UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 12-08
Statistics and Econometrics Series 05
April 2012

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

SENSOR SCHEDULING FOR HUNTING ELUSIVE HIDING TARGETS: A RESTLESS BANDIT INDEX POLICY

José Niño-Mora and Sofía S. Villar

Abstract

We consider a sensor scheduling model where a set of identical sensors are used to hunt a larger set of heterogeneous targets, each of which is located at a corresponding site. Target states change randomly over discrete time slots between "exposed" and 'hidden," according to Markovian transition probabilities that depend on whether sites are searched or not, so as to make the targets elusive. Sensors are imperfect, failing to detect an exposed target when searching its site with a positive misdetection probability. We formulate as a partially observable Markov decision process the problem of scheduling the sensors to search the sites so as to maximize the expected total discounted value of rewards earned (when targets are hunted) minus search costs incurred. Given the intractability of finding an optimal policy, we introduce a tractable heuristic search policy of priority index type based on the Whittle's index for restless bandits. Preliminary computational results are reported showing that such a policy is nearly optimal and can substantially outperform the myopic policy and other simple heuristics.

Keywords: Smart Targets; Sensor Management; Sensor Scheduling; Partially Observed Markov Decision Process; Bayes Filter; index policy, Whittle index, real-state Restless Bandits.

The authors are with the Department of Statistics, Carlos III University of Madrid, 28911 Leganés (Madrid), Spain. Email: jnino@est-econ.uc3m.es (José Niño-Mora) and svillar@est-con.uc3m.es (Sofía S. Villar). **Acknowledgements:** This work has been supported in part by the Spanish Ministry of Education and Science project MTM2007-63140 and by the Ministry of Science and Innovation project MTM2010-20808. A preliminary version of this work appeared in proceedings of the International conference on NETwork Games, COntrol and Optimization, NetGCOOP (2011).

Sensor Scheduling for Hunting Elusive Hiding Targets: a Restless Bandit Index Policy

José Niño-Mora and Sofía S. Villar[‡]

April 19, 2012

Abstract

We consider a sensor scheduling model where a set of identical sensors are used to hunt a larger set of heterogeneous targets, each of which is located at a corresponding site. Target states change randomly over discrete time slots between"exposed" and "hidden" according to Markovian transition probabilities that depend on whether sites are searched or not, so as to make the targets elusive. Sensors are imperfect, failing to detect an exposed target when searching its site with a positive misdetection probability. We formulate as a partially observable Markov decision process the problem of scheduling the sensors to search the sites so as to maximize the expected total discounted value of rewards earned (when targets are hunted) minus search costs incurred. Given the intractability of finding an optimal policy, we introduce a tractable heuristic search policy of priority index type based on the Whittle's index for restless bandits. Preliminary computational results are reported showing that such a policy is nearly optimal and can substantially outperform the myopic policy and other simple heuristics.

KEY WORDS: Smart targets; Sensor Management; Sensor Scheduling; Partially Observed Markov Decision Process; Bayes Filter; Index policy; Whittle's index; real-state Restless Bandits.

[*]This work has been supported in part by the Spanish Ministry of Education and Science project MTM2007-63140 and by the Ministry of Science and Innovation project MTM2010-20808.

[†]J. Niño-Mora and Sofía S. Villar are with the Department of Statistics, Carlos III University of Madrid, 28911 Leganés (Madrid), Spain. Email: jnino@est-econ.uc3m.es, svillar@est-econ.uc3m.es

1 Introduction

1.1 Background and Motivation

In recent years, the investigation of effective dynamic policies for operating wireless sensor networks has become an active research area. An issue that has received much attention is the design of scheduling policies to allocate over time a relatively small set of sensor resources to extract the required information about a scene containing a larger set of targets of interest, in order to optimize a system-wide performance objective. See, e.g., the survey [4].

The sensors provide error-prone measurements of the sensed targets, such as their location, or their presence (or absence) at a given location. The current knowledge on each target is represented by its information state, which evolves via Bayesian updates depending on whether or not the target is sensed at each time slot. This allows for the formulation of a variety of optimal sensor scheduling problems as a partially observable Markov decision process (POMDP) with special structure, which often fit into the framework of the real-state multi-armed bandit problem, either in its classic version or, more often, in its restless variant. See, e.g., [12].

Although the restless variant is, generally, computationally intractable, formulating a sensor scheduling problem in such a framework allows for the use of the indexation methodology reviewed in the previous chapter. Such an approach, further provides with a bound on the optimal problem value that can be used to assess the deviation from optimality of a given policy.

In certain situations, sensing actions do not only affect the system's information state (e.g. in terms of its precision) but also alter targets' behavior. This is the case when objective targets are *smart*, in the sense that they react to being sensed by changing their dynamics, so as to hinder their detection or tracking. Sensor scheduling problems complicate substantially when targets under surveillance are able to detect and respond to sensing activities yet, it is natural to expect that different types of reaction would require a different operating rule to optimize the system's performance.

Specifically, sensor scheduling to detect (and/or track) *smart* targets is an application that would strongly benefit from non-myopic decision rules, indicating the controller when it is better not to sense a site for the sake of the possible future gains obtained by influencing the target located at it accordingly. On the contrary, tractable myopic rules, of the type defined by solving a one-period ahead optimization problem, do not inform when a target should not be searched (specially in the case in which there are enough sensing resources available to do so). This is clearly undesirable if targets are elusive, as constantly searching for them makes them more and more elusive, resulting in larger use of system resources (especially in time) to successfully find them.

Despite all these problems, few papers have considered sensor scheduling problems with such reactive targets. Instead targets are typically assumed to follow dynamics that are unaffected by sensing decisions. In the recent literature, some sensor management models have been proposed for smart object localization disregarding such an unrealistic assumption. For instance, in [2] reinforcement learning is used to obtain a non-myopic policy for detection and tracking of smart targets, while [3] uses particle filter methods, and [11] presents a game theoretic analysis.

The model presented in this paper extends such a line of work by investigating a sensor scheduling model where a set of identical sensors are used to hunt a larger (or at least equal) set of heterogeneous targets, each of which is located at a corresponding site. As

in [2], target states change randomly over discrete time slots between *exposed* and *hidden*, according to Markovian transition probabilities that depend on whether sites are searched or not, so as to make the targets elusive. Sensors have a binary mode, so they can be either active or passive at a site, and they are imperfect, failing to detect an exposed target when searching its site with a positive misdetection probability

As a specific motivating application for such a model, we propose the problem investigated in [10], where the targets are mobile platforms (transporter-erector-launchers) for launching short-range ballistic missiles (known as Scuds), and the sites are areas where it is known that such platforms are hidden. In this setting, the sensors can be mounted on unmanned aerial vehicles (UAV). A metric frequently used to measure the effectiveness of such operations is the time to detect all targets. Hence, an effective sensor scheduling rule may be derived by designing a a search policy that aims at maximizing the expected discounted rewards of detecting and destroying all missile launchers, where the discount factor represents how future detections are penalized in a given mission.

1.2 Goals and Contributions

It is the goal of this paper to propose a dynamic and readily implementable index policy for a hunting elusive target model of POMDP type which exhibits a near-optimal performance both under the discounted and the total criterion.

We accomplish this by formulating the resulting POMDP as a real-state Multi-armed Restless Bandit Problem (MARBP) and deploying the recent extensions of the existing theoretical and algorithm results on discrete-state restless bandit indexation to the continuous-state case.

This work makes the following contributions: it successfully deploys the methodology announced in [6] to obtain a novel and dynamic index policy for the model of concern. The PCL-indexability of the model is shown for the expected total discounted problem for discount factors smaller than a critical value. For such a purpose, the lack of closed form expressions for the required performance measures becomes a severe technical difficulty introduced by considering a real state variable.

The remainder of the paper is organized as follows. Section 2 describes and formulates the model. Section 3 reviews the restless bandit indexation approach as it applies to the design of index policies for the present model. Section 4 outlines how to deploy such a methodology to compute the index and summarizes our main results regarding the indexability analysis of the model. Section 5 reports on several simulation experiments where the proposed index policy is compared with alternative heuristic policies. Finally, Section 6 ends the paper with concluding remarks. Detailed analysis and complete proofs will be included in a full version of this paper, which is currently under preparation. A preliminary version of this work appeared in [8].

2 Model description and MARBP Formulation

We consider a model where M sensors are available to hunt $N \geq M$ elusive hiding targets, where each target n is known to hide at a corresponding site $n = 1, \ldots, N$. We assume that the target present at site n alternates its visibility state $s_{n,t}$ at discrete time periods $t = 0, 1, \ldots$ over an infinite horizon between the *hidden* state ($s_{n,t} = 0$), in which it is invisible to sensors but cannot perform its tasks, and the *exposed* state ($s_{n,t} = 1$), in which it can perform its tasks but can be detected by a sensor surveying the site.

Figure 1: A model of a 2-state Markov chain. The arrows represent one-period transitions among the states 0 (hidden) and 1 (exposed) with given probabilities under actions 0 (on the left) and 1 (on the right).

The visibility state $s_{n,t}$ evolves according to Markovian transition probabilities depending on whether or not its site is searched. We assume that only one sensor can search a site at each time slot, and model sensing decisions by binary actions processes $a_{n,t}$, where $a_{n,t} = 1$ if site n is sensed at time t, and $a_{n,t} = 0$ otherwise. When the action taken on site n is $a_{n,t} = a$ the target moves from the hidden to the exposed state (resp. from the exposed to the hidden state, in case the target is not detected) with probability $p_n^{(a)}$ (resp. $q_n^{(a)}$). Those transitions probabilities are such that after a site is searched and the *unhunted* target on it is not detected, it is more likely that the target moves into or remains in the hidden state than if the site had not been searched, i.e., $q_n^{(1)} > q_n^{(0)}$ and $p_n^{(1)} > p_n^{(0)}$. Notice that such condition ensures also that after a site is not searched, it is more likely that the target moves into or remains in the exposed state than if the site had been searched. We further assume that the visibility state processes have positive autocorrelation or memory, so $\rho_n^{(a)} \triangleq 1 - p_n^{(a)} - q_n^{(a)} > 0$. Figure 1 illustrates the above description.

The target at site n can only be hunted if it is exposed when searched, yielding a reward r_n for completing the site's mission. Information on target n's visibility state is gained by sensing it, which provides a *sensor outcome* $o_{n,t} \in \{0, 1\} : o_{n,t} = 1$ if the target is detected and hunted, and $o_{n,t} = 0$ otherwise. Sensing is imperfect in that the target at site n will not be detected when it is exposed and its site was sensed with a positive *misdetection* probability of $\alpha_n = P(o_{n,t} = 0|s_{n,t} = 1)$. Hence, target n's visibility state $s_{n,t}$ is not directly observable, but it is tracked by the *information state* $X_{n,t} \in \mathbb{X} \triangleq [0, 1]$, giving the posterior probability that the target is *exposed* in period t conditioned on the history $\{X_{n,s}, a_{n,s} : 0 \leq s < t\} \cup X_{n,t}$.
Since successfully hunting a target completes the mission at its site, we assume that a site n whose target has been hunted ($x_n = 0$) is removed from further search. Hence, we partition a target state space $\mathbb{X}$ into the set $\mathbb{X}^{0,1} \triangleq (0, 1]$ of controllable states, where both actions $A \triangleq \{0, 1\}$ are available, and the uncontrollable state 0, where only action $a_n = 0$ is available.

