Sensor Scheduling for Hunting Elusive Hiding Targets via Whittle's Restless Bandit Index Policy

José Niño-Mora and Sofía S. Villar
Department of Statistics
Carlos III University of Madrid
28911 Leganés (Madrid), Spain
Email: jnimora@alum.mit.edu, svillar@est-econ.uc3m.es

***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 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.**

I. Introduction

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 set of sensors to track a larger set of moving targets, to optimize a system-wide performance objective. See, e.g., [1].

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 us to formulate the *optimal sensor scheduling problem* as a *partially observable Markov decision process* (POMDP) with special structure, which often fits into the framework of the continuous-state *multi-armed bandit problem*, either in its classic version or, more often, in its restless variant. See [2].

Although the *multi-armed restless bandit problem* (MARBP) is, generally, computationally intractable, formulating a sensor scheduling problem in such a framework allows us to use the general heuristic *index policy* proposed by Whittle in [3] for the former problem. If M sensors are available to track $N > M$ targets, such a policy attaches an index $\lambda_n^*(x_n)$ to each target $n = 1, \ldots, N$ as a function of its current information state x_n, and then senses at each time up to M targets with higher index values, among those targets, if any, whose current index value exceeds or equals the search cost c_n i.e., such that $\lambda_n^*(x_n) \geqslant c_n$.

Such an approach has been used to address sensor scheduling problems in, e.g., [4]–[9]. Besides a policy, it further provides an upper bound on the optimal problem value, which can be used to assess the optimality loss of heuristic policies. A growing body of numerical work suggests that the performance of the *Whittle index policy* is often nearly optimal.

The targets are typically assumed to follow dynamics that are unaffected by sensing decisions. Yet, in certain applications, targets are *smart*, in that they react to sensing actions by changing their dynamics to make them *elusive*. However, few papers consider sensor scheduling with smart targets. [10] uses reinforcement learning to obtain a non-myopic policy for detection and tracking of smart targets, while [11] uses particle filter methods, and [12] uses game theory. [13] uses agent-based modeling to address a model similar to ours.

This paper extends such a line of research by investigating 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.

As a specific motivating application, consider the problem investigated in [13], where the targets are mobile platforms (transporter-erector-launchers) for launching Scud missiles, and the sites are areas where it is known that such platforms are located. Sensors can be mounted on unmanned aerial vehicles (UAV). The optimal sensor scheduling problem is to devise a scheduling policy for minimizing the average time until all missile launchers are detected and destroyed.

The remainder of the paper is organized as follows. Section II describes and formulates the model. Section III reviews the restless bandit indexation approach as it applies to the design of index policies for the present model. Section IV outlines how to deploy such a methodology to compute the index.

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.

Section V reports on several simulation experiments where the proposed index policy is compared with alternative heuristic policies. Finally, section VI ends the paper with concluding remarks. Detailed analysis and proofs will be included in a full version of this paper, which is currently under preparation.

II. Model Description and Formulation

We consider a model where M sensors are available to hunt $N > M$ elusive targets, where each target n is located at a corresponding site $n = 1, \ldots, N$. The target at site n changes its *visibility state* $s_{n,t}$ at discrete time slots $t = 0, 1, \ldots$ 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 searching the site.

The visibility state $s_{n,t}$ of target n 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 $a_{n,t}$, where $a_{n,t} = 1$ if site n is searched at time t, and $a_{n,t} = 0$ otherwise. When $a_{n,t} = a$ 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. $q_n^{(a)}$). Such probabilities are such that, after a site is searched and the 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 $p_n^{(1)} < p_n^{(0)}$. We assume that the visibility state processes have positive autocorrelation, i.e., $\rho_n^{(a)} \triangleq 1 - p_n^{(a)} - q_n^{(a)} > 0$.

The target at site n is hunted when it is searched if it is exposed, which yields a reward r_n. Information on the visibility state of target n is gained by sensing it, which provides a *sensing 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 the site is searched and the target is exposed with a positive *misdetection* probability $\alpha_n = \mathsf{P}\{o_{n,t} = 0 \,|\, s_{n,t} = 1\}$. Hence, the visibility state $s_{n,t}$ is not observable, but it is tracked by the *information state* $X_{n,t} \in \mathsf{X} \triangleq [0, 1]$, giving the posterior probability that the target will be exposed in slot t, conditioned on the history $\{X_{n,s}, a_{n,s} \colon 0 \leq s < t\} \cup \{X_{n,t}\}$.

