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

Jerome Le Ny, Munther Dahleh, Eric Feron

October 8, 2007

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, although we operate here under 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.

Contents

1 Introduction

Unmanned aerial vehicles (UAVs) are already actively used for military operations and investigated for civilian applications such as environmental control and monitoring. Technological advances in this area have been impressive, yet it seems clear that a major challenge for future developments will be to increase the level of autonomy of these systems [UAV05]. For this we need solutions with acceptable levels of performance to difficult optimization problems. For example, variants of the weapon-target assignment problem have been recently studied in this context by [BTAH02]. Often, the problems solved are essentially static combinatorial optimization problem. 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 (UAVs) is tracking the states of $N > M$ sites. Each site can be in one of two states s_1 or s_2 at each period, but we only know the state of a site with certainty if we actually visit it with a sensor. However, we also know that the states of the sites evolve independently of each other and and as Markov chains with known transition matrices. Hence we can estimate the states of the sites that are not visited by any sensor at a given period. Every time a sensor visits site number i which happens to be in state s_1, we receive a reward R^i. No reward is received if the site turns out to be in state s_2. The goal is to allocate the sensors at each time period, in order to maximize an expected total discounted cost over an infinite horizon. Through this model, we capture the following trade-off. It is advantageous to keep a good estimate of the states of the sites in order to take a good decision about which sites to visit next. However, the estimation problem is not the end goal, and so there is a balance between visiting sites to gain more information and collecting rewards. For an application example, one can think of the following environmental monitoring problem. M UAVs are surveilling N ships for possible bilge water dumping. The rewards are associated with the detection of a dumping event (state s_1 for a ship). However, if the UAV is not present during the dumping, the event is missed.

The problem described above is related to various sensor management problems. These problems have a long history [Ath72, MPD67], but have enjoyed a renewed interest more recently due to the important research effort in sensor networks, see e.g. [FO90, Cas97, Cas05, GCHR06, Wil07]. Close to the ideas of this work, we mention the use by [KE01, KE03] of Gittins' solution to the multi-armed bandit problem to direct a radar beam towards multiple moving targets. One goal of this paper is to point out the relevance of Whittle's interesting extension to the multi-armed bandit problem, which he called the restless bandits problem. In fact, Whittle already mentioned the potential application to airborne sensor routing in his original paper [Whi88].

Let us start by briefly recalling the multi-armed bandit (MAB) and restless bandits (RB) models. The classical MAB problem concerns N sites or projects, where the state of project i at discrete time t is x_t^i, with values in a discrete state space. At each time t, only one project can be worked on. Then a reward $r^i(x_t^i)$ is received, and the state x_t^i evolves to x_{t+1}^i 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 [GJ74, Git89] 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 $\lambda^i(x_t^i) \in \mathbb{R}$ such that the optimal policy is to operate at each period the project with the maximal index.

The requirement of the discrete state space is not crucial. [KE01] adapted the model to a stochastic control problem with partial information. It is well-known that these problems can be solved by a transformation to a full state information control problem, where the new state however is continuous and represents a conditional probability on the original state space. The other assumptions made in the MAB model inhibit its applicability for the sensor management problem in a more fundamental way however. Suppose one has to track the state of N targets evolving independently. First, the MAB 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. Hence in [KE01] the authors have to assume that the dynamics of the targets are slow and that the propagation step of the filters can be neglected for unobserved targets. This might be a reasonable assumption for scheduling a radar beam, but not necessarily for our purpose of moving a limited number of airborne sensors to different regions.

To overcome the shortcomings of the MAB model, [Whi88] introduced the restless bandits model. 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 RB problem is now known to be intractable, in fact PSPACE-hard [PT99], 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 4 in our specific context. The relaxation technique in particular has been used more recently and apparently independently for more involved sensor management problems but with similar characteristics by Castañón [Cas97, Cas05]. Our problem can be seen as a particular case of the restless bandits problem, although under partial information. Whittle's heuristic for partial information problems has apparently not been studied previously in the literature, except for the particular case already mentioned of the multi-armed bandit problem.

The rest of the paper is organized as follows. In section 2, we give a precise formulation of our problem. In section 3, we provide a counter example showing that the obvious candidate solution to the problem is not optimal. Section 4 presents in a general setting our solution to this sensor routing problem, inspired by Whittle's method and more general work on constrained Markov decision processes [Alt99]. 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. An attractive feature of our problem however is that all computations can be carried out analytically. Hence in section 5, we show the indexability of the problem and obtain simultaneously a closed form expression of Whittle's indices. Finally in section 6, we verify experimentally the high performance of the index policy by comparing it to the upper bound for some problems involving a large number of targets and vehicles.