The dynamics of the information state for target n under each sensing action are obtained via Bayesian updates as follows. If the target at site's n has not yet been hunted at the beginning of period t, i.e. $X_{n,t} > 0$, and the site is searched ($a_{n,t} = 1$), then its next state will depend on wether the search was successful or not. Thus, if the sensor outcome is positive ($o_t^n = 1$), which happens with probability (w.p.) $(1-\alpha_n)X_{n,t}$, and the target is detected $o_{n,t} = 1$, which happens with probability $(1 - \alpha_n)X_{n,t}$, then the target has been hunted and hence site n is removed from the search objectives. We model such a situation by letting the target's information state drop to zero, i.e. $X_{n,t+1} = 0$.
On the other hand, if the target is not detected $o_{n,t} = 0$, which happens with probability

$1-(1-\alpha_n)X_{n,t}$, it is readily calculated that the information state changes to

$$X_{n,t+1} = p_n^{(1)} + \rho_n^{(1)} \left(\frac{\alpha_n X_{n,t}}{1-(1-\alpha_n)X_{n,t}} \right). \tag{1}$$

Hence, when site n is searched, its next information state is obtained in a randomized fashion depending on the sensing outcome.
Finally, if site n is not sensed ($a_{n,t} = 0$) in period t, with its information state being $X_{n,t} > 0$, i.e., as long as the target has not been hunted yet, its next information state is determined by

$$X_{n,t+1} = p_n^{(0)}(1 - X_{n,t}) + (1 - q_n^{(0)})X_{n,t}. \tag{2}$$

Yet, if the target has already been hunted $X_{n,t} = 0$, then its information state remains at 0 under both sensing actions. Thus, we summarize the information state dynamics for all controllable states $X_{n,t} \in \mathbb{X}^{0,1}$ as

$$X_{n,t+1} = \begin{cases} p_n^{(0)}(1 - X_{n,t}) + (1 - q_n^{(0)})X_{n,t}, & \text{if } a_{n,t} = 0 \quad \text{w.p } 1, \\ 0, & \text{if } a_{n,t} = 1 \quad \text{w.p } (1-\alpha_n)X_{n,t}, \\ p_n^{(1)} + \rho_n^{(1)} \left(\dfrac{\alpha_n X_{n,t}}{1-(1-\alpha_n)X_{n,t}} \right), & \text{if } a_{n,t} = 1 \quad \text{w.p } 1-(1-\alpha_n)X_{n,t}, \end{cases}$$

Sensing actions are prescribed by a *scheduling policy* drawn from the class of admissible policies $\mathbf{\Pi}(M)$, consisting of the nonanticipative policies (i.e., based on the history of states and actions) that search at most M sites per slot:

$$\sum_{n=1}^{N} a_{n,t} \leq M, \quad t = 0, 1, \ldots. \tag{3}$$

As for the economic results of sensing actions, taking action a_n on site n when it occupies the information state x_n yields the *expected one-slot net reward* $R_n(x_{n,t}, a_{n,t}) \triangleq (r_n\,(1-\alpha_n)\,x_n - c_n)\,a_n$, where $c_n \geq 0$ is a fixed site/target specific sensing cost.

The sensing system described by this model operates over time slots of equal length, assuming sensors are synchronized to operate over discrete time slots. The sequence of events within each slot is described in Figure 2. At the beginning of each slot, the system's manager given site's n current *information state* $X_{n,t}$, decides whether to sense that site or not, afterwards earning an expected reward $R_n(x_{n,t}, a_{n,t})$ which depends on the selected action and the current belief state. Afterwards target's n, if not hunted, changes its visibility state depending on the selected sensing action and hence by the end of the slot, site's n *belief state* is updated accordingly.

2.1 Performance Objectives

We will consider the following dynamic optimization problem: find a *β-discounted reward optimal* policy, i.e.,

$$\max_{\boldsymbol{\pi} \in \mathbf{\Pi}(M)} \mathsf{E}_{\mathbf{x_0}}^{\boldsymbol{\pi}} \left[\sum_{t=0}^{\infty} \sum_{n=1}^{N} \beta^t R_n\big(X_{n,t}, a_{n,t}\big) \right], \tag{4}$$

where $0 < \beta \leq 1$ is the discount factor, $\mathbf{x}_0 = (x_{n,0})_{n=1}^N$ is the initial joint belief state, for n in $\{1, 2, \ldots, N\}$, and $\mathsf{E}_{\mathbf{x_0}}^{\boldsymbol{\pi}}[\cdot]$ denotes expectation under policy $\boldsymbol{\pi}$ conditioned on the initial

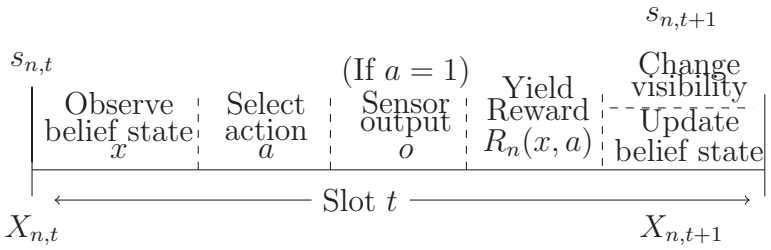


Figure 2: The sequence of events within a time slot for the elusive target hunt model.

joint state being equal to $\mathbf{x}_0$. Note that the undiscounted case $\beta = 1$, which corresponds to the *total expected reward* criterion, is well defined in the present setting given that the search plan terminates after a finite number of slots with probability one (but the number of slots until termination, i.e., the horizon, is uncertain and unbounded). Furthermore, when considering a discount factor $\beta = 1$ we may analize the case in which there is interest in finding targets regardless of how long it takes to do so. When there are reasons to penalise finding targets in a later futute, such as system's lifetime constraints, it makes sense to consider some $\beta < 1$.

Problem (4) is a POMDP of restless MARBP type, thus being notoriously hard to solve exactly. In the following section we shall present the results of deploying the real-state restless bandit Whittle's MP indexation approach to the model of concern.

3 MARBP formulation and Indexation

We will deploy the approach developed and applied in other real-state multi-armed restless bandit models in [6,7,9]. The following discussion reviews the key methodological aspects of such an approach.

3.1 Relaxed Problem, Lagrangian Relaxation and Performance Bound

Along the lines introduced in [13] for the equality-constrained case, we first construct a relaxation of (4), relaxing the hard sample path *peak resource-usage* constraint (3) by the averaged version that the expected total discounted (ETD) number of sensed sites does not exceed $M/(1-\beta)$, i.e.,

$$\mathsf{E}^{\boldsymbol{\pi}}_{\mathbf{x_0}}\left[\sum_{t=0}^{\infty}\sum_{n=1}^{N}\beta^t a_{n,t}\right] \leq \frac{M}{1-\beta}. \tag{5}$$

Denoting by Π the class of nonanticipative scheduling policies (which may sense any number of sites at any time), the *relaxed primal problem* is

$$\max_{(5),\boldsymbol{\pi}\in\boldsymbol{\Pi}} \mathsf{E}^{\boldsymbol{\pi}}_{\mathbf{x_0}}\left[\sum_{t=0}^{\infty}\sum_{n=1}^{N}\beta^t R_n\big(X_{n,t}, a_{n,t}\big)\right]. \tag{6}$$

Note that the optimal value of (6) $V^{\mathrm{R}}(\mathbf{x_0})$ gives an *upper bound* on the optimal value of (4) $V^*(\mathbf{x_0})$.

To address such a constrained MDP (6) we next deploy a Lagrangian approach, including coupling constraint (5) and attaching a multiplier $\lambda \geq 0$ to it. The resulting problem

$$\max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \mathsf{E}^{\boldsymbol{\pi}}_{\mathbf{x_0}} \left[\sum_{t=0}^{\infty} \sum_{n=1}^{N} \beta^t \left\{ R_n\big(X_{n,t}, a_{n,t}\big) - \lambda a_{n,t} \right\} \right] + \frac{M\lambda}{1-\beta} \tag{7}$$

is a *Lagrangian relaxation* of (6), whose optimal value $V^{\mathrm{L}}(\mathbf{x_0};\lambda)$ gives an upper bound on $V^{\mathrm{R}}(\mathbf{x_0})$. The *Lagrangian dual problem* is to find an optimal value $\lambda^*(\mathbf{x_0})$ of λ giving the best upper bound on $V^{\mathrm{R}}(\mathbf{x_0})$, which we denote by $V^{\mathrm{D}}(\mathbf{x_0})$:

$$V^{\mathrm{D}}(\mathbf{x_0}) = \min_{\lambda\geq 0} V^{\mathrm{L}}(\mathbf{x_0};\lambda) \tag{8}$$

Note that $\lambda^*(\mathbf{x_0})$ solves (8) which is a scalar convex optimization problem, since $V^{\mathrm{L}}(\mathbf{x_0};\lambda)$ is concave in λ.

3.2 Indexability and Whittle's Index Policy

Next, given the fact that target's state transitions are independent, we *decompose* problem (7) as

$$V^{\mathrm{L}}(\mathbf{x_0};\lambda) = \sum_{n=1}^{N} V_n^{\mathrm{L}}(\mathbf{x_{n,0}};\lambda) + \frac{M\lambda}{(1-\beta)}, \tag{9}$$

where each $V_{(n)}^{\mathrm{L}}(\mathbf{x_0};\lambda)$ is a single-project restless bandit subproblem, consisting of the following hunting problem considered for some site n in isolation,

$$\max_{\boldsymbol{\pi_n}\in\boldsymbol{\Pi_n}} \mathsf{E}^{\boldsymbol{\pi}^{(n)}}_{\mathbf{x_{n,0}}} \left[\sum_{t=0}^{\infty} \beta^t \{ R_n\big(X_{n,t}, a_{n,t}\big) - \lambda a_{n,t} \} \right], \tag{10}$$

where Π_n denotes the class of admissible policies for operating a single sensor on such site, i.e., deciding when it should be active ($a_{n,t}=1$) and passive ($a_{n,t}=0$) and with λ being a constant parameter representing an extra cost incurred per unit of time the sensor is active. In terms of these individual problems, multiplier λ represents an *additional cost*, to be added to the site's regular sensing cost c_n, that will be paid per time slot a sensor is searching site n.

The following defines a key structural property of such restless bandit subproblems, termed *indexability* by Whittle in [13].

Definition 1 The *single-site* hunting subproblem (10) is said to be *indexable* if there exists an *index* $\lambda^*(x_n)$ which is a scalar function of the site's information state $x_n \in \mathbb{X}^{0,1}$ such that, for any value of the cost $\lambda \in \mathbb{R}$, the active action $a_{n,t}=1$ (sensing the site) is optimal in state $X_{n,t}=x_n$ iff $\lambda^*(x_n) \geq \lambda$, regardless of the initial state.

If definition 1 holds for each subproblem, then the resulting index can be used as site's n sensing-priority to define a heuristics for problem (4). Clearly, by decoupling the whole problem into n individual subproblems, (10) is significantly easier to solve than (4), yet its computational tractability depends on that of individual subproblems.