Since we assume that a site n whose target that has been hunted ($x_n = 0$) is removed from further search, we partition a target state space X into the set $\bar{\mathsf{X}} \triangleq (0, 1]$ of *controllable states*, where both actions $a_n \in \{0, 1\}$ 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. If the site is searched in slot t ($a_{n,t} = 1$), with its information state being $X_{n,t}$, and the target is detected ($o_{n,t} = 1$), which happens with probability $(1 - \alpha_n) X_{n,t}$, then the target is hunted, and the site is removed. We model such a situation by letting the target information state drop to $X_{n,t+1} = 0$.

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

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

Thus, if 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 searched ($a_{n,t} = 0$) in slot t, with its information state being $X_{n,t}$, then, as long as the target has not yet been hunted (i.e., if $X_{n,t} > 0$), its next information state is determined by $X_{n,t+1} = p_n^{(0)} (1 - X_{n,t}) + \big(1 - q_n^{(0)}\big) X_{n,t} = p_n^{(0)} + \rho_n^{(0)} X_{n,t}$. Yet, if the target has already been hunted ($X_{n,t} = 0$), its information state remains at 0, i.e., $X_{n,t+1} = 0$.

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

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

As for the economic consequences of the actions, taking action a_n on site n when it occupies the information state x_n yields the expected one-slot net reward $R_n(x_n, a_n) \triangleq \big(r_n (1 - \alpha_n)\, x_n - c_n\big) a_n$, where $c_n \geqslant 0$ is the cost of searching site n.

Consider the problem of finding a *discounted-reward optimal* policy,

$$\max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}(M)} \mathsf{E}_{\mathbf{x}_0}^{\boldsymbol{\pi}} \left[\sum_{t=0}^{\infty} \sum_{n=1}^{N} \beta^t R_n\big(X_{n,t}, a_{n,t}\big) \right], \tag{2}$$

where $0 < \beta \leqslant 1$ is the discount factor, $\mathbf{x}_0 = (x_{n,0})_{n=1}^N$, and $\mathsf{E}_{\mathbf{x}_0}^{\boldsymbol{\pi}}[\cdot]$ denotes expectation under policy $\boldsymbol{\pi}$ conditioned on the initial joint state being equal to $\mathbf{x}_0$. Note that the case $\beta = 1$ (total expected reward criterion) is well defined here.

Problem (2) is a POMDP of MARBP type. Since such problems are hard to solve, our main goal is to develop tractable policies that are close to optimal.

Given their intuitive appeal and tractability, we will focus on heuristics of *priority-index* type. Such policies attach an index $\lambda_n(x_n)$ to each site n as a function of its state x_n. At time t, the index policy selects at most M sites to sense, using $\lambda_n(x_{n,t})$ as a priority index for sensing site n (where a larger index value means a higher priority), among those sites, if any, at which the current index value exceeds or equals the search cost ($\lambda_n\big(x_{n,t}\big) \geqslant c_n$), breaking ties arbitrarily.

III. MARBP formulation and the Whittle index

We will deploy the approach developed and applied in other real-state MARBP models in [6]–[8], as reviewed next.

A. Relaxed Problem, Lagrangian Relaxation and Performance Bound

Along the lines introduced in [3] for the equality-constrained case, we first construct a relaxation of (2), replacing the sample-path activity constraint (1) with the weaker constraint that the expected total discounted (ETD) number of sensed sites does not exceed $M/(1-\beta)$, i.e.,

$$\mathsf{E}^{\boldsymbol{\pi}}_{\mathbf{x_0}}\left[\sum_{t=0}^{\infty}\sum_{n=1}^{N}\beta^t a_{n,t}\right] \leqslant \frac{M}{1-\beta}. \tag{3}$$

Denoting by $\widehat{\boldsymbol{\Pi}}$ the class of nonanticipative scheduling policies that are allowed to search any number of sites at any time, the *relaxed primal problem* is

$$\max_{(3),\boldsymbol{\pi}\in\widehat{\boldsymbol{\Pi}}} \mathsf{E}^{\boldsymbol{\pi}}_{\mathbf{x_0}}\left[\sum_{t=0}^{\infty}\sum_{n=1}^{N}\beta^t R_n\big(X_{n,t},a_n\big)\right]. \tag{4}$$

Note that the optimal value $V^{\mathrm{R}}(\mathbf{x_0})$ of (4) gives an *upper bound* on the optimal value $V^*(\mathbf{x_0})$ of (2).

To address the constrained MDP (4) we use a Lagrangian analysis, dualizing the constraint (3) using a multiplier $\lambda \geqslant 0$. The resulting problem

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

is a *Lagrangian relaxation* of (4), 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 a value $\lambda^*(\mathbf{x_0})$ of λ giving the best such upper bound, which we denote by $V^{\mathrm{D}}(\mathbf{x_0})$:

$$V^{\mathrm{D}}(\mathbf{x_0}) = \min_{\lambda\geqslant 0} V^{\mathrm{L}}(\mathbf{x_0};\lambda) \tag{6}$$

Note that (6) is a scalar convex optimization problem, since $\lambda \mapsto V^{\mathrm{L}}(\mathbf{x_0};\lambda)$ is convex. Under suitable regularity conditions, *strong duality* holds, i.e., $V^{\mathrm{D}}(\mathbf{x_0}) = V^{\mathrm{R}}(\mathbf{x_0})$.

