Multi-UAV Dynamic Routing with Partial Observations using Restless Bandit Allocation Indices

Jerome Le Ny, Munther Dahleh and Eric Feron

Abstract— **Motivated by the type of missions currently performed by unmanned aerial vehicles, we investigate a discrete dynamic vehicle routing problem with a potentially large number of targets and vehicles. Each target is modeled as an independent two-state Markov chain, whose state is not observed if the target is not visited by some vehicle. The goal for the vehicles is to collect rewards obtained when they visit the targets in a particular state. This problem can be seen as a type of restless bandits problem with partial information. We compute an upper bound on the achievable performance and obtain in closed form an index policy proposed by Whittle. Simulation results provide evidence for the outstanding performance of this index heuristic and for the quality of the upper bound.**

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) are actively used for military operations and considered for civilian applications such as environmental monitoring. Technological advances in this area have been impressive, and it seems now that a major challenge for future developments will be to increase the degree of automation of these systems [1]. For this we need solutions with acceptable levels of performance to difficult optimization problems, such as variants of the weapon-target assignment problem [2]. Often, the problems solved are static combinatorial optimization problems resulting in open-loop policies. Yet, for most applications of UAVs, involving surveillance and monitoring, we would like to factor into the decision making process the (stochastic) evolution of the environment, which results in even harder stochastic control problems.

In this paper, we consider the following scenario. A group of M mobile sensors (also denoted agents in the following) is tracking the states of $N > M$ sites. We discretize time. At each period, each site can be in one of two states $\{s_1, s_2\}$, but we only know the state of a site with certainty if we actually visit it with a sensor. For $i \in \{1,\dots,N\}$, the state of site i changes from one period to the next according to a Markov chain with known transition probability matrix P^i, independently of the fact that a sensor is present or not, and independently of the other sites. To specify P^i, it is sufficient to give P^i_{11} and P^i_{21}, which are the probabilities of transition to state s_1 from state s_1 and s_2 respectively. When a sensor explores site i, it can observe its state *without measurement error*, and obtains a reward R^i if the site is in state s_1. There is no cost for moving the agents between the sites. We want to determine how we should allocate the agents at each time period, in order to maximize an expected total discounted cost over an infinite horizon.

This problem is related to various sensor management problems. These problems have a long history [3], [4], but have enjoyed a renewed interest more recently. Close to the ideas of this work, we mention the use by Krishnamurthy and Evans [5], [6] of Gittins' solution to the multi-armed bandit problem to direct a radar beam towards multiple moving targets. La Scala and Moran [7] suggest to use instead for a similar problem the restless bandits model, as we do here. However, in the restricted symmetric cases that [7] considers, the greedy solution is optimal and Whittle's indices and upper bound are not computed. Whittle already mentioned the potential application of restless bandits to airborne sensor routing in his original paper [8]. Recently, a slightly more general version of our problem was considered independently by Guha et al. [9], in the average-cost setting, to schedule transmissions on wireless communication channels in different states. These authors propose a policy that is different from Whittle's and offers a performance guarantee of 2.

Let us start by briefly recalling the multi-armed bandit problem (MABP) and restless bandits problem (RBP). The classical MABP concerns N sites or projects, where the state of project i at discrete time t is x^i_t. At each time t, only one project can be worked on. Then a reward $r^i(x^i_t)$ is received, and the state x^i_t evolves to x^i_{t+1} according to a known Markov rule specific to project i. The $N-1$ projects that are not operated produce no reward and their states do not change. The important result of Gittins [10], [11] is that the rich structure of this problem makes possible an efficient solution. Optimal policies turn out to have the form of an index rule. That is, we can compute independently for each project an index $\lambda^i(x^i_t) \in \mathbb{R}$ such that the optimal policy is to operate at each period the project with the maximal index.

The assumptions made in the MAPB inhibit its applicability for the sensor management problem. Suppose one has to track the state of N targets evolving independently. First, the MABP solution helps scheduling only one sensor, since only one target can be worked on at each period. Moreover, even if one does not make new measurements on a specific target, its information state still has to be updated using the known dynamics of the true state. This violates the assumption that the projects that are not operated remain frozen. To use Gittins' result for this problem, [5] must assume that the dynamics of the targets are slow and that the propagation step of the filters can be neglected for unobserved targets.

This work was supported by Air Force - DARPA - MURI award 009628-001-03-132 and Navy ONR award N00014-03-1-0171.