3.3 Sufficient Indexability Conditions and Index Evaluation

Whittle's indexability is a structural property from which a tractable and generally well-performing priority index policy can be derived, yet it needs to be established for each model at hand. The introduction of sufficient indexability conditions for discrete-state restless bandits based on satisfaction on *partial conservation laws* (PCLs), along with an index algorithm, reviewed in [5], provided with a methodology for such a purpose. Such conditions were extended to continuous-state restless bandits in [6], as reviewed next.

In this section we focus on a generic single site/target subproblem, and hence drop the superscript n from the above notation. We will evaluate the performance of admissible sensing policies $\pi \in \Pi$ along two dimensions: the *work measure* $g(x,\pi)$, giving the ETD number of times a site is sensed under policy π starting at $X_0 = x$; and the *reward measure* $f(x,\pi)$, giving the corresponding ETD reward earned. Thus,

$$g(x,\pi) \triangleq \mathsf{E}_x^\pi \left[\sum_{t=0}^{\infty} \beta^t a_t\right], \quad f(x,\pi) \triangleq \mathsf{E}_x^\pi \left[\sum_{t=0}^{\infty} \beta^t R(X_t, a_t)\right],$$

We can then formulate the single-site's optimal target hunting subproblem (10) as

$$V^*(x;\lambda) = \max_{\pi\in\Pi} f(x,\pi) - \lambda g(x,\pi). \tag{11}$$

Problem (11), is a continuous-state Markov Decision Process (MDP), whose optimal policy, under certain assumptions on $R(x,a)$ (e.g. that it is a bounded and measurable function), belongs to the family of *deterministic stationary policies*, naturally represented by their *active (state) sets*, i.e., the set of information states where the active action (in this case, sensing the site) is prescribed. For an active set $B \subseteq \mathbb{X}^{0,1}$, we shall refer to the *B-active policy*.

We will further focus attention on the family of *threshold policies*. For a given *threshold level* $z \in \mathbb{R}$, the *z-threshold policy* senses the site in information state x iff $x > z$, so its active set is $B(z) \triangleq \{x \in \mathbb{X}^{0,1} \colon x > z\}$. Note that $B(z) = (z,1]$ for $0 \le z < 1$, $B(z) = \mathbb{X}^{0,1} = (0,1]$ for $z < 0$, and $B(z) = \emptyset$ for $z \ge 1$. We denote by $g(x,z)$ and $f(x,z)$ the corresponding work and reward measures under a z-threshold policy.

In the following we will use the notation

$$\phi^{(0)}(x) \triangleq (p^{(0)} + \rho^{(0)} x), \quad \phi^{(1)}(x) \triangleq p^{(1)} + \rho^{(1)} \frac{\alpha x}{1-(1-\alpha)x}. \tag{12}$$

For some fixed z, the total work measure $g(x,z)$ for any $x \in \mathbb{X}^{0,1}$ is characterized by the the unique solution to

$$g(x,z) = \begin{cases} 1 + \beta\ [1-(1-\alpha)\,x\,]\, g\big(\phi^{(1)}(x), z\big), & x > z \\ \beta g\big(\phi^{(0)}(x), z\big), & x \le z \\ 0 & x = 0, \end{cases} \tag{13}$$

in the Banach space of Borel measurable bounded functions, endowed with the sup norm. See [1]. Whereas the total reward measure $f(x,z)$ for any $x \in \mathbb{X}^{0,1}$ is characterized by

$$f(x,z) = \begin{cases} R(x,1) + \beta\ [1-(1-\alpha)\,x\,]\, f\big(\phi^{(1)}(x), z\big), & x > z \\ \beta f\big(\phi^{(0)}(x), z\big), & x \le z \\ 0 & x = 0. \end{cases} \tag{14}$$

We will further use the marginal counterparts of such total evaluation measures. For threshold any fixed z and action a, denote by $\langle a, z\rangle$ the policy that takes action a in the initial period and adopts the z-threshold policy thereafter. Define the *marginal work measure* $w(x, z)$ and the *marginal reward measure* $r(x, z)$ as

$$\begin{aligned} w(x,z) &\triangleq g(x,\langle 1,z\rangle) - g(x,\langle 0,z\rangle), \qquad (15)\\ &= 1+\beta\ [1-(1-\alpha)\,x\,]\ g(\phi^{(1)}(x),z)\\ &\quad -\beta\, g(\phi^{(0)}(x),z)\\ r(x,z) &\triangleq f(x,\langle 1,z\rangle) - f(x,\langle 0,z\rangle),\\ &= R(x,1)+\beta\ [1-(1-\alpha)\,x\,]\ f(\phi^{(1)}(x),z)\\ &\quad -\beta\, f(\phi^{(0)}(x),z) \qquad (16)\end{aligned}$$

If $w(x, z) \neq 0$, define the *marginal productivity (MP) measure*

$$\lambda^{MP}(x,z) \triangleq \frac{r(x,z)}{w(x,z)}. \qquad (17)$$

We will invoke the following definition and theorem introduced in [6] .

Definition 2 Subproblem (11) is *PCL-indexable* (with respect to threshold policies) if:

(i) *positive marginal work*: $w(x, z) > 0, x \in \mathbb{X}^{0,1}, z \in \mathbb{R}$;

(ii) *nondecreasing index*: the index defined by

$$\lambda^{MP}(x) \triangleq \lambda^{MP}(x,x), \quad x \in \mathbb{X}^{0,1}. \qquad (18)$$

is monotone nondecreasing in x and continuous

Theorem 1 *If subproblem* (11) *is PCL-indexable, then it is indexable and the MP index* $\lambda^{MP}(x)$ *in* (18) *is its Whittle's index* $\lambda^*(x)$.

4 PCL-Indexability Analysis and MP Index Computation

4.1 Verification of PCL-indexability

As reviewed in subsection 3.3, establishing Whittle's indexability 1 by means of deploying sufficient indexability conditions 2 we focus on an individual site's subproblem (11), which is a single-project restless bandit subproblem, consisting of a hunting problem considered for some site n in isolation. Π_n denotes the class of admissible policies for operating a single sensor on such site, i.e., deciding when it should be active ($a_{n,t} = 1$) and passive ($a_{n,t} = 0$) and with λ being a constant parameter representing an extra cost incurred per unit of time the sensor is active.

Next, we would like to establish that each subproblem (11) has the key structural *indexability* property defined by 1. For such a purpose, we will deploy conditions 2 to establish that the problem is indexable with respect to the family of z-threshold policies, and thus we start by computing the performance measures under such class of policies

(13) and (14). In the remainder of this section we focus on a generic single site/target subproblem as (11), and hence drop the superscript n from the above notation.

We recall that problem (11), is a continuous-state Markov Decision Process (MDP), whose optimal policy belongs to the family of *deterministic stationary policies* Π^{SD}, naturally represented by their *active (state) sets* (in this case, that is the set of information states where sensing the site is prescribed). For an active set $B \subseteq \mathbb{X}^{0,1}$, we shall refer to the *B-active policy*.

The following section outlines how to solve the evaluation equations to perform a PCL-indexability analysis, and further shows how to use such solutions to compute in practice the index (18).

Total and Marginal Evaluation Measures

In order to analyze the PCL-indexability of (11) by means of the sufficient indexability conditions (SIC) stated in Theorem 2 we must first solve the evaluation measures $g(x, z)$ and $f(x, z)$ for any fixed threshold $z \in \mathbb{R}$ and any $x \in \mathbb{X}^{0,1}$.

In order to do so, we must address the problem posed by the fact that possible information state trajectories $\{X_t\}$ are naturally infinite, since X_t can take any value in $\mathbb{X}$ at each t. To do so, we will take advantage of the fact that under a z-threshold policy for any initial state x, possible information state trajectories $\{X_t\}$ are infinite but numerable, as they exhibit recurrent cyclical patterns depending on the threshold level. Yet, as we will next show, the total performance measures do not converge to a simple closed-form expression. In the cases in which such measures can be solved in closed form, as e.g. [6], both direct verification of the SIC and obtaining a closed-form index formula are possible. Yet, in the model addressed in this paper, a significant challenge to establish indexability and to derive an index policy, is to do so despite the fact that the evaluation equations do not admit a straightforward algebraic manipulation.

Next, we outline how to solve the evaluation measures to perform an indexability analysis and further shows how to use such solutions to evaluate the index $\lambda^*(x)$ and to establish the PCL-indexability of the model.

To solve for (13) and (14) in closed form we further define $\phi_t^{(a)}(x)$ for $a = 0, 1$ as the recursion generated by letting $\phi_0^{(a)}(x) \triangleq x$ and $\phi_t^{(a)}(x) \triangleq \phi_0^{(a)}(\phi_{t-1}^{(a)}(x))$ for $a = 0, 1$. Note that for any $x \in \mathbb{X}^{0,1}$, both recursions $\phi_t^{(a)}(x)$ converge as $t \to \infty$ to the respective limits

$$\phi_\infty^{(0)} \triangleq \frac{p^{(0)}}{1-\rho^{(0)}} \qquad \phi_\infty^{(1)} \triangleq \frac{\gamma - \sqrt{\gamma^2 - 4p^{(1)}(1-\alpha)}}{2(1-\alpha)}$$

with $\gamma \triangleq 1 - \rho^{(1)} + (p^{(1)} + \rho^{(1)})(1-\alpha)$.

Most importantly, notice that both functions (1) and (2) and their resulting iterated mappings can be seen as (non-linear) functions known as Möbius transformations (or also as Linear Fractional Transformations). This observation is crucial to the subsequent indexability analysis.

The definitions of (1) and (2) ensure that that $\phi_\infty^{(1)} < \phi_\infty^{(0)}$, and both limits are attractive fixed points of the active and passive dynamics respectively. This naturally divides the state space into three parts, where the active and passive actions (depending on the initial information state x and the threshold z) produce movements in the state space, which are either both increasing (if $x, z \in (0, \phi_\infty^{(1)})$) or both decreasing (if $x, z \in [\phi_\infty^{(0)}, 1]$) or moving in opposite directions (if $x, z \in [\phi_\infty^{(1)}, \phi_\infty^{(0)})$). Hence, we exploit this knowledge

to solve the evaluation equations by distinguishing among three $z-$threshold cases, as discussed below.
In the sequel we assume, without loss of generality, that $c = 0$.