B. Indexability and the Whittle Index Policy

Since target state transitions are independent, problem (5) decomposes into the N single-sensor single-site subproblems

$$\max_{\pi_n\in\Pi_n} \mathsf{E}^{\pi_n}_{x_{n,0}}\left[\sum_{t=0}^{\infty}\beta^t\{R_n\big(X_{n,t},a_{n,t}\big)-\lambda a_{n,t}\}\right], \tag{7}$$

where Π_n denotes the class of admissible policies for operating a single sensor on site n. Note that the Lagrange multiplier λ plays the role of an additional search cost.

Denoting by $V_n^{\mathrm{L}}(x_{n,0};\lambda)$ the optimal value of subproblem (7), the optimal value $V^{\mathrm{L}}(\mathbf{x}_0;\lambda)$ of (5) is decomposed as

$$V^{\mathrm{L}}(\mathbf{x}_0;\lambda) = \frac{M\lambda}{(1-\beta)} + \sum_{n=1}^{N} V_n^{\mathrm{L}}(x_{n,0};\lambda). \tag{8}$$

We will say that the single-site subproblem (7) is *indexable* if there exists an *index* $\lambda_n^*(x_n)$ attached to the controllable information states $x_n \in \bar{\mathsf{X}} \triangleq (0,1]$, which is such that, for any value of $\lambda\in\mathbb{R}$, it is optimal to search the site when it occupies state x_n, regardless of the initial state, iff $\lambda_n^*(x_n) \geqslant \lambda$.We will refer to $\lambda_n^*(x_n)$ as the *Whittle index*, or the *marginal productivity* (MP) index for site n.

If each subproblem (7) is indexable, then we can use the Whittle indices $\lambda_n^*(x_n)$ to obtain a corresponding priority-index policy, as described above.

C. Sufficient Indexability Conditions and Index Evaluation

The indexability property for restless bandits, introduced in [3], cannot be taken for granted, needing to be established for the model at hand. The introduction in [14], [15] of sufficient indexability conditions for discrete-state restless bandits, along with an index algorithm, based on satisfaction of *partial conservation laws* (PCLs), provided a methodology for such a purpose. See the review [16]. Such conditions were extended to real-state restless bandits in [6], as reviewed next.

We focus here on a generic site and target, dropping the subscript n from the notation. We will evaluate sensing policies $\pi\in\Pi$ along two dimensions: the *work measure* $g(x,\pi)$, giving the ETD number of times the site is searched under policy π starting from $X_0 = x$; and the *reward measure* $f(x,\pi)$, giving the ETD reward earned:

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

The single-site subproblem (7) is thus formulated as

$$\max_{\pi\in\Pi} f(x,\pi) - \lambda g(x,\pi). \tag{9}$$

By standard results, there exists an optimal policy for the discounted real-state MDP (9) that is *stationary deterministic*. Such policies are conveniently represented by their *active (state) sets*, i.e., the set of information states where the active action (search the site) is prescribed. For an active set $B\subseteq\bar{\mathsf{X}}$, we will thus refer to the B*-active policy*.

We will further focus on the family of *threshold policies*. Given a threshold $z\in\mathbb{R}$, the z*-threshold policy* prescribes to search the site in information state x iff $x > z$, so its active set is $B(z) \triangleq \{x\in\bar{\mathsf{X}}\colon x > z\}$. Note that $B(z) = (z,1]$ for $0\leqslant z < 1$, $B(z) = \bar{\mathsf{X}} = (0,1]$ for $z < 0$, and $B(z) = \emptyset$ for $z \geqslant 1$. We denote by $g(x,z)$ and $f(x,z)$ the work and reward measures under the z-threshold policy.

