), resulted in both cases also optimal in terms of the objective function.

Results displayed in Figure 3(a) show that both the Whittle index rule and the 3-limited Round Robin rules outperform all the other rules with the largest suboptimality gap (of around 43%) occurring again for $M \geq 70$. The Myopic and the Belief index rules are

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

statistically equivalent in practically all of the values of M, with their resulting mean performance deteriorating as M increases. All the policies are statistically equivalent (and moreover optimal) only when M takes very small values ($M \leq 3$) and the variability of the performance attained by the optimal rules is relatively constant over the values of M whereas it decreases for the suboptimal ones as M grows. It is also noteworthy that in almost all the range of values of M these two naive index rules do not result in statistically significant improvements over the random policy.

5.2. Experiment #2

In this experiment we consider a base instance with $N = 100$ identical processes, each with parameters $\phi^0 = 0.6$, $\alpha = 0.25$, $r = 1$ and $c = 0$, for the expected total criterion (i.e., for $\beta = 1$). This base instance was modified to asses the effect of varying of the hard sample-path resource constraint, specifically by letting M (the number of processes that can be observed at a time slot t) increase by 1 unit from 1 to N. Again, the rough upper bound on the objective function is in this case $rN = 100$. Results displayed in Figure 3(b).

As expected, the Myopic and the Belief index rules are statistically equivalent policies for the identical processes case and they are both suboptimal. Once more, their suboptimality gap grows as M increases, reaching its theoretical maximum value of $40\% = (1 - \phi^0)$(corresponding for $M = N$) for $M \approx 70$. Remarkably, for the range of values of M smaller than that and until the set of values for which all the policies are statistically equivalent and optimal (roughly $M \leq 3$), these two heuristic rules do statistically significant improve over the random policy, contrary to what occurred in the previous experiment for non-identical processes. Finally, notice that the variability of the performance achieved by all the rules exhibits the same behavior as in the previous experiment.

5.3. Experiment #3

In this experiment we study the effect of including a strictly positive observation cost $c > 0$ when the discount factor equals $\beta = 0.95$ and for a case of identical processes each with parameters $\phi^0 = 0.55$, $\alpha = 1/3$, $r = 1$, $M = 100$. This base instance was modified to asses the effect of including a strictly positive observation cost, specifically by letting c increase in 0.05 units from 0 to 0.75. We expect that the naive index rules will perform better as the cost increases, and to even outperform the 3-limited Round Robin when c becomes sufficiently large. The upper bound on the objective function is in this case computed for each instance using the Lagrangian relaxation approach described in Section 4.

Results displayed in Figure 3(c) show that the Whittle Index rule is statistically equivalent to the upper bound for any value of the observation cost c while, as expected, the suboptimality gap of the Myopic and Belief index rules decreases as c grows. Moreover, for small values of c (approximately less than 0.03) the 3-*limited* Round-Robin rule performs better than the other naive heuristics, while it performs statistically equivalent to them when c is around 0.06 and it is overperformed by them for larger values of c. The poor performance attained by the random rule is mainly explained by the fact that its definition does not take into account the observation cost.

In fact, it turns out that as c varies each of the index rules becomes equivalent to a simple $\bar{n}$–*limited* Round-Robin rule. Thus, to explain such results, we summarize the equivalence relations in Table 1. For $c = 0$ and $\beta = 0.95$ the Whittle index rule is equivalent to a 2-*limited* Round Robin rule, while the Myopic and Belief index rule are ∞-*limited*, therefore resulting in the maximum possible divergence among these index rules. When the observation cost is in the range $[0.2895, 0.367)$ all three index polices become equivalent to

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

TABLE 1. Equivalence between index rules and $\bar{n}$-limited round Robin rules as a function of c, with a is a integer such that $a \geq 4$