4.1.1 Case I: Threshold $z \in [0, \phi_\infty^{(1)})$ (*Low thresholds*)

In this case, the active set $B(z) = (z, 1]$ contains the attractive fixed points of the recursions associated to both actions, i.e. $\phi_\infty^{(0)}, \phi_\infty^{(1)}$. This implies that once the state reaches the active set $B(z)$ it stays in $B(z)$ as long as the target remains unhunted. For any $x \geq \phi_\infty^{(1)}$ then $\phi_t^{(1)}(x) \geq \phi_\infty^{(1)} > z$ for all $t \geq 0$. Further, for $z \in B^c(z) \triangleq [0, z]$: $\phi_t^{(0)}(x) \nearrow \phi_\infty^0$. Hence, after a finite number of passive slots $t_0^*(x, z) < \infty$: $\phi_{t_0^*(x,z)}^{(0)}(x) > z$, where we define the first (deterministic) hitting time to the active set as $t_0^*(x, z) \triangleq \min\{t \geq 1 : \phi_t^{(0)}(x) > z\}$
Also, denote by $\theta(x, z, t)$ the survival probability representing the probability that the target has not been hunted before time slot t under the z-threshold policy, starting from state x. Note that, for $x > z$

$$\theta(x, z, t) \triangleq \prod_{s=0}^{t-1} \left[1 - (1 - \alpha)\, \phi_s^{(1)}(x)\right] \tag{19}$$

where we let $\theta(x, z, 0) = 1$. Thus, the total work measure has the following evaluation:

$$g(x, z) = \begin{cases} \displaystyle\sum_{t=0}^{\infty} \beta^t\, \theta(x, z, t) & x \in (z, 1] \\ \beta^{t_0^*(x,z)} \left[\displaystyle\sum_{t=0}^{\infty} \beta^t\, \theta(y, z, t)\right] & x \in (0, z] \end{cases} \tag{20}$$

where $y \triangleq \phi_{t_0^*(x,z)}^{(0)}(x)$.
Similarly, we obtain the total reward evaluation

$$f(x, z) = \begin{cases} \displaystyle\sum_{t=0}^{\infty} \beta^t\, \theta(x, z, t)\, R(\phi_t^{(1)}(x, z), 1) & x \in (z, 1] \\ \beta^{t_0^*(x,z)} \displaystyle\sum_{t=0}^{\infty} \beta^t\, \theta(y, z, t)\, R(\phi_t^{(1)}(y, z), 1) & x \in (0, z] \end{cases} \tag{21}$$

The above infinite series are convergent, yet they do not admit closed form formulae. Hence, they must be truncated in practice to evaluate $w(x, z)$ and $r(x, z)$ via (15) and (16), and hence also for establishing that SIC conditions i) and ii) in 2 hold.

In the following we list our main results, drawing on the technical analysis of the marginal work measure that will be included in a full version of this paper, currently under preparation.We define β^* as the discount factor β such that:

$$\left(\sum_{t=0}^{\infty} (\beta^*)^t \theta(x, z, t) - \beta^* \sum_{t=0}^{\infty} (\beta^*)^t \theta(\phi^{(0)}(x), z, t)\right) = 0 \tag{22}$$

We further define $\beta^{(1)}$ as:

$$\beta^{(1)} \triangleq \frac{1}{1 + \left[1 - (1 - \alpha)(1 - \phi_\infty^{(1)})\right]}.$$

Proposition 3 *The marginal work measure $w(x,z)$ in problem* (11) *with $x \in \mathbb{X}^{0,1}$ and $z \in [0, \phi_\infty^{(1)})$ is strictly positive for $\beta < \beta^*$ with $\beta^* > \beta^{(1)}$.*

The strategy deployed for proving the positivity of marginal work measures in all the threshold cases of concern, despite the lack of a closed form formulae, is the following: for each z-threshold case and every possible initial state x, based on properties of the active and passive recursions as Möbius transformations, we derive a lower bound on $w(x,z)$ and then study its positivity (or the conditions under which its positivity it is ensured).

The proof of Proposition 3 is based on the following lemma which states lower bounds on $w(x,z)$ for this threshold case.

Lemma 4 *For all $z < \phi_\infty^{(1)}$,*

(a) $w(x,z) > \min\{1 - \frac{\beta\,(1-\alpha)\,x}{(1-\beta)+\beta\,(1-\alpha)\,z}, 1-\beta\} \geq 0$

for any $x \in (0,z]$, $\quad 0 \leq \beta \leq 1$.

(b) $w(x,z) > (1-\beta) \geq 0 \quad$ *for any* $x \in (z, \phi_\infty^{(0)}]$, $\quad 0 \leq \beta \leq 1$.

(c) $w(x,z) > 0 \quad$ *for any* $x \in (\phi_\infty^{(0)}, 1]$, *only if* $\beta < \beta^*$,

where β^* is defined as the discount factor β such that:

$$\left(\sum_{t=0}^{\infty}(\beta^*)^t\theta(x,z,t) - \beta^*\sum_{t=0}^{\infty}(\beta^*)^t\theta(\phi^{(0)}(x),z,t)\right) = 0 \tag{23}$$

As there is no closed form expression for those infinite sums, β^* cannot be computed exactly. $\beta^{(1)}$ is a lower bound on it obtained by imposing that the lowest bound on $w(x,z)$ for $x = 1$ is strictly positive. Further bounds can be obtained by truncation of the infinite sums in (23).

Proposition 3 ensures that condition (i) in the SIC holds for this case. Regarding the monotonicity condition of the index, first notice that it follows from the definition of $t_0^*(x,z)$ that: $t_0^*(\phi^{(1)}(x),x) = t_0^*(\phi^{(0)}(x),x) = 0$ given that $\phi^{(0)}(x) > x$ and $\phi^{(1)}(x) > 0$,which allows us to compute the index (17) for case I as follows:

$$\lambda^{MP}(x) = \frac{R\,(1-\alpha)\left[\sum_{t=0}^{\infty}\beta^t\left[\phi_t^{(1)}(x)\,\theta(x,x^-,t) - \beta\phi_t^1(\phi^{(0)}(x))\,\theta(\phi^{(0)}(x),x,t)\right]\right]}{\sum_{t=0}^{\infty}\beta^t\left[\theta(x,x^-,t) - \beta\,\theta(\phi^{(0)}(x),x,t)\right]}, \quad x \in (0,\phi_\infty^{(1)}) \tag{24}$$

where x^- stands for the sensing policy with active set equal to $B(x^-) = [x,1]$.

Next, to ensure indexability we must prove that this index is nondecreasing with respect to the information state. Notice, that for all $x \in [0,\phi_\infty^{(1)})$ the $\lambda^{MP}(x)$ is an infinite sum of continuous functions of the state. For such a purpose, we take the derivatives of the two infinite sums defining the index with respect to x. Since, there is no closed form formulae for those sums to manipulate it algebraically, the strategy to accomplish such a goal is to show that, provided continuity of the MP index is ensured, it holds that a) $\frac{\partial w(x,x)}{\partial x} < 0$ and b) $\frac{\partial r(x,x)}{\partial x} > 0$ by manipulating the derivative of each term in the infinite sum. Showing continuity of the index in this case calls for further research, but the experimental evidence suggests it holds.

Proposition 5 *The index* $\lambda^{MP}(x) = \frac{r(x,x)}{w(x,x)}$ *as defined in* (24) *for problem* (11) *is monotone increasing and a continuous in the information state* x *for* $x \in (0, \phi_\infty^{(1)})$.

The proof of proposition 5 is based on the following lemma.

Lemma 6 *For all* $x < \phi_\infty^{(1)}$, *it holds that:*

$$\sum_{t=1}^{\infty} \beta^t \left[\frac{\partial \theta(x, x^-, t)}{\partial x} - \beta \frac{\partial \theta(\phi^{(0)}(x), x, t)}{\partial x} \right] < 0 \tag{25}$$

$$\sum_{t=0}^{\infty} \beta^t \left[\frac{\partial \phi_t^{(1)}(x)\, \theta(x, x^-, t)}{\partial x} - \beta \frac{\partial \phi_t^{(1)}(\phi^{(0)}(x))\, \theta(\phi^{(0)}(x), x, t)}{\partial x} \right] > 0 \tag{26}$$

4.1.2 Case II: Threshold $z \in [\phi_\infty^{(1)}, \phi_\infty^{(0)})$ (*Intermediate thresholds*)

In this case, the passive set $B^c(z)$ contains the attractive fixed point of the recursion associated to the active action, i.e. $\phi_\infty^{(1)}$, whereas the active set $B(z)$ contains the attractive fixed point of the recursion associated to the passive action, i.e. ϕ_∞^0. Hence, the state X_t jumps above and below the threshold z, until the target is found. Following the argument introduced in [7], define the map $\phi(x, z) \triangleq 1_{x>z}\phi^{(1)}(x) + 1_{x\leq z}\phi^{(0)}(x)$, and let $\phi_0(x, z) = x$, $\phi_t(x, z) = \phi(\phi_{t-1}(x, z), z)$ for $t \geq 1$. Then, writing $a_t(x, z) \triangleq 1_{\phi_t(x,z)>z}$, $(\phi a)_t(x, z) \triangleq \phi_t(x, z)a_t(x, z)$. In this case, the survival probability has evaluation

$$\theta(x, z, t) \triangleq \prod_{s=0}^{t-1} [1 - (1 - \alpha)\, (\phi a)_s(x, z)] \tag{27}$$

with $\theta(x, z, 1) = 0$. Thus, total evaluation measures admit the following expressions

$$g(x, z) = \sum_{t=0}^{\infty} \beta^t\, a_t(x, z)\theta(x, z, t) \tag{28}$$

$$f(x, z) = \sum_{t=0}^{\infty} \beta^t\, R(\, (\phi a)_t(x, z), 1\,)\, \theta(x, z, t) \tag{29}$$

In this case also, since the expressions (28) and (29) cannot be calculated in a closed form, truncation is neccesary for evaluating them numerically. However, we are able to describe a *recurrent cyclical* pattern in the resulting information state X_t process under a z-threshold policy, which allows us to describe the possible trajectories of the information state to be considered. Specifically, using properties of the Möbius Transformations we are able to establish regularities , in terms of the sequence of active and passive slots until a target is hunted, that allow us to derive the corresponding lower bounds on $w(x, z)$ for this case, which is the most complex of the three threshold cases. In the following we list the main results for this case.

Proposition 7 *The marginal work measure* $w(x, z)$ *in problem* (11) *with* $x \in \mathbb{X}^{0,1}$ *and* $z \in [\phi_\infty^{(1)}, \phi_\infty^{(0)})$ *is positive for* $\beta < \beta^*$ *with* $\beta^* > \beta^{(1)}$.

The proof of Theorem 3 is based on the following lemma, providing lower bounds on $w(x,z)$.

Lemma 8 *For all* $z \in [\phi_\infty^{(1)}, \phi_\infty^0)$,

(a) $w(x,z) > \min\{1 - \frac{\beta\,(1-\alpha)\,x}{(1-\beta)+\beta\,(1-\alpha)\,\phi_\infty^{(1)}}, 1-\beta\} \geq 0$
for any $x \in (0,z]$, $\quad 0 \leq \beta \leq 1$.

(b) $w(x,z) > (1-\beta) \geq 0 \quad$ *for any* $x \in (z, \phi_\infty^0]$, $\quad 0 \leq \beta \leq 1$.

(c) $w(x,z) > 0 \quad$ *for any* $x \in (\phi_\infty^0, 1]$ *only if* $\beta < \beta^*$

Theorem 7 ensures that condition (i) in the SIC holds for this case. Further, Theorem 7 implies that Theorem 3 holds in this threshold case also. Next, we compute the index (17) in this case, using the fact that $t_0^*(x,x) = 1$ and, given that $\phi^{(1)}(x) < x < \phi^{(0)}(x)$, as follows:

$$\lambda^{MP}(x) = \frac{\displaystyle\sum_{t=0}^{\infty} \beta^t R\,(1-\alpha)\left[(\phi a)_t(x,x^-)\theta(x,x^-,t) - \beta(\phi a)_t(\phi^{(0)}(x),z)\,\theta(\phi^{(0)}(x),x,t)\right]}{\displaystyle\sum_{t=0}^{\infty} \beta^t \left[\,\theta(x,x^-,t)a_t(x,x^-) - \beta\,\theta(\phi^{(0)}(x),x,t)a_t(\phi^{(0)}(x),z)\right]}, \tag{30}$$

for $x \in [\phi_\infty^{(1)}, \phi_\infty^0)$, where x^- stands for the sensing policy with active set equal to $B(x^-) = [x,1]$.
Such an index can be expressed as an infinite sum of functions defined by a composition of the two the Möbius transformations describing the active and passive dynamics, depending on the concrete cycle that a given threshold x generates. Showing continuity of the index in this case calls for further research, but the experimental evidence suggests it holds.