J. Le Ny and M. Dahleh are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA jleny@mit.edu, dahleh@mit.edu.

E. Feron is with the School of Aerospace Engineering, Georgia Tech, Atlanta, GA 30332, USA eric.feron@aerospace.gatech.edu

To overcome the shortcomings of the MABP, Whittle introduced the RBP [8]. In this problem, we now allow for M projects to be simultaneously operated, rewards can be generated for the projects that are not active, and most importantly these projects are also allowed to evolve, possibly according to different transition rules. These less stringent assumptions are very useful for the sensor management problem, but unfortunately the RBP is now known to be intractable, in fact PSPACE-hard [12], even if $M = 1$ and we only allow deterministic transition rules. Nonetheless, Whittle investigated an interesting relaxation and index policy for this problem, which extends Gittins' and which we will review in section IV in our specific context. The relaxation technique has been used apparently independently for sensor management problems by Castañón [13], [14], who does not develop index policies however. [15] also investigates the relaxation technique in a more general setting.

The rest of the paper is organized as follows. In section II, we give a precise formulation of our problem. In section III, we provide a counter example showing that the obvious candidate greedy solution to the problem is not optimal. Section IV gives a general discussion of our proposed solution to this sensor routing problem. Whittle's method is discussed with an emphasis on computations and from the point of view of constrained Markov decision processes [16]. An upper bound on the achievable performance is obtained by solving a relaxed problem using a Lagrangian approach and subgradient optimization. A lower bound is obtained by computing Whittle's index policy. The computation of Whittle's indices is non trivial in general, and the indices may not always exist. However, in section V we show the indexability of our particular problem by obtaining a closed form expression of Whittle's indices, which is the main result of the paper. We also obtain in closed form the subgradient necessary for the computation of the upper bound on achievable performance. Finally in section VI, we verify experimentally the high performance of the index policy by comparing it to the upper bound for problems involving a large number of targets and vehicles.

II. Problem Formulation

For the dynamic optimization problem described in the introduction, the state of the N sites at time t is $x_t = (x_t^1, \ldots, x_t^N) \in \{s_1, s_2\}^N$, and the control is to decide which M sites to observe. An action at time t can only depend on the information state I_t which consists of the actions $a_0, \ldots, a_{t-1}$ at previous times as well as the observations $y_0, \ldots, y_{t-1}$ and the prior information y_{-1} on the initial state x_0. We represent an action a_t by the vector $(a_t^1, \ldots, a_t^N) \in \{0,1\}^N$, where $a_t^i = 1$ if site i is visited by a sensor at time t, and $a_t^i = 0$ otherwise.

Assume the following flow of events. Given our current information state, we make the decision as to which M sites to observe. The rewards are obtained depending on the states observed, and the information state is updated. Once the rewards have been collected, the states of the sites evolve according to the known transition probabilities.

Let p be a given probability distribution on the initial state x_0. We assume independence of the initial distributions, i.e.,

$$P(x_0^1 = s^1, \ldots, x_0^N = s^N) = p(s^1, \ldots, s^N)$$
$$= \prod_{i=1}^{N} (p_{-1}^i)^{1\{s^i = s_1\}} (1 - p_{-1}^i)^{1\{s^i = s_2\}},$$

for some given numbers $p_{-1}^i \in [0,1]$. We denote by $1\{\cdot\}$ the indicator function. For an admissible policy π, i.e., depending only on the information process, we denote E_p^π the expectation operator. We want to maximize over the set Π of admissible policies the expected infinite-horizon discounted reward (with discount factor α)

$$J(p, \pi) = E_p^\pi \left\{ \sum_{t=0}^{\infty} \alpha^t r(x_t, a_t) \right\}, \tag{1}$$

where

$$r(x_t, a_t) = \sum_{i=1}^{N} R^i \, 1\{a_t^i = 1, x_t^i = s_1\},$$

and subject to the constraint

$$\sum_{i=1}^{N} 1\{a_t^i = 1\} = M, \forall t. \tag{2}$$

It is well known that we can reformulate this problem as an equivalent Markov decision process (MDP) with complete information [17]. A sufficient statistic for this problem is given by the conditional probability $P(x_t | I_t)$, so we look for an optimal policy of the form $\pi_t(P(x_t | I_t))$. An additional simplification in our problem comes from the fact that the sites are assumed to evolve independently. Let p_t^i be the probability that site i is in state s_1 at time t, given I_t. A simple sufficient statistic at time t is then $(p_t^1, \ldots, p_t^N) \in [0,1]^N$.