In the sequel, we will write $b(x) \triangleq 1-(1-\alpha)x$, and

$$\phi^{(0)}(x) \triangleq p^{(0)} + \rho^{(0)}x,\ \phi^{(1)}(x) \triangleq p^{(1)} + \frac{\rho^{(1)}\alpha x}{b(x)}. \tag{10}$$

Given threshold z, the performance measures $g(x,z)$ and $f(x,z)$ are characterized as the unique solutions to the following functional equations:

$$g(x,z) = \begin{cases} 1+\beta b(x) g\big(\phi^{(1)}(x),z\big), & x\in(z,1] \\ \beta g\big(\phi^{(0)}(x),z\big), & x\in(0,z] \\ 0 & x=0, \end{cases} \tag{11}$$

$$f(x,z) = \begin{cases} R(x,1)+\beta b(x) f\big(\phi^{(1)}(x),z\big), & x\in(z,1] \\ \beta f\big(\phi^{(0)}(x),z\big), & x\in(0,z] \\ 0 & x=0. \end{cases} \tag{12}$$

Given threshold z and action a, denote by $\langle a,z\rangle$ the policy that takes action a in the initial time slot and then follows the z-threshold policy thereafter. Define the *marginal work* *measure* $w(x,z)$ and the *marginal reward measure* $r(x,z)$ for $x \in \bar{\mathsf{X}}$ by

$$\begin{aligned} w(x,z) &\triangleq g(x,\langle 1,z\rangle) - g(x,\langle 0,z\rangle) && (13)\\ &= 1 + \beta b(x) g(\phi^{(1)}(x),z) - \beta\, g(\phi^{(0)}(x),z), \\ r(x,z) &\triangleq f(x,\langle 1,z\rangle) - f(x,\langle 0,z\rangle) \\ &= R(x,1) + \beta b(x) f(\phi^{(1)}(x),z) \\ &\quad -\beta\, f(\phi^{(0)}(x),z). && (14)\end{aligned}$$

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

$$\lambda(x,z) \triangleq \frac{r(x,z)}{w(x,z)}. \qquad (15)$$

We say that subproblem (9) is *PCL-indexable* (with respect to threshold policies) if:

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

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

$$\lambda^*(x) \triangleq \lambda(x,x), \quad x \in \bar{\mathsf{X}}. \qquad (16)$$

is monotone nondecreasing in x

Theorem 1. *If subproblem* (9) *is PCL-indexable, then it is indexable and the* $\lambda^*(x)$ *in* (16) *is its Whittle index.*

IV. Index Computation

To apply the PCL-indexability conditions, we must first calculate the evaluation measures $g(x,z)$ and $f(x,z)$.

In some cases (see [6]), such measures can be evaluated in closed form, which allows for direct verification of the PCL-indexability conditions, and yields a closed-form index formula. Such is not the case, however, with the present model. This 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 $\lambda^*(x)$.

A. Total and Marginal Evaluation Measures

As a preliminary step for solving (11) and (12), we define $\phi_t^{(a)}(x)$ for $a \in \{0,1\}$ as the t-th iterate of the recursion $\phi_0^{(a)}(x) \triangleq x$ and $\phi_t^{(a)}(x) \triangleq \phi^{(a)}(\phi_{t-1}^{(a)}(x))$. Note that, for any $x \in \bar{\mathsf{X}}$, $\lim_{t\to\infty} \phi_t^{(a)}(x) = \phi_\infty^{(a)}$, where

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

with $\gamma \triangleq 1 - \rho^{(1)} + (1-\alpha)(p^{(1)} + \rho^{(1)})$. The above assumptions on the transition probabilities for the target visibility state ensure that $\phi_\infty^{(1)} < \phi_\infty^{(0)}$. Hence, to solve the evaluation equations we must distinguish three cases, as discussed below. In the sequel we assume, without loss of generality, that the search cost is $c = 0$.

Case I: $z \in \big[0, \phi_\infty^{(1)}\big]$*:* In this case, once the target state X_t reaches the active set $B(z) = (z,1]$, it remains in $B(z)$ as long as the target is not hunted. Thus, given the initial state $x \leqslant z$, let $t_0^*(x,z) \triangleq \min\{t \geqslant 1 \colon X_t > z\}$ be the first hitting time to $B(z)$, which can be computed as $t_0^*(x,z) \triangleq \min\{t \geqslant 1 \colon \phi_t^{(0)}(x) > z\}$, and let $y \triangleq \phi_{t_0^*(x,z)}^{(0)}(x)$. Also, denote by $\theta(x,z,t)$ the *survival probability* that the target has not been hunted before time slot t under the z-threshold policy, starting from state x. Note that, for $x > z$,

$$\theta(x,z,t) = \prod_{s=0}^{t-1} \big[1 - (1-\alpha)\, \phi_s^{(1)}(x)\big], \quad t \geqslant 1,$$

and $\theta(x,z,0) = 1$. We have the work measure evaluation

$$g(x,z) = \begin{cases} \displaystyle\sum_{t=0}^{\infty} \beta^t \theta(x,z,t), & x \in (z,1] \\ \beta^{t_0^*(x,z)} g(y,z), & x \in (0,z]. \end{cases} \qquad (17)$$

Similarly, we obtain the reward measure evaluation

$$f(x,z) = \begin{cases} \displaystyle\sum_{t=0}^{\infty} \beta^t \theta(x,z,t) R\big(\phi_t^{(1)}(x,z),1\big), & x \in (z,1] \\ \beta^{t_0^*(x,z)} f(y,z), & x \in (0,z]. \end{cases} \qquad (18)$$

The above infinite series are convergent, yet they do not admit closed form formulae. Hence, they must be truncated in practice to approximate $w(x,z)$ and $r(x,z)$ via (13)–(14), and for approximating the index $\lambda^*(x)$ in (16) for $0 < x \leqslant \phi_\infty^{(1)}$.