Proposition 9 *The index* $\lambda^{MP}(x) = \frac{r(x,x)}{w(x,x)}$ *as defined in 30 for problem* (11) *is monotone increasing and a continuous in the information state* x *for* $x \in [\phi_\infty^{(1)}, \phi_\infty^{(0)})$.

Lemma 10 *For all* $x < \phi_\infty^{(1)}$, *it holds that:*

$$\sum_{t=1}^{\infty} \beta^t \left[\frac{\partial\theta(x,x^-,t)a_t(x,x^-)}{\partial x} - \beta\frac{\partial\theta(\phi^{(0)}(x),x,t)a_t(\phi^{(0)}(x),z)}{\partial x}\right] \;<\; 0 \tag{31}$$

$$\sum_{t=0}^{\infty} \beta^t \left[\frac{\partial(\phi a)_t(x,x^-)\theta(x,x^-,t)}{\partial x} - \beta\frac{\partial(\phi a)_t(\phi^{(0)}(x),x)\,\theta(\phi^{(0)}(x),x,t)}{\partial x}\right] \;>\; 0 \tag{32}$$

4.1.3 Case III: Threshold $z \in [\phi_\infty^{(0)}, 1]$ (*High thresholds*)

In this case, the passive set $B^c(z)$ contains the attractive fixed points of the recursions associated to both actions, i.e. $\phi_\infty^{(0)}, \phi_\infty^{(1)}$. This, in turn, implies that once the information state reaches the passive set $B^c(z)$, it remains in it, regardless if the target has been hunted or not at that moment of time. For all $x > z$, $z \geq \phi_\infty^{(0)} \geq \phi_t^{(0)}(x)$ for all $t \geq 0$. Further, for $z \in [\phi_\infty^{(0)}, 1]$ and for $x > z$: $\phi_t^{(1)}(x) \to \phi_\infty^{(1)}$. Hence, after a finite number of active slots

$\tau^*(x,z) < \infty$, with $\tau^* \triangleq \min\{t \geq 1 : X_t \leq z\}$,: $\phi^1_{\tau^*}(x) \leq z$. Notice that $\tau^*(x,z)$ for some $x > z$ is a random variable with maximum value $t^*_1(x,z) \triangleq \min\{t \geq 1 : \phi^{(1)}_t(x) \leq z\}$. Then, we have that

$$g(x,z) = 1_{\{x>z\}} \left[\sum_{t=0}^{t^*_1(x,z)-1} \beta^t \theta(x,z,t) \right], \tag{33}$$

$$f(x,z) = 1_{\{x>z\}} \left[\sum_{t=0}^{t^*_1(x,z)-1} \beta^t \, R(\, \phi^{(1)}_t(x), 1\,) \theta(x,z,t) \right]. \tag{34}$$

where $\theta(x,z,t)$ is the survival probability as defined in Case I. For $x > z$, equations (33) and (34) are readily computed by evaluating finite sums with $t^*_1(x,z) - 1$ terms.

Proposition 11 *The marginal work measure $w(x,z)$ in problem* (11) *with $x \in \mathbb{X}^{0,1}$ and $z \in [\phi(0)_\infty, 1)$ is positive for $\beta < \beta^*$ with $\beta^* > \beta^{(1)}$.*

The proof of Proposition 11 is based on the following lemma, providing lower bounds on $w(x,z)$.

Lemma 12 *For all $z \geq \phi^0_\infty$,*

(a) $w(x,z) = 1$ for any $x \in (0,z]$, $\quad 0 \leq \beta \leq 1$.

(b) $w(x,z) > 0 \quad$ for any $x \in (z,1]$ for $\beta < \beta^*$.

Hence, for $x \leq z$ it is readily seen that $w(x,z) = 1$ and $r(x,z) = R(x,1)$. Therefore, the index in (18)

$$\lambda^{MP}(x) = R(x,1), \quad \phi^{(0)}_\infty \leq x \leq 1 \tag{35}$$

Proposition 13 *The index $\lambda^{MP}(x) = \frac{r(x,x)}{w(x,x)}$ as defined in 35 for problem* (11) *is a continuous and monotone increasing in the information state x for $x \in (\phi^{(0)}_\infty, 1]$.*

Proof:
Taking partial derivative to index (35), it follows that:

$$\frac{\partial \lambda^{MP}(x)}{\partial x} = \frac{\partial R(x,1)}{\partial x} = r(1-\alpha) > 0.$$

Notice that in case III, the MP index $\lambda^{MP}(x)$ coincides with the myopic index $\lambda^{myopic}(x)$, which results from optimizing the one-period expected reward.

4.1.4 Verification of PCL-indexability Sufficient Conditions

Based on propositions 3-13, we conclude:

Theorem 14 *The single-site elusive target hunt ETD problem* (11) *is PCL-indexable for $\beta \in [0, \beta^*)$, with*

$$\beta^* > \frac{1}{1 + \left[1 - (1-\alpha)(1-\phi^{(1)}_\infty)\right]}.$$

Therefore, it is indexable for $\beta \in [0, \beta^*)$, and the MP index $\lambda^{MP}(x)$ calculated above is its Whittle's index $\lambda^*(x)$.

Notice that once the information state process X_t reaches the set $[\phi_\infty^{(1)}, \phi_\infty^{(0)}]$, it never leaves it. Then, for $x \in [0, \phi_\infty^{(0)}]$ the ETD problem (11) is PCL-indexable for all discount values $\beta \in [0, 1]$, as shown by 3, 7 and 11. Hence, the set of information states for which PCL-indexability in ensured only if $\beta < \beta^*$, i.e. $x \in (\phi_\infty^{(0)}, 1]$, applies only to a set of states which the system will, with certainty, leave and never return to, since the subset $[\phi_\infty^{(1)}, \phi_\infty^{(0)}]$ contains the absorbing set of states of the system operated under any z-threshold policy.

4.2 Index Computation

The Whittle's MP index has evaluation given by (24), (30) and (35). As already mentioned during the indexability analysis, the index $\lambda^*(x)$, which is further equal to the MP index $\lambda^{MP}(x)$, for the information states $0 \leq x < \phi_\infty^{(0)}$ in practice must be computed by by truncating the infinite series defining them to a finite number of terms.

4.2.1 Performance Bound Computation

Once the indexability of subproblem (11) is ensured by Theorem 14 and having proposed a tractable procedure to compute its optimal value given any λ (i.e. the optimal active set $B^*(z)$ contains those information states x such that $\lambda^*(x) - \lambda \geq 0$), we can solve the Lagrangian dual problem (8) stated as

$$V^D(\boldsymbol{x_0}) = \min_{\lambda \geq 0} \sum_{n=1}^{N} \left[\max_{\pi_n \in \Pi_n} f(x_{n,0}, \pi) - \lambda g(x_{n,0}, \pi) \right] + \lambda \frac{M}{(1-\beta)} \tag{36}$$

Hence, we may use $V^D(\boldsymbol{x_0})$ as a upper bound on the best attainable performance for problem (4). In the next section we will compute such a bound for the simulated scenarios considered and use it to evaluate the suboptimality gap of our proposed policy and other possible heuristics.

5 Computational Experiments

In this section we clarify and extend the ideas on the MARB elusive target hunt model presented in Section 4. First, we discuss, through a series of computational experiments, index tractability, the validity of PCL-indexability conditions and of theorem Theorem 14, and relative and absolute performance of the Whittle's index policy. Throughout the analysis, we will seek to draw insightful interpretations of the results in terms of the search problem of concern.

5.1 Index Evaluation

As an example of the use of our index computation method, we have simulated 10^3 runs of a scenario involving a target instance with the following parametric specification: $q^{(0)} = 0.1$, $p^{(0)} = 0.5$, $\rho^{(0)} = 1 - p^{(0)} - q^{(0)}$, $q^{(1)} = 0.5$, $p^{(1)} = 0.3$, $\rho^{(1)} = 1 - p^{(1)} - q^{(1)}$, $R = 1$, and $\alpha = 0.05$. The fixed points dividing the state space $\mathbb{X}^{0,1} \triangleq (0, 1]$ into the three analyzed threshold cases are $\phi_\infty^{(1)} = 0.3043$ and $\phi_\infty^{(0)} = 0.8333$. The discount factor β varied over the range $\beta \in \{0, 0.1, 0.2, \ldots, 0.9, 0.99\}$ and the critical discount factor is in this case $\beta^{(1)} = 0.7468$.

The index was computed using a MATLAB script for index evaluation based on the expressions (24), (30), (35). For each β, the index $\lambda^*(x)$ was evaluated on a grid of x information state values of width 10^{-2} and the infinite sums of cases I and II were approximately evaluated by truncating them to $T = 10^4$.

Figure 3 plots the results. As required by the PCL-indexability conditions, in each case the index $\lambda^*(x)$ is monotone nondecreasing in x. Note that the index is continuous in x and piecewise differentiable and it converges as $\beta \nearrow 1$ to a limiting index that can be used for the expected total criterion. For each x the time required to compute the index is negligible.