We have the following recursion:

$$p_{t+1}^i = \begin{cases} P_{11}^i, \text{ if site } i \text{ is visited at time } t \text{ and found} \\ \text{in state } s_1. \\ P_{21}^i, \text{ if site } i \text{ is visited at time } t \text{ and found} \\ \text{in state } s_2. \\ f^i(p_t^i) := p_t^i P_{11}^i + (1 - p_t^i) P_{21}^i \\ = P_{21}^i + p_t^i (P_{11}^i - P_{21}^i), \text{ if site } i \text{ is not visited} \\ \text{at time } t. \end{cases} \tag{3}$$

III. Non-Optimality of the Greedy Policy

We can first try to solve the problem formulated above with a general purpose solver for partially observable MDPs. However, the computations become quickly intractable, since the size of the underlying state space increases exponentially with the number of sites. Moreover, this approach would not take advantage of the structure of the problem, notably the independent evolution of the sites. We would like to use this structure to design optimal or good suboptimal policies more efficiently.

There is an obvious candidate solution to this problem, which consists in selecting at each period the M sites for which $p_t^i R^i$ is the highest. Let us call this policy the "greedy

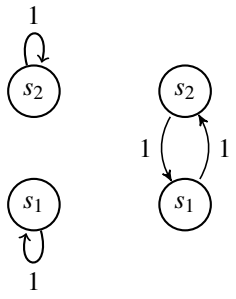

Fig. 1. Counter Example.

policy". It is not optimal in general. To see this, it is sufficient to consider a simple example with completely deterministic transition rules but uncertainty on the initial state. This underlines the importance of exploring at the right time.

Consider the example shown on Fig. 1, with $N=2$, $M=1$. Assume that we know already at the beginning that site 1 is in state s_1, i.e., $p^1_{-1}=1$. Hence we know that every time we select site 1, we will receive a reward R^1, and in effect this makes state s_2 of site 1 obsolete. Assume $R^1 > p^2_{-1}R^2$, but $(1-p^2_{-1})R^2 > R^1$, i.e., $R^2 - R^1 > p^2_{-1}R^2$. Let us denote $p^2_{-1} := p^2$ for simplicity. The greedy policy, with associated reward-to-go J_g, first selects site 1, and we have

$$J_g(1,p^2) = R^1 + \alpha J_g(1,1-p^2).$$

During the second period the greedy policy chooses site 2. Hence

$$J_g(1,1-p^2) = (1-p^2)R^2 + \alpha(1-p^2)J_g(1,0) + \alpha p^2 J_g(1,1).$$

Note that $J_g(1,0)$ and $J_g(1,1)$ are also the optimal values for the reward-to-go at these states, because the greedy policy is obviously optimal once all uncertainty has been removed. It is easy to compute

$$J_g(1,0) = \frac{R^1 + \alpha R^2}{1-\alpha^2}, \quad J_g(1,1) = \frac{R^2 + \alpha R^1}{1-\alpha^2}.$$

Now suppose we sample first at site 2, removing the uncertainty, and then follow the greedy policy, which is optimal. We get for the associated reward-to-go:

$$J(1,p^2) = p^2R^2 + \alpha p^2 J_g(1,0) + \alpha(1-p^2)J_g(1,1).$$

We compute the difference and obtain after some calculations:

$$J - J_g = p^2R^2 - R^1 + \alpha p^2 R^1.$$

For example, we can take $R^2 = 3R^1$, $p^2 = (1-\varepsilon)/3$, for a small $\varepsilon > 0$. We get $p^2R^2 = R^1(1-\varepsilon) < R^1$ and $(1-p^2)R^2 = (2-\varepsilon)R^1 > R^1$ so our assumptions are satisfied. Then $J - J_g = \frac{\alpha}{3}R^1(1-\varepsilon-\frac{3\varepsilon}{\alpha})$, which can be made positive for ε small enough, and as large as we want by simply scaling the rewards. Hence in this case it is better to first inspect site 2 than to follow the greedy policy from the beginning.