Case II: $z \in \big(\phi_\infty^{(1)}, \phi_\infty^{(0)}\big)$*:* In this case, 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 $\phi(x,z) \triangleq 1_{\{x>z\}}\phi^{(1)}(x) + 1_{\{x\leqslant z\}}\phi^{(0)}(x)$, and let $\phi_0(x,z) = x$, $\phi_t(x,z) = \phi(\phi_{t-1}(x,z),z)$ for $t \geqslant 1$. 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 $\theta(x,z,t)$ has the evaluation

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

with $\theta(x,z,0) = 1$. We have that

$$g(x,z) = \sum_{t=0}^{\infty} \beta^t \theta(x,z,t) a_t(x,z), \qquad (19)$$

$$f(x,z) = \sum_{t=0}^{\infty} \beta^t \theta(x,z,t) (\phi a)_t(x,z). \qquad (20)$$

Since the series (19) and (20) cannot be calculated in closed form, we must evaluate them numerically by truncating them. From this, we can approximately compute the index $\lambda^*(x)$ via (16) for $\phi_\infty^{(1)} < x < \phi_\infty^{(0)}$.

Case III: $z \in \big[\phi_\infty^{(0)}, 1\big]$*:* In this case, X_t remains in the passive set $B^c(z) = (0,z]$ after first hitting it. For $x > z$, let $\tau^*(x,z) \triangleq \min\{t \geqslant 1 \colon X_t \leqslant z\}$ be the first hitting time to $B^c(z)$ starting from x. Note that $\tau^*(x,z)$ is a random variable with maximum value $t_1^*(x,z) \triangleq \min\{t \geqslant 1 \colon \phi_t^{(1)}(x) \leqslant z\}$. Then, we have that, for $x > z$,

$$g(x,z) = \sum_{t=0}^{t_1^*(x,z)-1} \beta^t \theta(x,z,t), \tag{21}$$

$$f(x,z) = \sum_{t=0}^{t_1^*(x,z)-1} \beta^t \theta(x,z,t) R(\phi_t^{(1)}(x), 1), \tag{22}$$

where $\theta(x,z,t)$ is the survival probability as in Case I. From the above and (13)–(14), we can compute the $w(x,z)$ and $r(x,z)$ for $x > z$ by computing finite sums. Further, for $x \leqslant z$ it is readily seen that $w(x,z) = 1$ and $r(x,z) = R(x,1)$. Therefore, the index in (16) reduces to

$$\lambda^*(x) = R(x,1), \quad \phi_\infty^{(0)} \leqslant x \leqslant 1 \tag{23}$$

B. Verification of PCL-indexability and Index Evaluation

Based on the results in Section IV-A and on further work not shown here, we present the following conjecture.

Conjecture 1. *The single-site search problem (9) is PCL-indexable for $\beta \in [0, \beta^*)$, with*

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

Therefore, under Conjecture 1, the index $\lambda^*(x)$ calculated above is the Whittle index.

V. Computational Experiments

A. Index Evaluation

The index was computed using a Matlab script based on the results in Sec. IV for a target instance with $q^{(0)} = 0.1$, $p^{(0)} = 0.5$, $q^{(1)} = 0.5$, $p^{(1)} = 0.3$, $r = 1$, and $\alpha = 0.05$. The fixed points are thus $\phi_\infty^{(1)} = 0.3043$ 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 $\beta^* = 0.7472$. For each β, the index $\lambda^*(x)$ 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 $T = 10^4$.

Fig. 1 plots the results. Note that the index $\lambda^*(x)$ is continuous in x and piecewise differentiable, converging as $\beta \nearrow 1$ to a limiting index. Note also that, for small enough x, the index $\lambda^*(x)$ is negative, reflecting the intuition that it is counterproductive to search a site when it is very unlikely that the target is visible, as doing so will only drive the target into hiding, delaying the hunt. For each x, the time expended to compute $\lambda^*(x)$ was negligible.