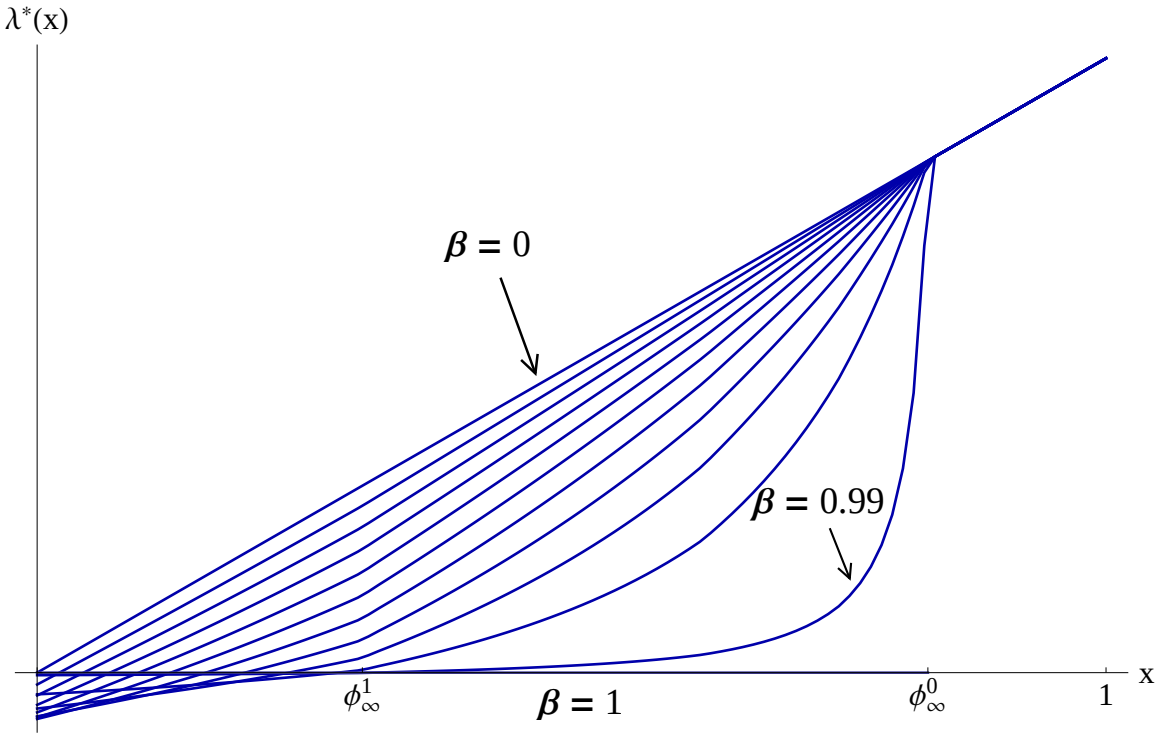


Figure 3: MP index for different discount factors β

From Figure 3 we derive the following relevant conclusions regarding the intuition of the optimal search policy for one elusive target in isolation.
For small enough x, (i.e., for $x \leq \phi_\infty^{(1)}$) the index $\lambda^*(x)$ may be negative for large values of β, reflecting the fact that it is unproductive to search a site when it is very unlikely that the target is visible, as both actions result in an increased probability that it is exposed (further, this increase is larger if we do not search for it).

For x within the *absorbing* set of states ($\phi_\infty^{(1)} \leq x \leq \phi_\infty^{(0)}$), as $\beta \nearrow 1$, the marginal profit of searching the target practically vanishes. This reflects the fact that as the system's lifetime grows, it becomes counterproductive to try to hunt a target which is unlikely to be exposed, as doing so will only drive the target into hiding, delaying the hunt.
By the same reasoning, the fact that the $\lambda^*(x)$ is decreasing in the discount factor β within the *absorbing* set $\phi_\infty^{(0)}$, suggests that as the moment in which the target is hunted is less important, then the best search strategy is to let the target be unsensed so that its probability of being exposed raises (up to its maximum value if $\beta = 1$), and only then attempt to hunt it. In simpler terms, if we have enough time to hunt the target, it is best to wait for the moment in which it becomes the most likely to be exposed, and only

then try to hunt it. For larger values of x (i.e., for $x > \phi_{\infty}^{(0)}$), it is optimal to behave myopically, since in those states the target is most likely to be exposed, yet those states are only *transient*.

5.2 PCL-indexability

As required by the PCL-indexability condition (ii), Figure 3 shows that in each case the index $\lambda^*(x)$ is monotone nondecreasing and continuous in x (in fact, it is strictly increasing in x). This section reports some computational evidence on the validity of condition (i), regarding the positivity of the marginal work measures, considering 10^3 runs of the target instance analyzed in in the previous section.

Figure 4 shows the results of computing the marginal work measure $w(x, z)$ fixing the z threshold value in $\{0.05, 0.5, 0.85\}$ and letting x vary in $\mathbb{X}^{0,1}$, analyzing a z value for each of the possible threshold cases described in subsection 5.1. The discount factor β varied over the range $\beta \in \{0, 0.1, 0.2, \ldots, 0.9, 0.99, 0.999\}$. For each β and z, the index $w(x, z)$ was evaluated on a grid of x values of width 10^{-2} and the infinite sums of cases I and II were approximately evaluated by truncating them to $T = 10^4$. Figure 4 illustrates how $w(x, z)$ differs for each threshold case considered. Further, notice that in these examples of case I ($z = 0.05$) and case II ($z = 0.5$), the marginal work measure positivity condition only holds for $\beta \leq 0.8$.

Notice that these simulation results are in accordance with the indexability analysis described in Section 4. Also, in light of the interpretations provided in subsection 5.1, note that since the target never returns to the largest values of x (i.e., $x > \phi_{\infty}^{(0)}$), then the total expected search effort to hunt it will be larger (in time) if we miss the opportunity to hunt it in those states than if we do not. Hence, the marginal work measure becomes negative for this range of x as the time horizon of the search increases.

5.3 Alternative Index Policies

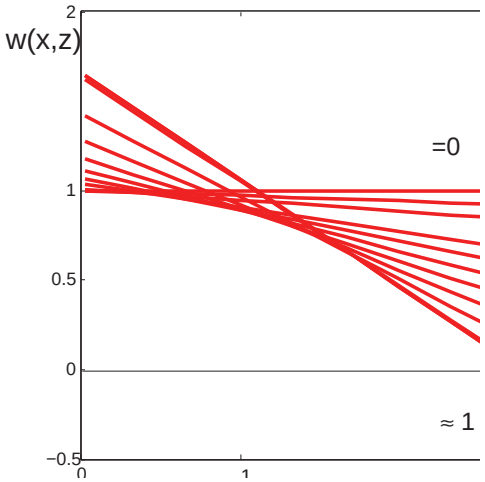
In this section we define some alternative heuristics for the MARB elusive target hunt problem (4) as stated in subsection 2.1. In the following section we will report simulation studies that compare the performance of Whittle's MP index policy against these simpler alternatives.

5.3.1 The Myopic Index Policy

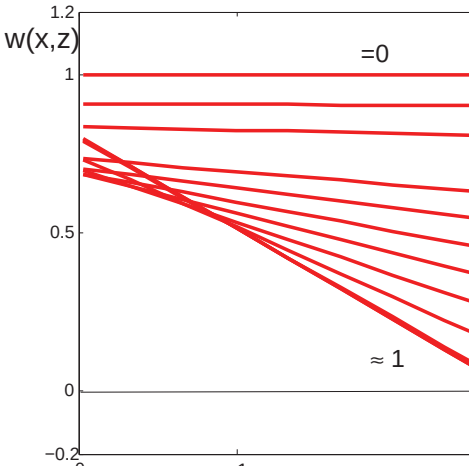
The *myopic* policy is based on index $\lambda^{Myopic}(x) = R(x, 1)$ for all $x \in \mathbb{X}^{0,1}$. Notice from Figure 3 that this index also corresponds with Whittle's MP index $\lambda^*(x)$ for the case $\beta = 0$ and also for all discount factors β when x is in the range $(\phi_{\infty}^{(0)}, 1]$.

5.3.2 The Belief State Index Policy

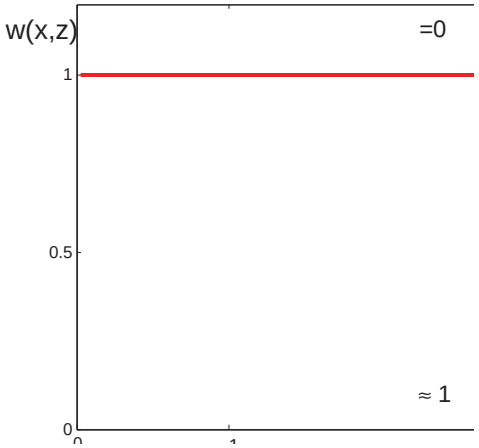
The *belief state* policy is based on index $\lambda^B(x) = x$, for all $x \in \mathbb{X}^{0,1}$. At this point, it is worth pointing out that since $\lambda^{Myopic}(x)$, $\lambda^B(x)$ and $\lambda^*(x)$ are monotone increasing functions of the information state x, in instances of identical targets the three policies result in equivalent sensing decisions, as the higher the information state the greater the priority a target receives under all search rules.



(a) $w(x, z = 0.05)$ *(Low Threshold)*



(b) $w(x, z = 0.5)$ *(Intermediate Threshold)*



(c) $w(x, z = 0.85)$ *(High Threshold)*

Figure 4: Marginal work measure for the z-threshold cases

5.3.3 The Random Search Policy

The *random* selection policy is based on picking a site to search (among the ones that contain an uhunted target) at random, with each site having the same probability of being selected.

5.4 Bechmarking the Whittle Index Policy

We have performed some small-scale preliminary simulation studies, based on MATLAB implementations we have developed to compare the performance of the proposed Whittle's MP index policy against the the *myopic* policy, the *belief state* policy, and the *random* selection policy.
Further, we have computed an upper bound on the optimal value (4) based on the ideas discussed in subsection 3.1.

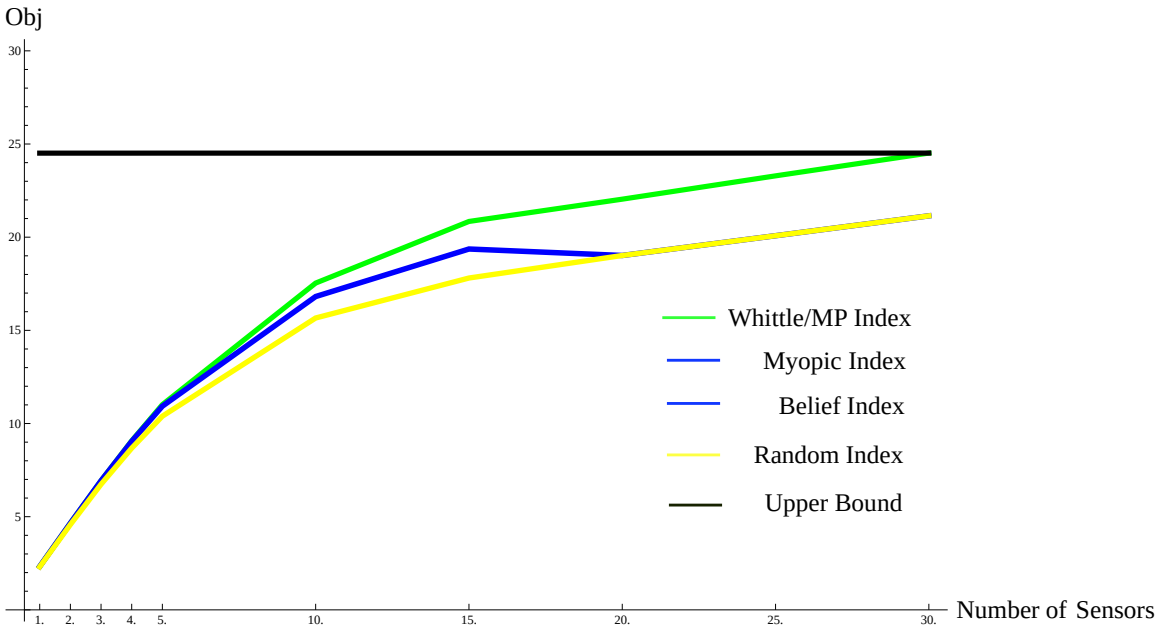
5.4.1 Cautious and Reckless targets

In this experiment we assess the relative performance of the Whittle's MP index policy against the other heuristics distinguishing target instances between *reckless* and *cautious*. We call reckless those targets which "*after not being searched, are highly likely to expose themselves*", i.e. with $p^{(0)} \approx 1$, while cautious targets display the opposite behavior, i.e. with $p^{(0)} \approx 0$ (while having $p^{(0)} > p^{(1)}$).
Each base instance has a single sensor $M = 1$ for searching within $N = 30$ sites, in one instance all targets are reckless with $p_n^{(0)} = 0.95$, while in the other instance all targets are cautious with $p_n^{(0)} = 0.35$. In both instances, $p_n^{(1)} = 10^{-3}$, $q_n^{(1)} = 0.97$, $q_n^{(0)} = 0.003$, $\alpha_n = 0.30$ and $R_n = 1$ for all n. Thus, for both targets $\phi_\infty^{(1)} = 0.0010$ while for reckless $\phi_\infty^{(0)} = 0.9694$ and for cautious $\phi_\infty^{(0)} = 0.9211$.