2 Problem Formulation

We consider the following discrete time problem. We have N sites, each of which can be in one of two states $\{s_1, s_2\}$. 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 an agent is present or not, and independently of the other sites. To specify $P^{(i)}$, it is sufficient to give $P_{11}^{(i)}$ and $P_{21}^{(i)}$, which are the probabilities of transition to state s_1 from state s_1 and s_2 respectively. We have M agents to observe the sites and obtain rewards. When an agent explores site i, it can observe its state without measurement error, and obtain 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.

The state of the N sites at time t is $x_t = (x_t^1, \dots, x_t^N) \in \{s_1, s_2\}^N$, 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 $a_0, \dots, a_{t-1}$ at previous times as well as the observations $y_0, \dots, y_{t-1}$ and the prior information y_{-1} on the initial state x_0. We represent an action a_t by the vector $(a_t^1, \dots, a_t^N) \in \{0, 1\}^N$, 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.,

$$\begin{aligned} P(x_0^1 = s^1, \dots, x_0^N = s^N) &= p(s^1, \dots, s^N) \\ &= \prod_{i=1}^{N} (p_{-1}^i)^{\mathbb{1}(s^i = s_1)} (1 - p_{-1}^i)^{\mathbb{1}(s^i = s_2)}, \end{aligned}$$

for some given numbers $p_{-1}^i \in [0, 1]$. 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 of admissible policies Π the expected infinite-horizon discounted reward

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

where

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

and subject to the constraint

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

It is well known that we can reformulate this problem as an equivalent Markov decision process (MDP) with complete information [Ber01]. 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 $(p_t^1, \dots, p_t^N) \in [0,1]^N$.

Remark. The state representation chosen here involves a uncountable state space, for which the MDP theory is usually more technical. However, in our case, little additional complexity will be introduced. It is possible to adopt a state representation with a countable state space, by keeping track for each site of the number of periods since last visit as well as the state of the site at that last visit. In addition, we need to treat separately the time periods before we visit a site for the first time. This state representation, although potentially simpler from the theoretical point of view, is notationally more cumbersome and will not be used.

We have the following recursion:

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

3 Non-Optimality of the Greedy Policy

We can first try to solve the problem formulated above with a general purpose POMDP solver. However, the computations become quickly intractable, since the size of the 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. This policy is not optimal in general, however. To show this, it is sufficient to consider a simple case with completely deterministic transition

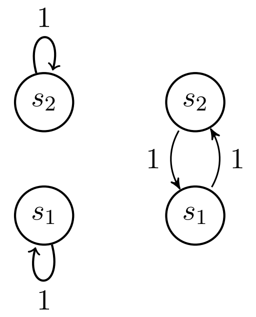

Figure 1: Counter Example.

rules but uncertainty on the initial state. This underlines the importance of removing the uncertainty 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:

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

Let us take the difference:

$$\begin{aligned}
J - J_g &= p^2 R^2 + \alpha p^2 J_g(1,0) + \alpha(1-p^2) J_g(1,1) - R^1 \\
&- \alpha(1-p^2)R^2 - \alpha^2(1-p^2)J_g(1,0) - \alpha^2 p^2 J_g(1,1) \\
&= p^2 R^2 - R^1 - \alpha(1-p^2)R^2 \\
&+ \alpha J_g(1,0)(p^2 - \alpha + \alpha p^2) + \alpha J_g(1,1)(1 - p^2 - \alpha p^2) \\
&= p^2 R^2 - R^1 - \alpha(1-p^2)R^2 \\
&+ \frac{\alpha}{1-\alpha^2}[(R^1 + \alpha R^2)(p^2 - \alpha + \alpha p^2) \\
&+ (R^2 + \alpha R^1)(1 - p^2 - \alpha p^2)] \\
&= p^2 R^2 - R^1 - \alpha(1-p^2)R^2 + \\
&\frac{\alpha}{1-\alpha^2}[R^1(p^2 - \alpha + \alpha p^2 + \alpha - \alpha p^2 - \alpha^2 p^2) \\
&+ R^2(\alpha p^2 - \alpha^2 + \alpha^2 p^2 + 1 - p^2 - \alpha p^2)] \\
&= p^2 R^2 - R^1 - \alpha(1-p^2)R^2 + \alpha p^2 R^1 + \alpha R^2(1-p^2) \\
&= p^2 R^2 - R^1 + \alpha p^2 R^1.
\end{aligned}$$

For example, we can take $R^2 = 3R^1$, $p^2 = (1-\epsilon)/3$, for a small $\epsilon > 0$. We get $p^2 R^2 = R^1(1-\epsilon) < R^1$ and $(1-p^2)R^2 = (2-\epsilon)R^1 > R^1$ so our assumptions are satisfied. Then $J - J_g = \frac{\alpha}{3} R^1 (1 - \epsilon - \frac{\alpha\epsilon}{3})$, 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.