B. PCL-indexability

This section presents computational evidence for the validity of Conjecture 1 for the target instance analyzed in Sec. V-A. As required by the PCL-indexability condition (ii), Fig. 1 shows that in each case the index $\lambda^*(x)$ is strictly increasing in x. Regarding condition (i), Fig. 2 shows the marginal work measure $w(x,z)$ for fixed threshold values z in $\{0.05, 0.5, 0.85\}$, letting x vary in $\bar{\mathsf{X}}$, analyzing a z value for each of the three cases described in Sec. IV-A. The discount factor β varies 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 $T = 10^4$. Fig. 2 illustrates how $w(x,z)$ differs for each threshold case considered. Further, notice that, in the examples for Case I ($z = 0.05$) and Case II ($z = 0.5$), the marginal work measure positivity condition only holds for $\beta \leqslant 0.8$.

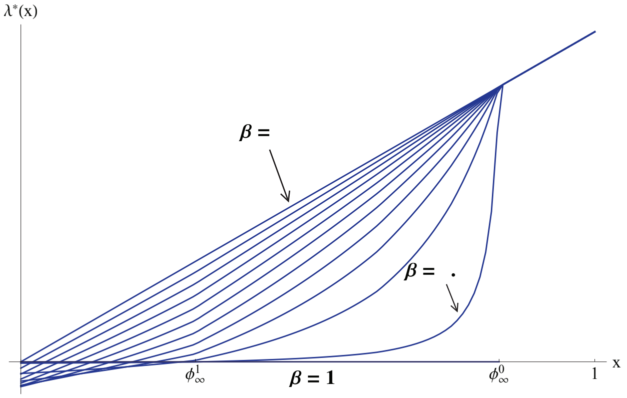


Fig. 1: The index $\lambda^*(x)$ for different discount factors β.

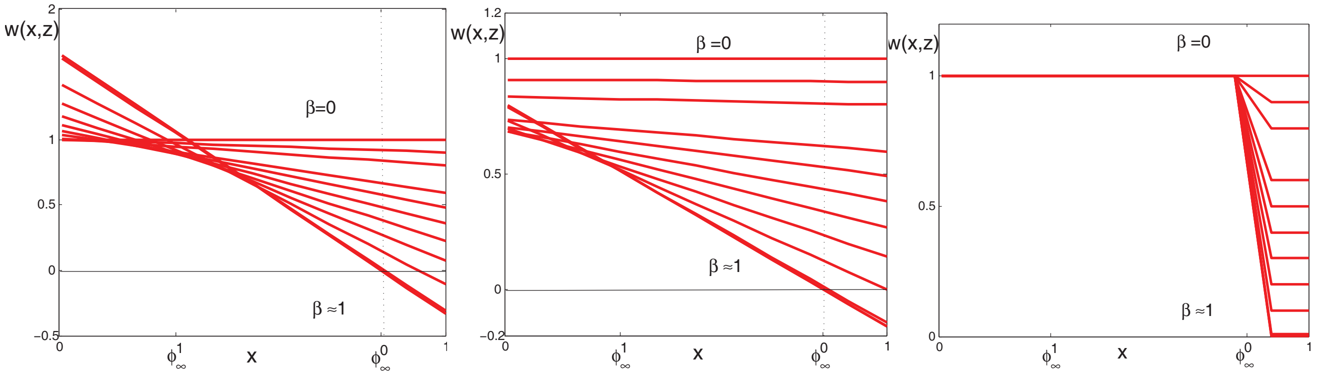
C. Benchmarking the Whittle Index Policy

We have performed some small-scale preliminary simulation studies, where the performance of the proposed Whittle index policy is compared against simpler policies: the *myopic* policy, based on index $\lambda^M(x) = R(x,1)$, which corresponds with the case $\beta = 0$, the *belief state* policy, based on index $\lambda^B(x) = x$, and the *random* selection policy which picks a site at random, with each site having the same probability of being selected.