IV. Restless Bandits

The optimization problem (1) subject to the resource constraint (2) seems difficult to solve directly. However one can obtain an upper bound on the achievable performance by relaxing the constraint (2) to enforce it only on average. More specifically, we replace it by the following constraint

$$E^\pi_p\left\{\sum_{t=0}^{\infty}\alpha^t\sum_{j=1}^{N}1\{a^i_t=1\}\right\} = \frac{M}{1-\alpha},$$

or equivalently by

$$D(p,\pi) = E^\pi_p\left\{\sum_{t=0}^{\infty}\alpha^t\sum_{i=1}^{N}1\{a^i_t=0\}\right\} = \frac{N-M}{1-\alpha}. \quad (4)$$

Clearly (4) is implied by (2), so solving the optimization problem (1) with relaxed constraint (4) indeed provides an upper bound on the achievable performance. This relaxed problem can now be solved using the tools available for constrained MDPs. The two main (dual) approaches are a direct linear programming formulation on the set of occupation measures, or a Lagrangian approach using dynamic programming ideas [16]. In addition to solving the relaxed problem, we would also like to use its solution to obtain a feasible policy for the original problem. We do this by using the additional restless bandits structure.

To study the restless bandits problem, Whittle used the Lagrangian approach for the constrained MDP, which we also follow here. The following results can be found in [16, chapter 3]. Define the Lagrangian

$$L(p,\pi,\lambda) = J(p,\pi) + \lambda\left(D(p,\pi) - \frac{N-M}{1-\alpha}\right),$$

with $\lambda \in \mathbb{R}$ a Lagrange multiplier. Then the optimal reward for the problem with averaged constraint satisfies

$$J^*(p) = \sup_{\pi\in\Pi}\inf_\lambda L(p,\pi,\lambda) = \sup_{\pi\in\Pi_S}\inf_\lambda L(p,\pi,\lambda),$$

where Π_S is the set of stationary Markov (randomized) policies. Since we allow for randomized policies, a classical minimax theorem allows us to interchange the sup and the inf to get

$$J^*(p) = \inf_\lambda\left\{J^*(p;\lambda) - \lambda\frac{N-M}{1-\alpha}\right\} \quad (5)$$

where

$$J^*(p;\lambda) = \sup_{\pi\in\Pi_D}\{J(p,\pi) + \lambda D(p,\pi))\} \quad (6)$$
$$= \sup_{\pi\in\Pi_D} E^\pi_p\left\{\sum_{t=0}^{\infty}\alpha^t\sum_{i=1}^{N}R^i\,1\{a^i_t=1, x^i_t=s_1\} + \lambda 1\{a^i_t=0\}\right\},$$

and Π_D is now the set of stationary deterministic policies. For a fixed λ, $J^*(p;\lambda)$ can be computed using dynamic programming, and the possibility to restrict to deterministic policies is a classical result for unconstrained dynamic programming. Moreover, the computation of $J^*(p;\lambda)$ has the interesting

property of being separable by site. Hence we can solve the dynamic programming problem for each site separately:

$$J^*(p;\lambda) = \sum_{i=1}^{N} J^{*,i}(p;\lambda)$$
$$J^{*,i}(p^i;\lambda) = \max\left\{\lambda + \alpha J^{*,i}(f^i p^i;\lambda), \right.$$
$$\left. p^i R^i + \alpha p^i J^{*,i}(P_{11}^i;\lambda) + \alpha(1-p^i) J^{*,i}(P_{21}^i;\lambda)\right\},$$

the second equation being Bellman's equation for site i.

We can now finish the computation of the upper bound using standard dual optimization methods. Suppose that we are given a prior p on the initial states of the sites. The dual function, which we would like to minimize over λ, is

$$G(p;\lambda) = J^*(p;\lambda) - \lambda \frac{N-M}{1-\alpha}.$$

G is a convex function of λ, although in general not differentiable. We can solve the minimization problem (5) using the subgradient method, although an even simpler method such as a line search would also be possible. We have the following well-known result, see e.g. [18]:

Theorem 1: A subgradient of $G(p;\cdot)$ at λ is

$$D(p,\pi_\lambda^*) - \frac{N-M}{1-\alpha} = \sum_{i=1}^{N} D^i(p^i,\pi_\lambda^{*,i}) - \frac{N-M}{1-\alpha}, \qquad (7)$$

where π_λ^* is an optimal policy for the problem (6) (which can be decomposed into optimal policies $\pi_\lambda^{*,i}$ for each site), and

$$D^i(p^i,\pi_\lambda^{*,i}) = E_{p^i}^{\pi_\lambda^{*,i}}\left\{\sum_{t=0}^{\infty} \alpha^t 1\{a_t^i = 0\}\right\}.$$

We will see in section V that an expression for $D(p,\pi_\lambda^*)$ is obtained at no additional cost once we have an expression for $J^*(p;\lambda)$.