4 Restless Bandits

The optimization problem (1) subject to the ressource 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

$$E_p^\pi \left\{ \sum_{t=0}^{\infty} \alpha^t \sum_{j=1}^{N} \mathbb{1}\{a_t^i = 1\} \right\} = \frac{M}{1-\alpha},$$

or equivalently by

$$D(p,\pi) = E_p^\pi \left\{ \sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{N} \mathbb{1}\{a_t^i = 0\} \right\} = \frac{N-M}{1-\alpha}. \tag{4}$$

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 [Alt99]. 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. Linear programming in the context of sensor management has also been used, see e.g. [YW00, LDF06]. The following results can be found in [Alt99, 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

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

where

$$\begin{aligned} J^*(p; \lambda) &= \sup_{\pi \in \Pi_D} \; \{ J(p, \pi) + \lambda D(p, \pi)) \} \qquad (6) \\ &= \sup_{\pi \in \Pi_D} \; E_p^{\pi} \left\{ \sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{N} R^i \; \mathbb{1}\{a_t^i = 1, x_t^i = s_1\} + \lambda \mathbb{1}\{a_t^i = 0\} \right\}, \end{aligned}$$

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:

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

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. Supose 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, althouth an even simpler method such as a line search would also be possible. We have the following well-known result, see e.g. [Ber99]:

Theorem 4.1. *A subgradient of $G(p; \cdot)$ at λ is*

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

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

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

We will see in section 5 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 [Whi88] 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 $\mathcal{P}(\lambda) \subset [0,1]$ the set of information states p of the bandit such that the passive action is optimal, i.e.,

$$\mathcal{P}(\lambda) = \left\{ \lambda \in \mathbb{R} : \lambda + \alpha J^*(fp;\lambda) \geq pR + \alpha p J^*(P_{11};\lambda) + \alpha(1-p) J^*(P_{21};\lambda) \right\}.$$

Definition 4.1. A bandit is *indexable* if $\mathcal{P}(\lambda)$ is monotonically increasing from $\emptyset$ to $[0,1]$ as λ increases from $-\infty$ to $+\infty$, i.e.,

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

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. [NM01, GRHK06]. However, when this property could be established, Whittle's index policy, which we now describe, was found empirically to perform outstandingly well. [WW90] also studied a form of asymptotic optimality for this heuristic.

Definition 4.2. If a bandit is indexable, its *Whittle index* is given, for any $p \in [0,1]$, by

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

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, restauring 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 Whittle indices $\lambda^1(p_t^1), \ldots, \lambda^N(p_t^N)$. Then Whittle's index heuristic applies at each period t the active action to the M projects with largest index $\lambda^i(p_t^i)$, and the passive action to the remaining $N - M$ projects.

5 Indexability and Computation of Whittle's Indices

5.1 Preliminaries

In this section we study the indexability property for each site. 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 concerns a single site, we will again drop the superscript i. A site has its dynamics completely specified by P_{21} and P_{11}. We denote the information state by p_t, i.e., p_t is the probability that the site is in state s_1, conditioned on the past. If we visit the site and it is in state 1, we get a reward $R > 0$, but if it is in state s_2, we get no reward. In the following, we call visiting the site the *active* action. Finally, if we do not visit the site, which is called the *passive* action, we collect a reward λ with probability 1. The indexability property that we would like to verify means that as λ increases, the set of information states where it is optimal to choose the passive action grows monotonically (in the sense of inclusion), from $\emptyset$ when $\lambda \to -\infty$ to $[0, 1]$ when $\lambda \to +\infty$.

For reference we rewrite Bellman's equation of optimality for this problem. If J is the optimal value function, then

$$\begin{aligned} & J(p) = \max\{\lambda + \alpha J(fp), pR + \alpha p J(P_{11}) + \alpha(1-p)J(P_{21})\} \\ & \text{where } fp := pP_{11} + (1-p)P_{21} = P_{21} + p(P_{11} - P_{21}). \end{aligned} \tag{8}$$

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 5.1. *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. [Son78]. 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 5.2. *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:*

$$\begin{aligned} J(p) &= \alpha J(P_{21}) + p[R + \alpha(J(P_{11}) - J(P_{21}))] \\ &= \frac{(\alpha P_{21} + p(1-\alpha))R}{(1-\alpha)(1-\alpha(P_{11} - P_{21}))}. \end{aligned} \tag{9}$$

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

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

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 5.3. *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

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

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 Belmann's equation (8) that

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

By convexity of J, we have

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

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

Lemma 5.4. *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 5.2. □

In the following, we emphasize the dependence of p^* on λ by writing $p^*(\lambda)$. It is a direct consequence of lemma 5.4 and the continuity of J that $p^*(\lambda)$ is the unique value where the passive and the active actions are equally attractive. We also see that to show the indexability property of definition 4.1, it is sufficient to show that $p^*(\lambda)$ is an increasing function of λ. Then, Whittle's index is obtained by inverting the relation $\lambda \to p^*(\lambda)$, i.e.,

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

In the following, we will compute $p^*(\lambda)$ explicitely, distinguishing between various cases depending on the values of the parameters P_{11} and P_{21} of the bandit. In addition we also compute the value function $J(p) := J^*(p;\lambda)$ and the following "discounted passivity measure" for each bandit:

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

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

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

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

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

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 $\lambda > 0$.