Experiment #1: *Cautious and Reckless Targets*

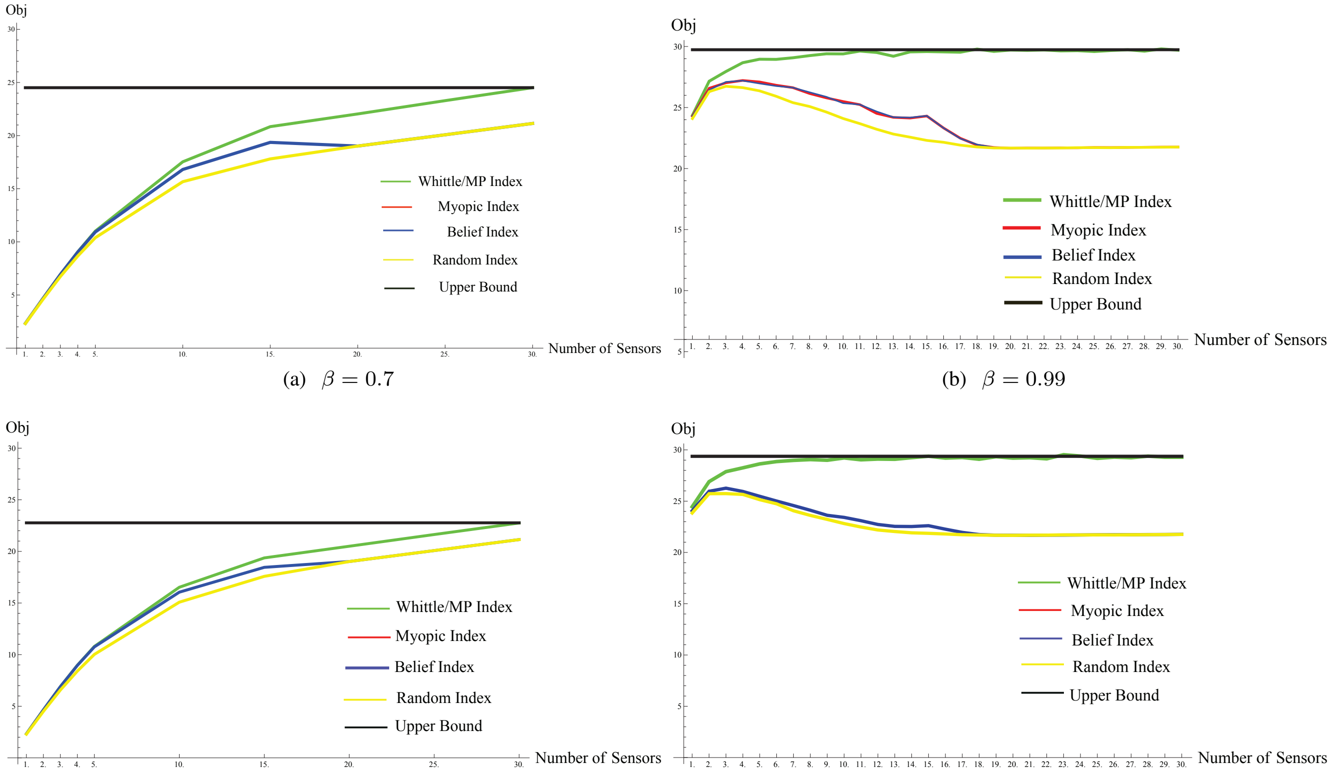
In this experiment we assess the relative performance of the Whittle index policy against the other heuristics distinguishing target instances between *reckless* and *cautious*. We call reckless those targets that "*after not being searched, are highly likely to expose themselves*", i.e., with $p^{(0)} \approx 1$, while cautious targets display the opposite behavior, i.e., have $p^{(0)} \approx 0$ (while having $p^{(0)} > p^{(1)}$).

Each base instance has a single sensor $M = 1$ for $N = 30$ 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. Also, 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 0.9491 both for the reckless and cautious instance. Both base instances were modified, letting the number of sensors increase from $M = 1$



(a) $w(x, z = 0.05)$ (b) $w(x, z = 0.5)$ (c) $w(x, z = 0.85)$

Fig. 2: Marginal work measure for the three threshold cases.



(a) $\beta = 0.7$ (b) $\beta = 0.99$

(c) $\beta = 0.7$ (d) $\beta = 0.99$

Fig. 3: Experiment 1: (3a) & (3b) *Reckless Targets* instances and (3c) & (3d): *Cautious Targets* instances

up to $M = N = 30$. For each instance, 10^3 independent runs were performed on a horizon of $T = 10^4$ time slots.

Fig 3 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 . The Whittle index policy outperforms other heuristic policies for any number of sensors with the performance improvement increasing as $M \to N$. An interesting result is that the Whittle index policy optimality loss goes to 0 for a relatively small number of sensors when $\beta \to 1$, while the largest sensor network size is required for the Whittle index policy to be nearly optimal for smaller β (i.e., when hunting targets is urgent). Notice that as the number of sensors grows, all other policies perform worse, since they *overuse* the network resources thus making targets more elusive and hence, more difficult to hunt.