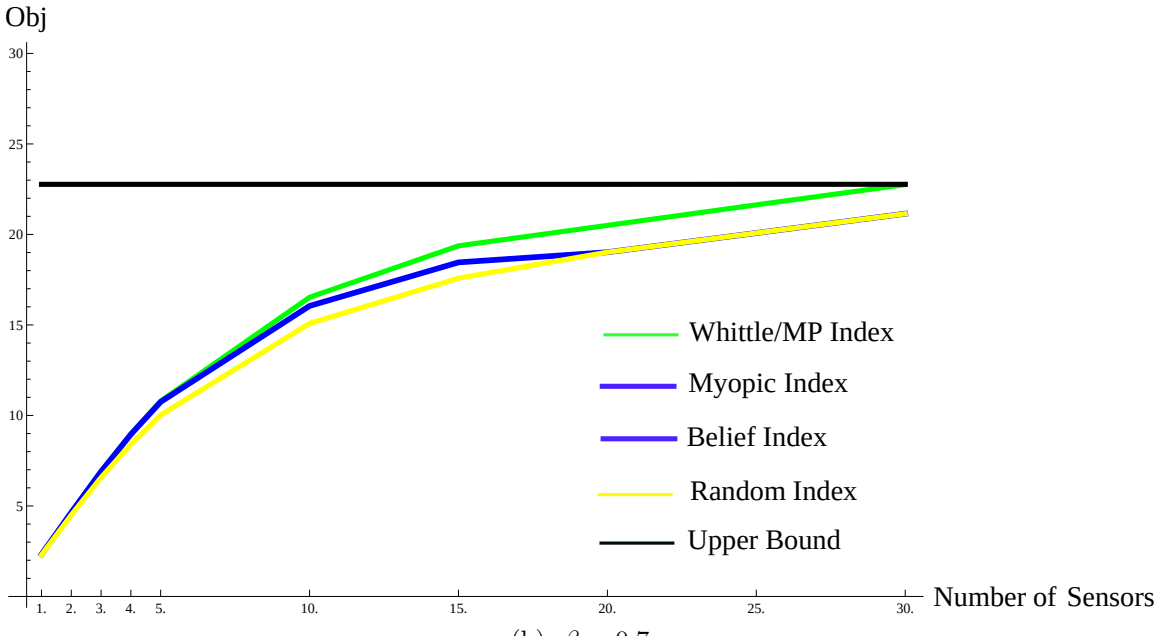
We take the initial state $x_n = 1$, which corresponds to exact knowledge of N exposed targets at the start of the search. Sensing costs were taken to be zero and we consider two possible discount factors $\beta \in \{0.7, 0.99\}$, where β^* is equal to 0.7688 both for the reckless and cautious instance. Both base instances were modified, letting the number of sensors increase from $M = 1$ up to $M = N = 30$. For each instance, 10^3 independent runs were performed on a horizon of $T = 10^4$ time slots.

Figure 6 shows the ETD net rewards under each policy as the number of sensors in the network grows. The upper bound from the relaxation for all the instances with reckless targets was of 24.510 and 29.735 for discount factors 0.7 and 0.99, respectively, whereas for cautious targets those values were 22.767 and 29.374. Note that the bound on the best result of the search is always less when targets are cautious, since they are harder to hunt.

As depicted by Figure 6, the Whittle's MP policy outperforms other heuristic policies for any number of sensors with the performance improvement increasing as $M \nearrow N$. In fact, as the number of sensors grows all policies perform worse, except for the Whittle's MP policy for which the opposite occurs. The explanation of this results is that all other heuristics *overuse* the network resources as they become available, searching more and more sites possibly containing a target, thus making targets more elusive and hence, more difficult to hunt. This is a salient result, since it points out a severe drawback that myopic or simpler heuristic have for allocating resources in cases in which idling is expected to have a greater impact on the system's expected returns.

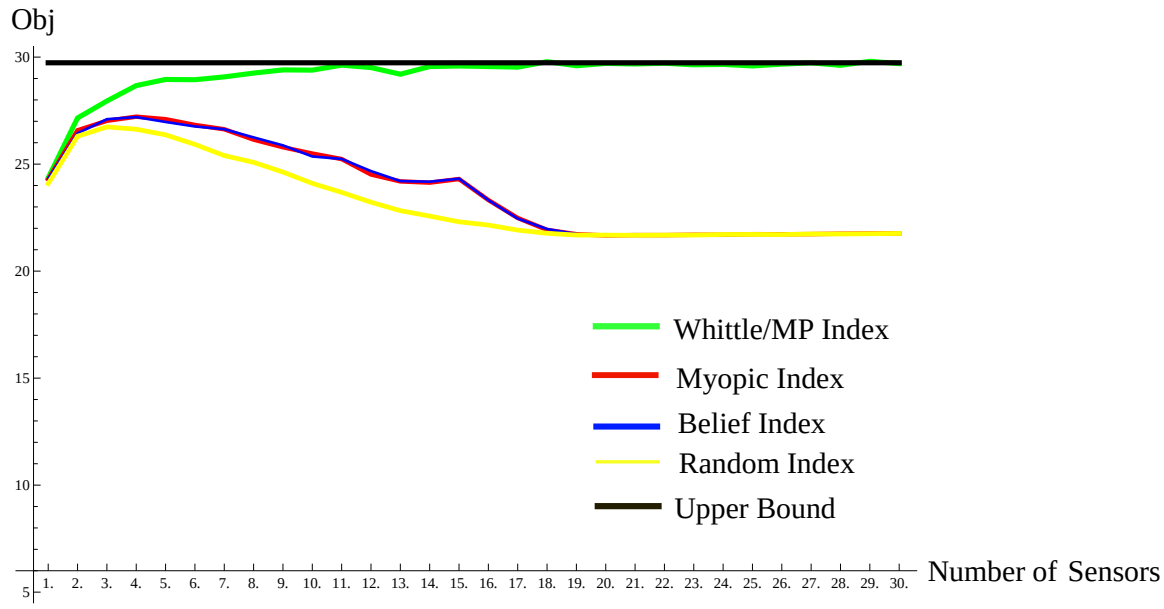


(a) $\beta = 0.7$

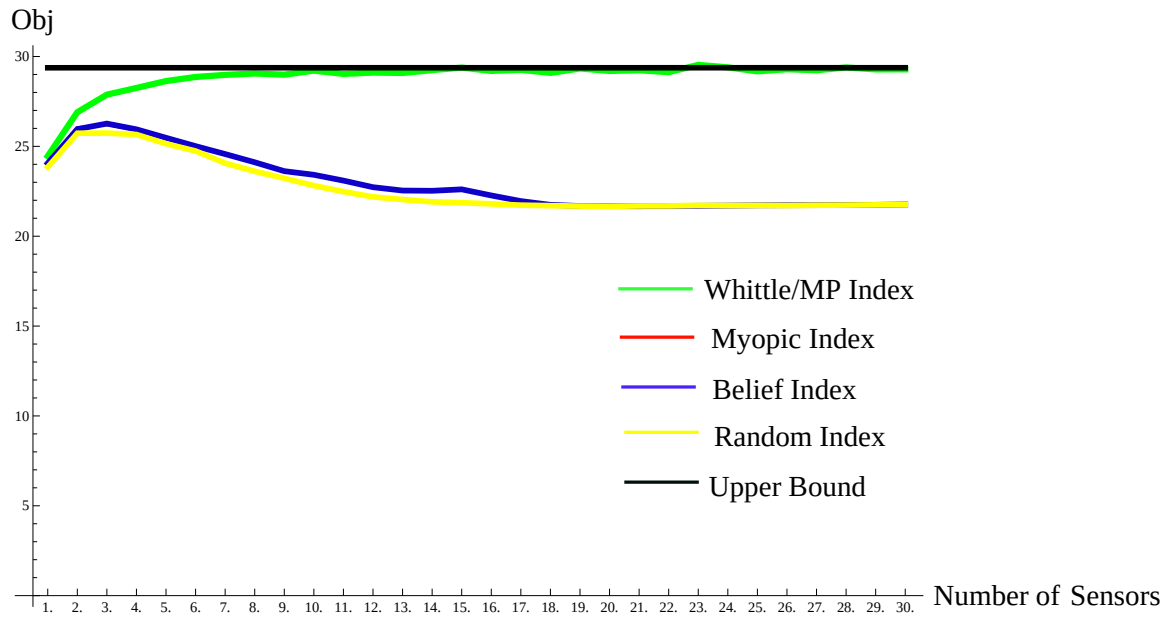


(b) $\beta = 0.7$

Figure 5: Experiment 1 - (5a) & (5b), *Reckless and Cautious Targets* instances



(a) $\beta = 0.99$



(b) $\beta = 0.99$

Figure 6: Experiment 2 - (6a) & (6b), *Reckless and Cautious Targets* instances

Table 1: Average System's Operating Time

M / Reckless	$\bar{T}^{MP}$	$\bar{T}^{My}$	$\bar{T}^{B}$	$\bar{T}^{R}$
1	6.175	9.768	7.083	31.039
2	5.778	18.994	12.021	42.475
3	2.405	49.074	45.718	95.318
4	3.344	36.678	33.071	70.652
5	3.034	90.074	76.949	102.643
15	1.928	122.554	155.053	371.901
30	1.924	373.586	366.239	458.258
M / Cautious	$\bar{T}^{MP}$	$\bar{T}^{My}$	$\bar{T}^{B}$	$\bar{T}^{R}$
1	10.770	43.833	37.630	44.864
2	6.384	29.845	33.414	80.638
3	4.841	66.301	55.581	87.306
4	3.828	74.822	76.426	138.993
5	3.970	135.726	86.134	182.447
15	3.277	281.945	264.044	410.706
30	3.073	448.593	465.127	423.586

Another interesting result is that the Whittle's MP policy suboptimality gap tends to 0 for a relatively small number of sensors when $\beta \approx 1$, while the largest sensor network size is required for the Whittle's MP policy to be nearly optimal for smaller β (i.e. when hunting targets is urgent). Such a result is related to the fact that all policies successfully find the N targets, yet they differ significantly in the time they take to do so. Thus, if the hunt mission is urgent a large sensor network (operated under the Whittle's MP index policy) will result nearly optimal whereas if the mission is just to find the objects but not urgently a relatively small sensor network is required.