The next paragraphs of this section present the explicit computations.

5.2 Case $P_{21} = P_{11}$

This case is very easy. Let $P_{11} = P_{21} = P$. We have in particular $fp = P$, for all $p \in [0,1]$. Bellman's equation gives

$$J(p) = \max\{\lambda + \alpha J(P), pR + \alpha J(P)\},$$

so in this case, we have immediately

$$p^*(\lambda) = \frac{\lambda}{R}$$

and the project is indexable.

To give the complete expression of the value function, we only need to determine $J(P)$. There are two cases:

1. Case $P \leq (\lambda/R)$: then $J(P) = \frac{\lambda}{1-\alpha}$, and so

$$J(p) = \begin{cases} \frac{\lambda}{1-\alpha} & \text{if } p \leq \frac{\lambda}{R} \\ pR + \frac{\alpha\lambda}{1-\alpha} & \text{if } p > \frac{\lambda}{R}. \end{cases}$$

2. Case $P > (\lambda/R)$: then $J(P) = \frac{PR}{1-\alpha}$ and so

$$J(p) = \begin{cases} \lambda + \frac{\alpha PR}{1-\alpha} & \text{if } p \leq \frac{\lambda}{R} \\ \left(p + \frac{\alpha P}{1-\alpha}\right) R & \text{if } p > \frac{\lambda}{R}. \end{cases}$$

5.3 Case $P_{21} < P_{11}$

This case is more involved. We will need to consider the evolution of the state p_t.

5.3.1 Case $P_{11} = 1, \ P_{21} = 0$

Suppose first $P_{11} = 1$, $P_{21} = 0$, and recall that $0 < \lambda < R$. Then it is clear that $J(0) = \lambda/(1-\alpha)$ and $J(1) = R/(1-\alpha)$. By continuity of J, at p^* we are indifferent between the passive and active actions, so

$$J(p^*) = \lambda + \alpha J(p^*) = p^* R + \alpha p^* J(1) + \alpha(1-p^*)J(0),$$

from which $J(p^*) = \lambda/(1-\alpha)$ follows, and then

$$p^* = \frac{\lambda(1-\alpha)}{R - \alpha\lambda}.$$

Thus $p^*(\lambda)$ an increasing function of λ, since its derivative is

$$\frac{dp^*}{d\lambda} = \frac{(1-\alpha)R}{(R-\alpha\lambda)^2} \geq 0.$$

As for the value function, we get

$$J(p) = \begin{cases} \frac{\lambda}{1-\alpha} & \text{if } p \leq p^* \\ \frac{\alpha\lambda}{1-\alpha} + p\frac{R-\alpha\lambda}{1-\alpha} & \text{if } p > p^*. \end{cases}$$

5.3.2 Case $0 < P_{11} - P_{21} < 1$

In this case there is a unique point of intersection between the diagonal line and the line which is the graph of $p \to fp$, see Fig. 2. We will denote by I the abscissa in $[0,1]$ of this intersection point. Then I is defined by

$$I = fI = P_{21} + I(P_{11} - P_{21}),$$

that is,

$$I = \frac{P_{21}}{1 - (P_{11} - P_{21})},$$

which is well defined as long as we are not in the case of paragraph 5.3.1. Note that $P_{21} \leq I \leq P_{11}$. Also, we have

$$f^n p_1 - f^n p_2 = (p_1 - p_2)(P_{11} - P_{21})^n, \; \forall\, p_1, p_2 \in [0,1],$$

and in particular

$$f^n p - f^n I = f^n p - I = (p - I)(P_{11} - P_{21})^n. \tag{11}$$

So as long as $|P_{11} - P_{21}| < 1$, as in this paragraph or in paragraph 5.4.2, the distance between $f^n p$ and I decreases strictly at each iteration and $f^n p \to I$ as $n \to \infty$.

Case $p^* \geq I$:

Assume first that $p^* \geq I$, and consider p belonging to the passive region $[0, p^*]$. From (11) and the assumption $0 < P_{11} - P_{21} < 1$, we obtain a sequence of iterates $f^n p$ which remains in the passive region while converging to I. Hence we get, since J is continuous at I,

$$J(p) = \lambda + \alpha J(fp) = \lambda + \alpha\lambda + \alpha^2 J(f^2 p) = \ldots = \frac{\lambda}{1-\alpha}.$$

So for instance, $J(P_{21}) = J(p^*) = \frac{\lambda}{1-\alpha}$, since $P_{21} \leq I \leq p^*$ implies that P_{21} belongs to the passive region.