Table I shows the average time that the system takes to hunt all targets operated under each policy. Such results illustrate

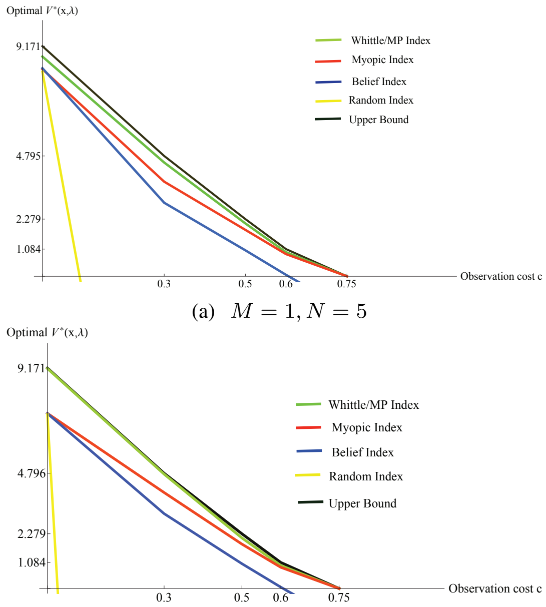
TABLE I: Average Time to Hunt All Targets

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

the fact that a large sensor network which is constantly searching will spend a larger slot of time to hunt targets. However, all policies succeed at finding the N targets at some slot. The Whittle 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.

Experiment #2: *Sensing Cost & Sensor Network Size*

In this experiment we assess the relative performance of the Whittle index policy against the other heuristics when the sensing cost increases. We consider two base instances of $N = 30$ sites with $M = 1$ and $N = 5$ 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 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. Fig 4 shows the ETD net rewards under each policy and the upper bound as c grows. Results show that the Whittle index policy outperforms the other policies in all instances. The resulting performance and its upper bound decrease with c with all policies yielding 0 rewards for $c > 0.75$. Notice that the Whittle index policy is nearly optimal for all values of the sensing cost when $M = 5$ while the optimality loss of the other heuristics is larger for $M = 5$ than for $M = 1$.



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

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

Fig. 4: Experiment 2: Sensing Cost Effect with: $M/N = 1/5$ (4a) and $M/N = 5/5$ (4b)

VI. 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] W. Moran, S. Suvorova, and S. Howard, "Application of sensor scheduling concepts to radar," in *Foundations and Applications of Sensor Management*, A. O. Hero, D. Castañón, D. Cochran, and K. Kastella, Eds. New York, NY: Springer, 2008, ch. 10, pp. 221–256.

[2] R. B. Washburn, "Application of multi-armed bandits to sensor management," in *Foundations and Applications of Sensor Management*, A. O. Hero, D. Castañón, D. Cochran, and K. Kastella, Eds. New York, NY: Springer, 2008, ch. 10, pp. 153–175.

[3] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.*, vol. 25A, pp. 287–298, 1988.

[4] B. F. La Scala and B. Moran, "Optimal target tracking with restless bandits," *Digit. Signal Process.*, vol. 16, pp. 479–487, 2006.

[5] J. Le Ny, M. Dahleh, and E. Feron, "Multi-UAV dynamic routing with partial observations using restless bandit allocation indices," in *Proc. 2008 American Control Conf.* New York, NY: IEEE, 2008, pp. 4220–4225.

[6] J. Niño-Mora, "An index policy for dynamic fading-channel allocation to heterogeneous mobile users with partial observations," in *Proc. NGI 2008, 4th Euro-NGI Conf. Next Generation Internet Networks (Krakow, Poland)*. New York, NY: IEEE, 2008, pp. 231–238.

[7] ——, "A restless bandit marginal productivity index for opportunistic spectrum access with sensing errors," in *Proc. NET-COOP 2009, 3rd Euro-NG Conf. Network Control and Optimization (Eindhoven, The Netherlands)*, ser. Lecture Notes Comput. Sci. Berlin: Springer, 2009, pp. 60–74.

[8] J. Niño-Mora and S. S. Villar, "Multitarget tracking via restless bandit marginal productivity indices and Kalman filter in discrete time," in *Proc. 2009 CDC/CCC, Joint 48th IEEE Conf. Decision and Control and 28th Chinese Control Conf. (Shanghai, China)*. New York, NY: IEEE, 2009, pp. 2905–2910.

[9] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, pp. 5547–5567, 2010.

[10] C. Kreucher, D. Blatt, A. Hero, and K. Kastella, "Adaptive multimodality sensor scheduling for detection and tracking of smart targets," *Digit. Signal Process.*, vol. 16, pp. 546–567, 2006.

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

[12] C. O. Savage and B. F. La Scala, "Sensor management for tracking smart targets," *Digit. Signal Process.*, vol. 19, pp. 968–977, 2009.

[13] J. E. Rucker, "Using agent-based modeling to search for elusive hiding targets," Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, 2006.

[14] J. Niño-Mora, “Restless bandits, partial conservation laws and indexability,” *Adv. Appl. Probab.*, vol. 33, pp. 76–98, 2001.

[15] ——, “Dynamic allocation indices for restless projects and queueing admission control: A polyhedral approach,” *Math. Program.*, vol. 93, pp. 361–413, 2002.

[16] ——, “Dynamic priority allocation via restless bandit marginal productivity indices,” *Top*, vol. 15, pp. 161–198, 2007.