Table 1 shows the average time that the system takes to hunt all targets operated under each policy. Such results illustrate the fact that a large sensor network which is constantly searching will spend a larger period of time to hunt targets. However, all policies succeed at finding the N targets at some period. The Whittle's MP index policy takes significantly less time to hunt targets than the alternative polices for both Reckless and Cautious targets, yet hunting the Cautious targets naturally takes longer for all policies. These results also show the overuse under other heuristics since their average operating time substantially increases as the number of sensors grows. Results in Table 1 are of particular relevance in terms of the specific motivating application proposed in subsection 1.1 and investigated in [10]. The main goal in that case was to have a scheduling policy which minimizes the average time until all missile launchers are detected and destroyed. As Table 1 shows, the Whittle's MP policy is the heuristic that manages to find all targets in the least time.

To sum up, the proposed policy is always as good as the other heuristics, yet in many instances it does yield important performance improvements. These performance improvements of the Whittle's MP policy are significant from a statistical point of view,and from a practical point of view (performance gains can be up to 36,48%). Further, the performance improvements become more important as the size of the sensor network increases. In fact, the Whittle's MP policy is even nearly optimal in both scenarios when

$M \nearrow N$. Also, the Myopic and the Belief policy are not significantly different in these scenarios, nor do they improve significantly on the random policy. Further, all the policies produce the same results when $M = 1$, basically because they are all equally forced to not search the remaining unhunted targets. Performance differences are observed when the system has the possibility of searching a site and a index policy prescribes not to do so.

5.4.2 Sensing Costs

In this experiment we assess the relative performance of the Whittle's MP index policy against the other heuristics as the sensing cost c increases. We consider two base instances of $N = 10$ sites with $M = 1$ and $M = 5$ sensors. In both instances targets parameters are: $p_n^{(1)} = 10^{-3}$, $q_n^{(1)} = 0.97$, $p_n^{(0)} = 0.05$, $q_n^{(0)} = 0.003$, $\alpha_n = 0.30$, $x_n = 1$, $\beta = 0.99$ and $R_n = 1$ for all n. Both base instances were modified, letting sensing costs for all sites vary as $c \in \{0, 0.3, 0.5, 0.75\}$. For each instance, 10^3 independent runs were performed on a horizon of $T = 10^4$ time slots.

Figure 7 plots the ETD net rewards under each policy and the upper bound as c grows. Results show that the Whittle's MP index policy outperforms the other policies in all instances. The random policy performs significantly worse in this case, basically because it prescribes to search sites, provided there are enough sensors, regardless of the sensing cost, while the other two heuristics have been defined in such a way that they prescribe to search only if their index value exceeds c.

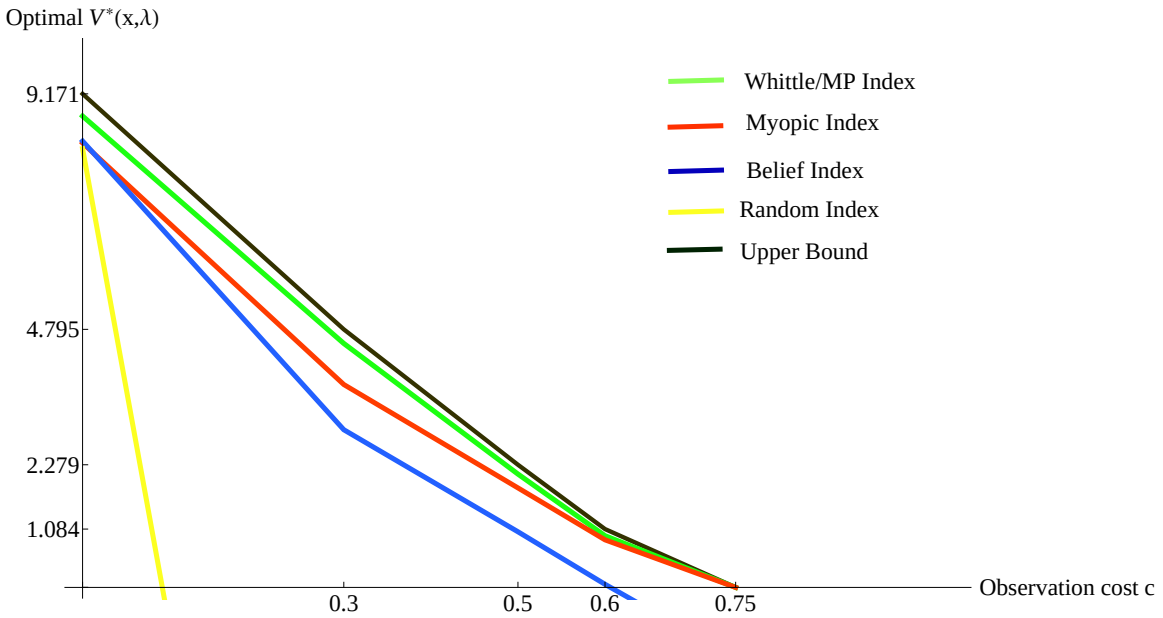
Naturally, as searching becomes expensive, both the resulting performance under all policies and its upper bound decrease. In the Figure 7 we observe that the system yields 0 rewards for $c > 0.75$. Actually, the optimal value function vanishes when $c = R(1, 1)$, which in this case is $c = 0.7$.

Notice that the Whittle's MP policy is nearly optimal for all values of the sensing cost when $M = 5$ while the suboptimality gap of the other heuristics is larger for $M = 5$ than for $M = 1$, a result consistent with the overuse of the simpler heuristics pointed out before.

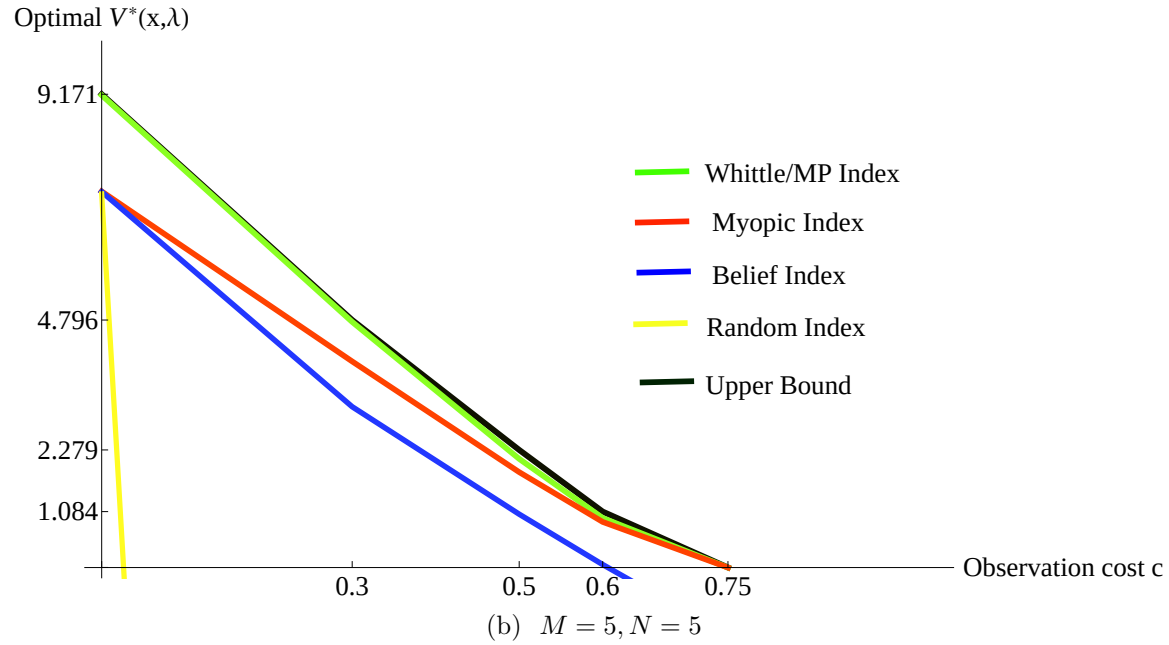
5.5 Sensor Network Size

Perhaps one of the most notorious results obtained, with special consequence for the design of sensing systems for hunting such elusive targets, is that if the horizon is long enough (i.e. as $\beta \nearrow 1$), operating a system's under the Whittle's MP policy requires a few sensors to optimally hunt a larger set of targets. In the instances plotted in 6a we observe that a sensor network of $M \approx 12$ or more sensors is enough to achieve the best possible expected reward under the Whittle's MP policy provided that targets are reckless. If targets are cautious, as in 6b, a sensor network of $M \approx 8$ is enough to achieve optimality, as the system spends less time actively searching targets.

The results also suggest that if the optimal scheduling policy is not tractable, and we are forced to operate the system under a simple heuristics, if we define heuristics which do not advise the system to idle, it will take longer to find all targets. Thus, for this kind of problems it makes more sense to define heuristics of *round-robin* type, specifying how to alternate between searching and not searching a target, as the Whittle's MP index does, than myopically operating the system.



(a) $M = 1, N = 5$



(b) $M = 5, N = 5$

Figure 7: Experiment 2: Sensing Cost Effect with: $M/N = 1/10$ (7a) and $M/N = 1/2$ (7b)

6 Concluding Remarks

This paper has introduced a novel dynamic index policy for a relevant sensor network scheduling problem where the goal is to hunt a fixed number of smart targets, in which the theory of restless bandit indexation is applied to a POMDP setting. The resulting policy has been shown in simulation experiments to outperform simpler heuristics.

References

[1] O. Hernández-Lerma and J. B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*. Springer, New York, NY, 1999.

[2] C. Kreucher, D. Blatt, A. Hero, and K. Kastella. Adaptive multi-modality sensor scheduling for detection and tracking of smart targets. *Digital Signal Processing*, 16(5):546–567, 2006.

[3] B. Liu, C. JI, Y. Zhang, and C. Hao. Blending sensor scheduling strategy with particle filter to track a smart target. *Wireless Sensor Network*, 1:300–305, 2009.

[4] W. Moran, S. Suvorova, and S. Howard. Application of sensor scheduling concepts to radar. *Foundations and Applications of Sensor Management*, pages 221–256, 2008.

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

[6] J. Niño-Mora. An index policy for dynamic fading-channel allocation to heterogeneous mobile users with partial observations. In *Next Generation Internet Networks, 2008. NGI 2008*, pages 231–238. IEEE, 2008.

[7] J. Niño-Mora. A restless bandit marginal productivity index for opportunistic spectrum access with sensing errors. *Network Control and Optimization. Lecture Notes in Computer Science.*, Volume 5894,:60–74, 2009.

[8] J. Niño-Mora and Sofia Villar, S. Sensor scheduling for hunting elusive hiding targets via whittle's restless bandit index policy. In *NetGCOOP 2011 : International conference on NETwork Games, COntrol and OPtimization*. IEEE, 2011.

[9] J. Niño-Mora and S.S. Villar. Multitarget tracking via restless bandit marginal productivity indices and Kalman filter in discrete time. In *Proceedings of the 48th IEEE Conference on Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009*, pages 2905–2910. IEEE, 2009.

[10] J.E. Rucker. Using agent-based modeling to search for elusive hiding targets. Technical report, DTIC Document, 2006.

[11] CO Savage and BF La Scala. Sensor management for tracking smart targets. *Digital Signal Processing*, 19(6):968–977, 2009.

[12] R. Washburn. Application of multi-armed bandits to sensor management. *Foundations and Applications of Sensor Management*, pages 153–175, 2008.

[13] P. Whittle. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988.