c	Whittle index rule	Myopic index rule	Belief index rule
$[0.55, \infty)$	0-*limited*	0-*limited*	0-*limited*
$[0.367, 0.55)$	0-*limited*	0-*limited*	1-*limited*
$[\mathbf{0.2895}, \mathbf{0.367})$	**1-*limited***	**1-*limited***	**1-*limited***
$[0.1923, 0.2895)$	1-*limited*	1-*limited*	2-*limited*
$[0.1195, 0.1923)$	1-*limited*	2-*limited*	2-*limited*
$[0.07971, 0.1195)$	1-*limited*	2-*limited*	3-*limited*
$[\mathbf{0.06}, \mathbf{0.07971})$	1-*limited*	**3-*limited***	**3-*limited***
$[\mathbf{0.0433}, \mathbf{0.06})$	2-*limited*	**3-*limited***	**3-*limited***
$[0.02887, 0.0433)$	2-*limited*	3-*limited*	4-*limited*
$[0.0148, 0.02887)$	2-*limited*	4-*limited*	4-*limited*
$(0, 0.0148)$	2-*limited*	a-*limited*	a-*limited*
0	2-*limited*	∞-*limited*	∞-*limited*

a 1-*limited* Round Robin rule but the total expected discounted rewards are reduced to an approximate value of 20. Also, from the table it follows that for values of c less than 0.0433 both naive index policies are overperformed by the 3-*limited* Round-Robin rule.

5.4. Experiment #4

In this final experiment we study the effect of differences in the misdetection error probability α in an instance of identical processes for the case $\beta = 0.9$. Once more, we have considered a total of $N = 100$ processes, where in the base instance all of them have $\phi^0 = 0.5$, $\alpha = 0.05$, $r = 1$ and $c = 0$ and $M = N$. Then we modify the base instance by letting the misdetection probability vary from $\alpha = 0.05$ to 0.5 in increments of 0.05. The upper bound on the objective function in each case is computed using the Lagrangian relaxation value.

Results displayed in Figure 3(d) show that the performance measure decreases for all rules as the misdetection error probability grows. The naive index rules are statistically equivalent to the random policy and their variability is smaller than the variability of Whittle index rule and the 3-limited round robin rule.

Further, the Whittle index rule is statistically equal to the upper bound over all values of the parameter α, overperforming the all the other rules. The 3-limited round robin rule results equivalent to the Whittle index policy for values of α larger than 0.35 and it further overperforms the other index rules for all values of α.

6. CONCLUDING REMARKS

In this paper, we have proposed a simple yet intractable POMDP model with application in surveillance systems dealing with the detection and expulsion of smart intruders. The model admits other applications in similar contexts. For example, consider a supervisory control system in which multiple processes are monitored and controlled. The state 1 represents an abnormal state and 0 the normal state. While a processes is being monitored, its state can only change if an abnormality is detected and corrected. The objective is to control the processes, identifying and rectifying processes that are in the abnormal state to ensure the quality or security of the system.

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

For solving the original model, we reformulated it as a MARBP with reinitializing states and we introduced a novel dynamic scheduling policy based on an index function which was further shown to be optimal in some special cases. Moreover, this optimality property was observed in more general scenarios through simulations. For the proposed model, the paper established analytically the existence of the Whittle index, obtaining a closed-form expression for it and analytically showing the suboptimality of other widely used heuristics under some conditions.

Besides the above mentioned theoretical results that are of concern for the model introduced by this work, we believe two important conclusions can be drawn from these results which have a relevance that goes beyond the scope of this paper. The first one is in relation to the design of simple tractable heuristics based on myopic approaches which is in stark contrast to the results valid for the MARBP models studied in [10,11] in which myopic rules where optimal or nearly optimal. This paper shows that in problems in which the passive action has a *recovery* effect on the states of arms (a situation likely to occur when arms suffer from exhaustion, as human resources), those polices which do not advice arms to rest are very likely to be substantially far from the optimal. Hence, if we must use a heuristic policy in such instances, it is more reasonable to deploy a heuristic that is defined in such a way that it will cyclically alternate between working and resting every arm. As shown in this paper, for the present model and in the case $\beta = 1$ and $c = 0$ any *cycled* policy of this sort, regardless of the cycle composition, will be optimal in terms of the objective function.

The second conclusion is regarding the potentially good performance of the Whittle index policy as an approximate solution method for POMDP models. In this particular case, given the simplicity of the model, this rule turns out to be optimal in many instances. We believe that the results reported in this paper suggest that in more complex cases (such as the model in [16]), the application of this approach may lead to designing a rule that may offer significant performance gains at a feasible computational cost. We regard this direction as a highly fruitful one to continue research.

Finally, it is our hope that this work will contribute to stimulate the development of tractable decision-making rules for models that extend the present formulation. For example, to solve a more complex problem that includes false positives, the case in which targets can re-appear in the sites after being detected or even a model in which smart targets have a different reaction that the one modelled here.