Now for $p > p^*$, we have

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

There are two subcases. First if $P_{11} \leq p^*$, then we get

$$J(p) = pR + \frac{\alpha\lambda}{1-\alpha}.$$

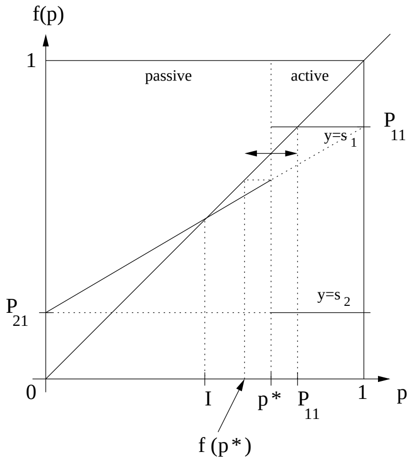

Figure 2: Case $I < p^* < P_{11}$. The line joining P_{21} and P_{11} is $p \to f(p)$. In the active region, we have $p_{t+1} = P_{11}$ or P_{21} depending on the observation.

Continuity of J at p^*, gives $p^* R + \frac{\alpha\lambda}{1-\alpha} = \frac{\lambda}{1-\alpha}$ and so

$$p^*(\lambda) = \frac{\lambda}{R}. \tag{12}$$

This is the expression in the case $\frac{\lambda}{R} \geq P_{11}$.

It is also possible that $P_{11} > p^*$. Then

$$J(P_{11}) = P_{11}R + \alpha P_{11} J(P_{11}) + \alpha(1 - P_{11}) J(P_{21}),$$

which gives

$$J(P_{11}) = \frac{P_{11}R + \alpha(1 - P_{11})\frac{\lambda}{1-\alpha}}{1 - \alpha P_{11}}.$$

Using the continuity of J at p^* and the fact that $fp^* \leq p^*$, we get

$$p^* \left[R + \alpha \frac{P_{11}R + \alpha(1 - P_{11})\frac{\lambda}{1-\alpha}}{1 - \alpha P_{11}} - \frac{\alpha\lambda}{1-\alpha} \right] = \lambda,$$

which after simplifications gives

$$p^*(\lambda) = \frac{(1 - \alpha P_{11})\lambda}{R - \alpha\lambda}. \tag{13}$$

Then the condition $p^* < P_{11}$ translates to $\frac{\lambda}{R} < P_{11}$, which is coherent. As before, it is easy to see that p^* is an increasing function of λ in this subcase. This completes the case $p^* \geq I$.

Remark. The expressions (12) and (13) give a continuous and monotonically increasing function $p^*(\lambda)$ on the interval $[\lambda_I, R] = [\lambda_I, P_{11}R] \cup [P_{11}R, R]$, where λ_I is the point where $p^*(\lambda_I) = I$, with $p^*(\lambda_I)$ given by (13). This can be rewritten

$$\lambda_I := \frac{P_{21}R}{1 - (P_{11} - P_{21})(1 + \alpha - \alpha P_{11})}.$$

Summarizing the results above, we have then

1. If $R > \lambda \geq P_{11}R$, then $p^*(\lambda) = \frac{\lambda}{R}$ and

$$J(p) = \begin{cases} \frac{\lambda}{1-\alpha} & \text{if } p \leq p^* \\ pR + \frac{\alpha\lambda}{1-\alpha} & \text{if } p > p^*. \end{cases}$$

2. If $P_{11}R > \lambda \geq \lambda_I$, then $p^*(\lambda) = \frac{(1-\alpha P_{11})\lambda}{R-\alpha\lambda}$ and

$$J(p) = \begin{cases} \frac{\lambda}{1-\alpha} & \text{if } p \leq p^* \\ \frac{\alpha\lambda}{1-\alpha} + p\,\frac{R-\alpha\lambda}{1-\alpha P_{11}} & \text{if } p > p^*. \end{cases}$$

Case $p^* < I$:

This subcase is the most involved. We have by continuity of J at p^*:

$$J(p^*) = \lambda + \alpha J(fp^*) = p^*R + \alpha p^* J(P_{11}) + \alpha(1-p^*)J(P_{21}).$$

Since $p^* < I$ we have $fp^* > p^*$, i.e., fp^* is in the active region. So we can rewrite the second equality as:

$$\lambda + \alpha((fp^*)R + \alpha(fp^*)J(P_{11}) + \alpha(1-fp^*)J(P_{21})) = p^*R + \alpha p^* J(P_{11}) + \alpha(1-p^*)J(P_{21}). \quad (14)$$

Expanding the left hand side gives

$$J(p^*) = \lambda + \alpha^2 J(P_{21}) + \alpha(P_{21} + p^*(P_{11} - P_{21}))[R + \alpha(J(P_{11}) - J(P_{21}))].$$