So far however, we have only provided a means to compute an upper bound on the achievable performance. It remains to find a good policy for the original, path constrained problem. Whittle proposed an index policy which generalizes Gittins' policy for the multi-armed bandit problem and emerges naturally from the Lagrangian relaxation. We underline here only the key ideas and refer the reader to [8] for more details and motivations behind this heuristic.

To compute Whittle's indices, we consider the bandits (or targets) individually. Hence we isolate bandit i, consider the computation problem for $J^{*,i}(p^i;\lambda)$ and drop the superscript identifier i for simplicity. λ can be viewed as a "subsidy for passivity", which parametrizes a collection of MDPs. Let us denote by $\mathscr{P}(\lambda) \subset [0,1]$ the set of information states p of the bandit such that the passive action is optimal, i.e.,

$$\mathscr{P}(\lambda) = \{p \in [0,1] : \lambda + \alpha J^*(fp;\lambda) \geq pR + \alpha p J^*(P_{11};\lambda) + \alpha(1-p) J^*(P_{21};\lambda)\}.$$

Definition 2: A bandit is *indexable* if $\mathscr{P}(\lambda)$ is monotonically increasing from $\emptyset$ to $[0,1]$ as λ increases from $-\infty$ to $+\infty$, i.e.,

$$\lambda_1 \leq \lambda_2 \Rightarrow \mathscr{P}(\lambda_1) \subseteq \mathscr{P}(\lambda_2).$$

Hence a bandit is indexable if the set of states for which it is optimal to take the passive action increases with the subsidy for passivity. This requirement seems very natural. Yet Whittle provided an example showing that it is not always satisfied, and typically showing the indexability property for particular cases of the RB problem is challenging, see e.g. [19], [20]. However, when this property could be established, Whittle's index policy, which we now describe, was found empirically to perform outstandingly well. [21] also studied a form of asymptotic optimality for this heuristic.

Definition 3: If a bandit is indexable, its *Whittle index* is given, for any $p \in [0,1]$, by

$$\lambda(p) = \inf\{\lambda \in \mathbb{R} : p \in \mathscr{P}(\lambda)\}.$$

Hence, if the bandit is in state p, $\lambda(p)$ is the value of the subsidy λ which renders the active and passive actions equally attractive. Then, restoring the superscripts i for the N bandits, and assuming that each bandit is indexable, we obtain for state $(p_t^1,\ldots,p_t^N)$ a set of indices $\lambda^1(p_t^1),\ldots,\lambda^N(p_t^N)$. The index heuristic applies at each period t the active action to the M projects with largest indices $\lambda^i(p_t^i)$, and the passive action to the remaining $N-M$ projects.

V. Indexability and Computation of Whittle's Indices

A. Preliminaries

In this section we give an overview of the study of the indexability property for each site. Due to space constraints, most of the computations are not presented. The interested reader can find them in our technical report [22]. For the sensor management problem considered in this paper, we show that the bandits are indeed indexable and compute the Whittle indices in closed form.

Since the discussion is concerned with a single site, we drop the superscript i. For reference we rewrite Bellman's equation of optimality for this problem. If J is the optimal value function, then

$$J(p) = \max\{\lambda + \alpha J(fp), pR + \alpha p J(P_{11}) + \alpha(1-p) J(P_{21})\} \qquad (8)$$

where $fp := pP_{11} + (1-p)P_{21} = P_{21} + p(P_{11} - P_{21})$.

Note that for simplicity, we dropped the λ and the $*$ from the previous notation, i.e., $J(p) := J^*(p;\lambda)$. First we have

Theorem 4: J is a convex function of p, continuous on $[0,1]$.

Proof: It is well known that we can obtain the value function by value iteration as a uniform limit of cost functions for finite horizon problems, which are continuous, piecewise linear and convex, see e.g. [23]. The uniform convergence follows from the fact that the discounted dynamic programming operator is a contraction mapping. The convexity of J follows, and the continuity on the closed interval $[0,1]$ is a consequence of the uniform convergence. ■

Lemma 5: 1) When $\lambda \leq pR$, it is optimal to take the active action. In particular, if $\lambda \leq 0$, it is always optimal

to take the active action and J is affine:

$$J(p) = \alpha J(P_{21}) + p[R + \alpha(J(P_{11}) - J(P_{21}))] = \frac{(\alpha P_{21} + p(1-\alpha))R}{(1-\alpha)(1-\alpha(P_{11}-P_{21}))}. \quad (9)$$

2) When $\lambda \geq R$, it is always optimal to take the passive action, and

$$J(p) = \frac{\lambda}{1-\alpha}. \quad (10)$$

Proof: By convexity of J, $J(fp) \leq pJ(P_{11}) + (1-p)J(P_{21})$ and so for $\lambda \leq pR$, it is optimal to choose the active action. The rest of 1 follows by easy calculation, solving first for $J(P_{11})$ and $J(P_{21})$. To prove 2, use value iteration, starting from $J_0 = 0$. ∎

With this lemma, it is sufficient to consider from now on the situation $0 < \lambda < R$.