Acknowledgements

The author is grateful for the contributions of Peter Jacko to my thinking about bandit problems and for his immense encouragement for the completion of this paper. This work was partially supported by grant SA-2012/00331 of the Department of Industry, Innovation, Trade and Tourism (Basque Government).

References

1. Ahmad, S., Liu, M., Javidi, T., Zhao, Q., & Krishnamachari, B. (2009). Optimality of myopic sensing in multichannel opportunistic access. *IEEE Transactions on Information Theory* 55(9): 4040–4050.
2. Gittins, J.C. & Jones, D.M. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gani (ed.), *Progress in Statistics*. North–Holland, Amsterdam, pp. 241–266.
3. Gittins, J.C. & Jones, D.M. (1979). Bandit processes and dynamic allocation indices (with discussion). *Journal of the Royal Statistical Society B* 41: 148–177.
4. Glazebrook, K.D., Ruiz-Hernandez, D., & Kirkbride, C. (2006). Some indexable families of restless bandit problems. *Advances in Applied Probability* 38(3): 643–672.
5. Glazebrook, K.D., Hodge, D.J., & Kirkbride, C. (2013). Monotone policies and indexability for bidirectional restless bandits. *Advances in Applied Probability* 45(1): 51–85.

6. Jacko, P. (2011). Optimal index rules for single resource allocation to stochastic dynamic competitors. *Proceedings of ValueTools*, pp. 425–433.
7. Jacko, P. & Sanso, B. (2012). Optimal anticipative congestion control of flows with time-varying input stream *Performance Evaluation* 69(2): 86–101.
8. Jacko, P. & Villar, S.S. (2012). Opportunistic schedulers for optimal scheduling of flows in wireless systems with ARQ Feedback. *24th International Teletraffic Conference (ITC)* IEEE, pp. 1–8.
9. Kreucher, C., Blatt, D., Hero, A., & Kastella, K. (2006). Adaptive multi-modality sensor scheduling for detection and tracking of smart targets. *Digital Signal Processing*, 16(5): 546–567.
10. Liu, K. & Zhao, Q. (2010). Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory* 56(11): 5547–5567.
11. Liu, K., Weber, R., & Zhao, Q. (2011). Indexability and whittle index for restless bandit problems involving reset processes. *In proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, pp. 7690–7696.
12. Lovejoy, W.S. (1991). Computationally feasible bounds for partially observed Markov decision processes, *Operations Research* 39(1): 162–175.
13. Mansourifard, P., Javidi, T., & Krishnamachari, B. (2012). Optimality of myopic policy for a class of monotone affine restless multi-armed bandits. *In the Proceedings of the 51th IEEE International Conference on Decision and Control (CDC)*.
14. Niño-Mora, J. (2001). Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability* 33(1): 76–98.
15. Niño-Mora, J. (2007). Dynamic priority allocation via restless bandit marginal productivity indices. *Top* 15(2): 161–198.
16. Niño-Mora, J. & Villar, S.S. (2011). Sensor scheduling for hunting elusive hiding targets via Whittle's Restless Bandit Index Policy. *5th International Conference on Network Games, Control and Optimization (NetGCooP)*, pp. 1–8.
17. Papadimitriou, C.H. & Tsitsiklis, J.N. (1999). The complexity of optimal queuing network control. *Mathematics of Operations Research* 24(2): 293–305.
18. Sondik, E.J. (1978). The optimal control of partially observable Markov processes over the infinite horizon: discounted costs. *Operations Research* 26(2): 282–304.
19. Villar, S.S. (2012). *Restless bandit models for sensor management*. LAP LAMBERT Academic Publishing, Germany.
20. Weber, R.R. & Weiss, G. (1990). On an index policy for restless bandits. *Journal of Applied Probability* 27(3): 637–648.
21. Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability* 25: 287–298.
22. Zhao, Q., Krishnamachari, B., & Liu, K. (2008). On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance. *IEEE Transactions on Wireless Communications* 7(12): 5431–5440.
23. Mathpages, Algebra, Linear Fractional Transformations. Available from http://www.mathpages.com/home/kmath464/kmath464.htm