It is clear that $P_{11} \geq I$ and since $I > p^*$ we have

$$J(P_{11}) = P_{11}R + \alpha P_{11} J(P_{11}) + \alpha(1-P_{11})J(P_{21}),$$

so

$$J(P_{11}) = \frac{P_{11}R + \alpha(1-P_{11})J(P_{21})}{1-\alpha P_{11}},$$

which gives

$$J(P_{11}) - J(P_{21}) = \frac{P_{11}R - (1-\alpha)J(P_{21})}{1-\alpha P_{11}},$$

and

$$R + \alpha(J(P_{11}) - J(P_{21})) = \frac{R - \alpha(1-\alpha)J(P_{21})}{1-\alpha P_{11}}. \quad (15)$$

We now use (15) on both sides of the continuity condition (14):

$$\lambda + \alpha^2 J(P_{21}) + \alpha(P_{21} + p^*(P_{11} - P_{21}))\left[\frac{R - \alpha(1-\alpha)J(P_{21})}{1-\alpha P_{11}}\right] = \alpha J(P_{21}) + p^*\left[\frac{R - \alpha(1-\alpha)J(P_{21})}{1-\alpha P_{11}}\right].$$

We obtain:

$$p^* = \frac{\lambda - \alpha(1-\alpha)J(P_{21}) + \alpha P_{21}\left[\frac{R-\alpha(1-\alpha)J(P_{21})}{1-\alpha P_{11}}\right]}{(1-\alpha(P_{11}-P_{21}))\left[\frac{R-\alpha(1-\alpha)J(P_{21})}{1-\alpha P_{11}}\right]}.$$

We can rewrite this expression as

$$p^* = \frac{(1-\alpha(P_{11}-P_{21}))\left[\lambda - \alpha(1-\alpha)J(P_{21})\right] + \alpha P_{21}(R-M)}{(1-\alpha(P_{11}-P_{21}))\left[R - \alpha(1-\alpha)J(P_{21})\right]}$$

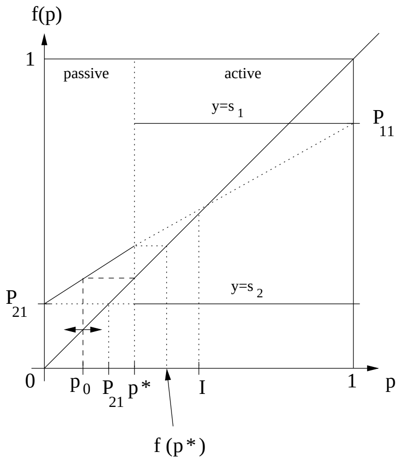

Figure 3: Case $P_{21} < p^* < I$.

or

$$p^* = 1 - \frac{(R-\lambda)(1-\alpha P_{11})}{(1-\alpha(P_{11}-P_{21}))\,[R-\alpha(1-\alpha)J(P_{21})]}. \tag{16}$$

So essentially the problem is solved if we can have an expression for $J(P_{21})$. The case where $p^* \leq P_{21}$ can be solved the most easily. Then

$$J(P_{21}) = P_{21}R + \alpha P_{21}J(P_{11}) + \alpha(1-P_{21})J(P_{21}),$$

which, combined with the similar equation for $J(P_{11})$, gives:

$$J(P_{11}) = \frac{R(P_{11}-\alpha(P_{11}-P_{21}))}{(1-\alpha)[1-\alpha(P_{11}-P_{21})]} \tag{17}$$

$$J(P_{21}) = \frac{P_{21}R}{(1-\alpha)[1-\alpha(P_{11}-P_{21})]}. \tag{18}$$

After some calculation, we get:

$$p^*(\lambda) = \frac{\lambda}{R}.$$

and this is valid for the case $p^*(\lambda) \leq P_{21}$, i.e., $\lambda \leq P_{21}R$.

Now from figure 3, if $P_{21} < p^* < I$, it can be seen that the iterates $f^k P_{21}$, initially in the passive region, eventually reach the active region and at that point one can evaluate $J(P_{21})$. The number of iterations however depends on the position of p^*.

It is easy to see that

$$f^n P_{21} = P_{21}\frac{1-(P_{11}-P_{21})^{n+1}}{1-(P_{11}-P_{21})}.$$

The rest of the analysis, computing $J(P_{21})$ for $P_{21} < p^* < I$, will distinguish different cases by the unique integer k such that:

$$f^k P_{21} < p^* \leq f^{k+1} P_{21}.$$

By definition of p^* separating the passive and active regions, we have:

$$J(P_{21}) = \lambda + \alpha\lambda + \alpha^2\lambda + \ldots + \alpha^k\lambda + \alpha^{k+1}J(f^{k+1}P_{21}) = \frac{1-\alpha^{k+1}}{1-\alpha}\lambda + \alpha^{k+1}J(f^{k+1}P_{21}), \tag{19}$$