Lemma 6: The set of $p \in [0,1]$ where it is optimal to choose the active action is convex, i.e., an interval in $[0,1]$.

Proof: In the set where the active action is optimal, we have

$$J(p) = pR + \alpha pJ(P_{11}) + \alpha(1-p)J(P_{21}).$$

Consider p_1 and p_2 in this set. We want to show that for all $\beta \in [0,1]$, it is also optimal to choose the active action at $p = \beta p_1 + (1-\beta)p_2$. We know from Bellman's equation (8) that

$$pR + \alpha pJ(P_{11}) + \alpha(1-p)J(P_{21}) \leq J(p).$$

By convexity of J, we have

$$\begin{aligned} J(p) \leq & \beta J(p_1) + (1-\beta)J(p_2) \\ J(p) \leq & \beta\left(p_1 R + \alpha p_1 J(P_{11}) + \alpha(1-p_1)J(P_{21})\right) + \\ & (1-\beta)\left(p_2 R + \alpha p_2 J(P_{11}) + \alpha(1-p_2)J(P_{21})\right) \\ J(p) \leq & pR + \alpha pJ(P_{11}) + \alpha(1-p)J(P_{21}). \end{aligned}$$

Combining the two inequalities, we see that the active action is optimal at p. ∎

Lemma 7: The sets of $p \in [0,1]$ where the passive and active actions are optimal are of the form $[0,p^*]$ and $[p^*,1]$, respectively.

Proof: This follows from the convexity of the active set and the fact that the active action is optimal for $p \geq \frac{\lambda}{R}$ by lemma 5. ∎

In the following, we emphasize the dependence of p^* on λ by writing $p^*(\lambda)$. It is a direct consequence of lemma 7 and the continuity of J that $p^*(\lambda)$ is a value where the passive and the active actions are equally attractive. We also see that to show the indexability property of definition 2, it is sufficient to show that $p^*(\lambda)$ is an nondecreasing function of λ. Then, Whittle's index is obtained by inverting the relation $\lambda \to p^*(\lambda)$, i.e.,

$$\lambda(p) = \inf\{\lambda : p^*(\lambda) = p\}.$$

An interesting feature of our problem is that it is possible to compute $p^*(\lambda)$ in closed form. In addition we can also compute the value function $J(p) := J^*(p;\lambda)$ and the "discounted passivity measure" for each bandit:

$$D(p,\pi_\lambda^*) = E_p^{\pi_\lambda^*}\left\{\sum_{t=0}^{\infty} \alpha^t 1\{a_t = 0\}\right\}.$$

This last quantity is necessary to compute the subgradient (7). Its computation is a policy evaluation problem. $D(p,\pi_\lambda^*)$ obeys the equations

$$D(p,\pi_\lambda^*) = \begin{cases} \alpha pD(P_{11},\pi_\lambda^*) + \alpha(1-p)D(P_{21},\pi_\lambda^*), \\ \text{ for } p > p^*(\lambda) \\ 1 + \alpha D(fp,\pi_\lambda^*), \text{ for } p \leq p^*(\lambda). \end{cases}$$

These equations can be compared to those verified by $J^*(p;\lambda)$ once $p^*(\lambda)$ is known:

$$J^*(p;\lambda) = \begin{cases} R + \alpha pJ^*(P_{11},\lambda) + \alpha(1-p)J^*(P_{21},\lambda), \\ \text{ for } p > p^*(\lambda) \\ \lambda + \alpha J^*(fp,\lambda), \text{ for } p \leq p^*(\lambda). \end{cases}$$

Hence it is sufficient to have a closed form solution for $J^*(p;\lambda)$. To compute $D(p,\pi_\lambda^*)$, we simply formally set $R = 0$ and $\lambda = 1$ in the corresponding expression for $J^*(p;\lambda)$. For example, starting from expressions (9) and (10), we recover the (trivial) result that $D(p,\pi_\lambda^*) = 0$ if $\lambda \leq 0$ and $D(p,\pi_\lambda^*) = 1/(1-\alpha)$ if $\lambda \geq R$.