APPENDIX

Proof of Lemma 3.3

Lemma 3.3 states the validity of the following results:

$$(a)\prod_{s=0}^{t-1}\Big[1-(1-\alpha)\phi_s^1(p)\Big]=1-(1-\alpha^t)p;\quad (b)\phi_t^1(p)\prod_{s=0}^{t-1}\Big[1-(1-\alpha)\phi_s^1(p)\Big]=\alpha^t p.$$

Proof: We show these results by induction. We start with part (a), and show that

$$(a)\prod_{s=0}^{t-1}\Big[1-(1-\alpha)\phi_s^1(p)\Big]=1-(1-\alpha^t)p,\quad \forall t\in\mathbb{Z}^+\triangleq\{1,2,\ldots\}$$

For $t=1$, by definition of $\phi_0^1(p)$ it holds $\prod_{s=0}^{0}\Big[1-(1-\alpha)\phi_s^1(p)\Big]=1-(1-\alpha)p$.

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press

Next, if we let it be true for some $t \geq 1$, we then have that $\prod_{s=0}^{t-1} \left[1-(1-\alpha)\phi_s^1(p) \right] = 1-(1-\alpha^t)p$.

Thus, it holds that $\prod_{s=0}^{t} \left[1-(1-\alpha)\phi_s^1(p) \right] = \left[1-(1-\alpha)\phi_t^1(p) \right] \left(1-(1-\alpha^t)p\right)$.

Next, using the expression (7) that

$$\phi_t^1(p) = \frac{\alpha^t\, p}{1-(1-\alpha^t)\, p}, \tag{A.1}$$

we conclude that

$$\prod_{s=0}^{t} \left[1-(1-\alpha)\phi_s^1(p) \right] = \left[1-(1-\alpha)\frac{\alpha^t\, p}{1-(1-\alpha^t)\, p} \right] \left(1-(1-\alpha^t)p\right).$$

Finally, straightforward algebra yields that

$$\prod_{s=0}^{t} \left[1-(1-\alpha)\phi_s^1(p) \right] = 1-(1-\alpha^{t+1})p,$$

which completes the proof of part (a).
Next, we shall prove part (b).

$$(b) \quad \phi_t^1(p) \prod_{s=0}^{t-1} \left[1-(1-\alpha)\phi^i t1_s(p) \right] = \alpha^t p, \quad \forall t \in \mathbb{Z}^+.$$

Notice that $\phi_t^1(p)$ admits expression (7) because is the tth iterate of a Möbius Transformation defined by $\phi^1(p) \triangleq \frac{\alpha\, p}{1-(1-\alpha)\, p}$ whose associated matrix is

$$\Phi^1 = \begin{pmatrix} \alpha & 0 \\ 1 & -(1-\alpha) \end{pmatrix}.$$

Thus, by properties of Möbius Transformations, $\phi_t(p)$ is also a Möbius Transformation with associated matrix $(\Phi^1)^t$.

We shall also prove part (b) by induction. By definition $\phi_0^1(p)$ and $\phi_1^1(p)$, it holds for $t = 1$:

$$\phi_t^1(p) \prod_{s=0}^{t-1} \left[1-(1-\alpha)\phi_s^1(p) \right] = \frac{\alpha\, p}{(1-(1-\alpha)\, p)} \left(1-(1-\alpha)p\right) = \alpha p$$

Next, let it be true for $t \geq 1$, that is, $\phi_t^1(p) \prod_{s=0}^{t-1} \left[1-(1-\alpha)\phi_s^1(p) \right] = \alpha^t p$.

Then, it holds that

$$\phi_{t+1}^1(p) \prod_{s=0}^{t} \left[1-(1-\alpha)\phi_s^1(p) \right] = \phi_{t+1}^1(p) \left[1-(1-\alpha)\phi_t^1(p) \right] \prod_{s=0}^{t-1} \left[1-(1-\alpha)\phi_s^1(p) \right].$$

Next, substituting $\phi_{t+1}(p)$ by its equal according to expression (7), we get that

$$\phi_{t+1}^1(p) \prod_{s=0}^{t} \left[1-(1-\alpha)\phi_s^1(p) \right] = \alpha\, \phi_t^1(p) \prod_{s=0}^{t-1} \left[1-(1-\alpha)\phi_s^1(p) \right] = \alpha^{t+1}\, p.$$

which shows part (b), and hence completes the proof of Lemma 3.3 holds for all natural t. ■

https://doi.org/10.1017/S026996481500025X Published online by Cambridge University Press