$$\begin{aligned} J(f^{k+1}P_{21}) &= (f^{k+1}P_{21})R + \alpha(f^{k+1}P_{21})J(P_{11}) + \alpha(1-f^{k+1}P_{21})J(P_{21}) \\ &= \alpha J(P_{21}) + (f^{k+1}P_{21})[R + \alpha(J(P_{11}) - J(P_{21}))] \\ &= \alpha J(P_{21}) + (f^{k+1}P_{21})\frac{R - \alpha(1-\alpha)J(P_{21})}{1-\alpha P_{11}}, \end{aligned} \tag{20}$$

where the last line was obtained using (15). Solving this system of equations, we obtain

$$J(P_{21}) = \frac{1}{1-\alpha}\,\frac{(1-\alpha^{k+1})(1-\alpha P_{11})\lambda + \alpha^{k+1}(1-\alpha)(f^{k+1}P_{21})R}{(1-\alpha^{k+2})(1-\alpha P_{11}) + \alpha^{k+2}(1-\alpha)(f^{k+1}P_{21})}. \tag{21}$$

We can now use this expression in (16). As an intermediate step, we compute:

$$R - \alpha(1-\alpha)J(P_{21}) = \frac{(1-\alpha P_{11})[R(1-\alpha^{k+2}) - \lambda(\alpha - \alpha^{k+2})]}{(1-\alpha P_{11})(1-\alpha^{k+2}) + \alpha^{k+2}(1-\alpha)(f^{k+1}P_{21})}.$$

Then we obtain:

$$p^* = 1 - \frac{(R-\lambda)[(1-\alpha P_{11})(1-\alpha^{k+2}) + \alpha^{k+2}(1-\alpha)(f^{k+1}P_{21})]}{(1-\alpha(P_{11}-P_{21}))[R(1-\alpha^{k+2}) - \lambda(\alpha-\alpha^{k+2})]}.$$

We rewrite this expression in a more readable form as:

$$\begin{aligned} p^* &= 1 - A_k\frac{R-\lambda}{B_kR - C_k\lambda} \\ A_k &= \frac{(1-\alpha P_{11})(1-\alpha^{k+2}) + \alpha^{k+2}(1-\alpha)(f^{k+1}P_{21})}{1-\alpha(P_{11}-P_{21})} \\ B_k &= 1-\alpha^{k+2} \\ C_k &= \alpha - \alpha^{k+2}. \end{aligned} \tag{22}$$

The condition

$$f^kP_{21} < p^* \leq f^{k+1}P_{21}$$

can then be rewritten as

$$\frac{A_k - (1-f^kP_{21})B_k}{A_k - (1-f^kP_{21})C_k} < \frac{\lambda}{R} \leq \frac{A_k - (1-f^{k+1}P_{21})B_k}{A_k - (1-f^{k+1}P_{21})C_k}. \tag{23}$$

As a sanity check, we can verify that the successive thresholds agree:

$$\frac{A_k - (1-f^{k+1}P_{21})B_k}{A_k - (1-f^{k+1}P_{21})C_k} = \frac{A_{k+1} - (1-f^{k+1}P_{21})B_{k+1}}{A_{k+1} - (1-f^{k+1}P_{21})C_{k+1}}.$$

Taking the cross-products, this amounts to verifying:

$$(1-f^{k+1}P_{21})(B_{k+1}C_k - B_kC_{k+1}) + A_k(C_{k+1} - B_{k+1}) + A_{k+1}(B_k - C_k) = 0.$$

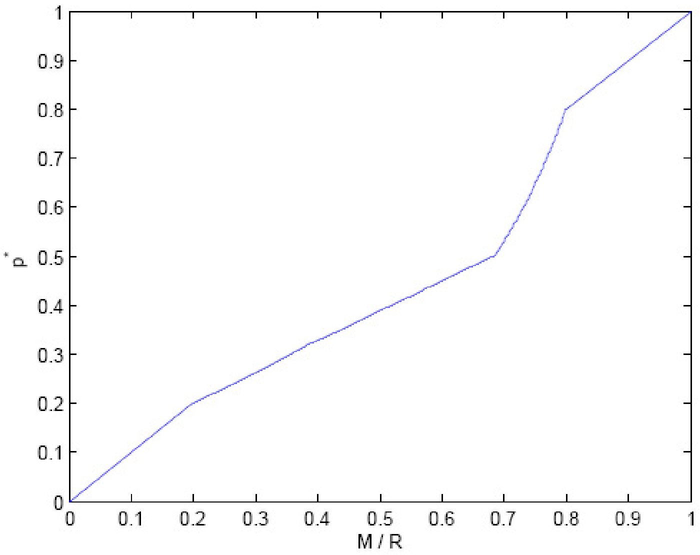

Figure 4: Plot of $p^*(\lambda)$ for $P_{21} = 0.2, P_{11} = 0.8, \alpha = 0.9$.