The computation of $p^*(\lambda)$, $J^*(p;\lambda)$ and $D(p,\pi_\lambda^*)$ for each bandit can be performed by distinguishing between various cases depending on the value of the parameters P_{11} and P_{21}. The computations are rather long and we omit them in this paper. Here we only give the main result, which is the expression of the Whittle indices.

Theorem 8: A two-state restless bandit as considered in this section is indexable. The index $\lambda(p)$ can be computed as follows. Let $s = P_{11} - P_{21}$ (then $-1 \leq s \leq 1$), $f^n P_{21} = P_{21}\frac{1-s^{n+1}}{1-s}$, and $I = \frac{P_{21}}{1-s}$.

1) Case $s = 0$: $\lambda(p) = pR$.
2) Case $s = 1$: $\lambda(p) = \frac{pR}{1-\alpha(1-p)}$.
3) Case $0 < s < 1$:
   - If $p \geq P_{11}$ or $p \leq P_{21}$: $\lambda(p) = pR$.
   - If $I \leq p < P_{11}$: $\lambda(p) = \frac{pR}{1-\alpha(P_{11}-p)}$.
   - If $P_{21} < p < I$: Let $k := k(p) = \lceil \frac{\ln(1-\frac{p}{I})}{\ln s} \rceil - 2$. Then let

     $$B_k = 1 - \alpha^{k+2},\ C_k = \alpha - \alpha^{k+2},$$
     $$A_k = \frac{(1-\alpha P_{11})B_k + \alpha^{k+2}(1-\alpha)(f^{k+1}P_{21})}{1-\alpha s}$$

     We have

     $$\lambda(p) = \frac{A_{k(p)} - (1-p)B_{k(p)}}{A_{k(p)} - (1-p)C_{k(p)}}R.$$

4) Case $s = -1$:
   - If $p \geq 1/2$: $\lambda(p) = \frac{\alpha + p(1-\alpha)}{1+\alpha(1-\alpha)(1-p)}R$.
   - If $p < 1/2$: $\lambda(p) = \frac{p}{1-\alpha p}R$.
5) Case $-1 < s < 0$:
   - If $p \geq P_{21}$ or $p \leq P_{11}$: $\lambda(p) = pR$.

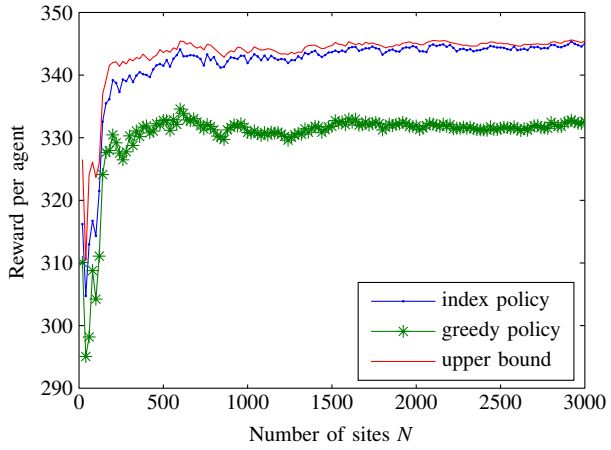

Fig. 2. Monte-Carlo Simulation for Whittle's index policy and the greedy policy. The upper bound is computed using the subgradient optimization algorithm. We fixed $\alpha = 0.95$.

- If $fP_{11} \leq p < P_{21}$: $\lambda(p) = \frac{p+\alpha(P_{21}-p)}{1+\alpha(P_{21}-p)}R$.
- If $I \leq p < fP_{11}$: $\lambda(p) = \frac{p+\alpha(P_{21}-p)}{1+\alpha(1-\alpha)(P_{21}-p)-\alpha^2 P_{11}s}R$.
- If $P_{11} < p < I$: $\lambda(p) = \frac{p}{1-\alpha(p-P_{11})}R$.

VI. Simulation Results

In this section, we briefly present some simulation results illustrating the performance of the index policy and the quality of the upper bound. We generate sites with random rewards R^i within given bounds and random parameters P_{11}, P_{21}. We progressively increase the size of the problem by adding new sites and UAVs to the existing ones. We keep the ratio M/N constant, in this case $M/N = 1/20$. When generating new sites, we only ensure that $|P_{11} - P_{21}|$ is sufficiently far from 0, which is the case where the index policy departs significantly from the simple greedy policy. The upper bound is computed for each value of N using the subgradient optimization algorithm. The expected performance of the index policy and the greedy policy are estimated via Monte-Carlo simulations.

Fig. 2 shows the result of simulations for up to $N = 3000$ sites. We plot the reward per agent, dividing the total reward by M, for readability. We can see the consistently stronger performance of the index policy with respect to the simple greedy policy, and in fact its almost optimality.

VII. Conclusion

We have proposed the application of Whittle's work on restless bandits in the context of a UAV routing problem with partial information. For given problem parameters, we can compute an upper bound on the achievable performance, and experimental results show that the performance of Whittle's index policy is often very close to the upper bound. This is in agreement with existing work on restless bandit problems for different applications. Some directions for future work include a better understanding the asymptotic performance of the index policy and the computation of the indices for more general state spaces.

References

[1] "Unmanned aircraft systems roadmap 2005-2030," Office of the Secretary of Defense, Tech. Rep., 2005. [Online]. Available: http://www.acq.osd.mil/usd/Roadmap%20Final2.pdf

[2] J. Bellingham, M. Tillerson, M. Alighanbari, and J. How, "Cooperative path planning for multiple UAVs in dynamic and uncertain environments," in *Proceedings of the 41st IEEE Conference on Decision and Control*, 2002.

[3] M. Athans, "On the determination of optimal costly measurement strategies," *Automatica*, vol. 8, pp. 397–412, 1972.

[4] L. Meier, J. Peschon, and R. Dressler, "Optimal control of measurement systems," *IEEE Transactions on Automatic Control*, vol. 12, no. 5, pp. 528–536, 1967.

[5] V. Krishnamurthy and R. Evans, "Hidden markov model multiarm bandits: a methodology for beam scheduling in multitarget tracking," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 2893 – 2908, December 2001.

[6] ——, "Correction to "hidden markov model multiarm bandits: a methodology for beam scheduling in multitarget tracking"," *IEEE Transactions on Signal Processing*, vol. 51, no. 6, pp. 1662–1663, June 2003.

[7] B. F. L. Scala and B. Moran, "Optimal target tracking with restless bandits," *Digital Signal Processing*, vol. 16, pp. 479–487, 2006.

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

[9] S. Guha, K. Munagala, and P. Shi, "On index policies for restless bandit problems," November 2007. [Online]. Available: http://arxiv.org/abs/0711.3861

[10] J. Gittins and D. Jones, "A dynamic allocation index for the sequential design of experiments," in *Progress in Statistics*, J. Gani, Ed. Amsterdam: North-Holland, 1974, pp. 241–266.

[11] J. Gittins, *Multi-armed Bandit Allocation Indices*, ser. Wiley-Interscience series in Systems and Optimization. New York: John Wiley and sons, 1989.

[12] C. Papadimitriou and J. Tsitsiklis, "The complexity of optimal queueing network control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, 1999.

[13] D. Castañón, "Approximate dynamic programming for sensor management," in *Proceedings of the 36th Conference on Decision and Control*, December 1997, pp. 1202–1207.

[14] ——, "Stochastic control bounds on sensor network performance," in *Proceedings of the 44th IEEE Conference on Decision and Control*, 2005.

[15] D. Adelman and A. J. Mersereau, "Relaxations of weakly coupled stochastic dynamic programs," *Operations Research*, 2008, accepted.

[16] E. Altman, *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

[17] D. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Athena Scientific, 2001, vol. 1 and 2.

[18] ——, *Nonlinear Programming*. Athena Scientific, 1999.

[19] J. Niño-Mora, "Restless bandits, partial conservation laws and indexability," *Advances in Applied Probability*, vol. 33, pp. 76–98, 2001.

[20] K. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride, "Some indexable families of restless bandit problems," *Advances in Applied Probability*, vol. 38, pp. 643–672, 2006.

[21] R. Weber and G. Weiss, "On an index policy for restless bandits." *Journal of Applied Probability*, vol. 27, pp. 637–648, 1990.

[22] J. Le Ny, M. Dahleh, and E. Feron, "Multi-UAV dynamic routing with partial observations using restless bandit allocation indices," LIDS, Massachusetts Institute of Technology, Tech. Rep., 2007. [Online]. Available: http://web.mit.edu/jleny/www/publications.html

[23] E. Sondik, "The optimal control of partially observable markov decision processes over the infinite horizon: Discounted costs," *Operations Research*, vol. 26, no. 2, pp. 282–304, March-April 1978.