The left hand side gives

$$\begin{aligned}
&(1 - f^{k+1}P_{21})[\alpha(1-\alpha^{k+3})((1-\alpha^{k+1}) - \alpha(1-\alpha^{k+2})^2] + (1-\alpha)(A_{k+1} - A_k) \\
=& -\alpha^{k+2}(1-\alpha)^2(1 - f^{k+1}P_{21}) + (1-\alpha)\frac{(1-\alpha P_{11})\alpha^{k+2}(1-\alpha) + \alpha^{k+2}(1-\alpha)(\alpha f^{k+2}P_{21} - f^{k+1}P_{21})}{1 - \alpha(P_{11} - P_{21})} \\
=&\alpha^{k+2}(1-\alpha)^2\left[-(1 - f^{k+1}P_{21}) + \frac{(1-\alpha P_{11}) + \alpha P_{21} + [\alpha(P_{11} - P_{21}) - 1]f^{k+1}P_{21}}{1 - \alpha(P_{11} - P_{21})}\right] \\
=&0.
\end{aligned}$$

On each interval, we verify that p^* is an increasing function of λ. We just compute

$$\begin{aligned}
\frac{dp^*}{d\lambda} &= -A_k\frac{-(B_kR - C_k\lambda) + C_k(R-\lambda)}{(B_kR - C_k\lambda)^2} \\
\frac{dp^*}{d\lambda} &= A_k\frac{(1-\alpha)R}{(B_kR - C_k\lambda)^2} \geq 0.
\end{aligned}$$

Remark. It can be verified that the first threshold also coincides with the previous case studied, that is:

$$\frac{A_0 - (1 - fP_{21})B_0}{A_0 - (1 - fP_{21})C_0} = P_{21}$$

Also, as $k \to \infty$, we can easily verify that the thresholds in (23) converge to λ_I/R.

A plot of $p^*(\lambda)$ is given on Fig. 4 and 5. To conclude this case, i.e., $\lambda < \lambda_I$, let us summarize the computational procedure.

1. If $\lambda \leq P_{21}R$, then $p = \frac{\lambda}{R}$. $J(P_{11})$ and $J(P_{21})$ are given by (17) and (18) respectively.

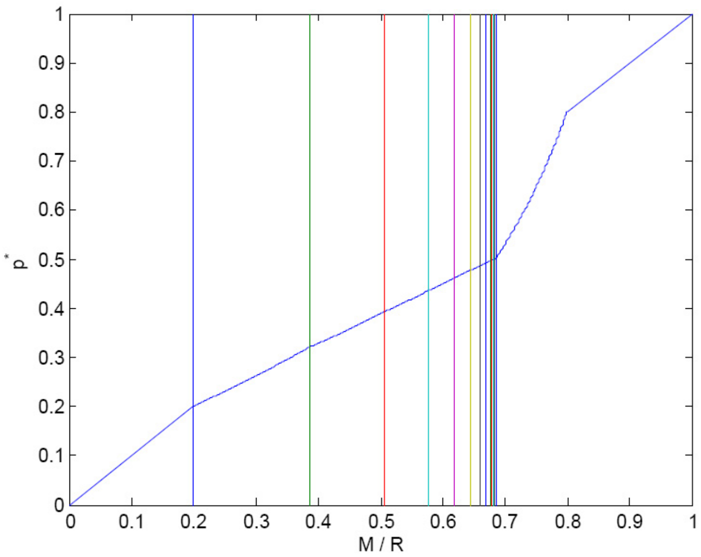

Figure 5: Plot of $p^*(\lambda)$ for $P_{21} = 0.2, P_{11} = 0.8, \alpha = 0.9$. The vertical lines show the separation between the regions corresponding to different values of k in the analysis for $p^* < I$. λ_I is an accumulation point, i.e., there are in fact infinitely many such lines converging to λ_I.

2. If $P_{21}R < \lambda < \lambda_I$, we have first to find the unique k such that the condition (23) is verified. Once this is done, p^* is given by (22), $J(P_{21})$ is given by (21), and $J(P_{11})$, or equivalently $R + \alpha(J(P_{11}) - J(P_{21}))$, is given by (15).

With this, we have all the elements to actually compute $J(p)$ for given values of p and λ. When $p \geq p^*$, we have $J(p) = \alpha J(P_{21}) + p[R + \alpha(J(P_{11}) - J(P_{21}))]$ and so we are done. When $p < p^*$ however, to finish the computation we need to proceed as in the computation of $J(P_{21})$. We first find the unique integer l such that $f^l p < p^* \leq f^{l+1}p$ (it exists since here $p^* < I$ and $f^l p \to I$ as $l \to \infty$). Let $s = P_{11} - P_{21}$. Since

$$f^l p = P_{21}\frac{1-s^l}{1-s} + ps^l = I(1-s^l) + ps^l = I - s^l(I-p),$$

we obtain

$$l = \left\lceil \frac{1}{\ln s} \ln \frac{I-p^*}{I-p} \right\rceil - 1.$$

Then we have

$$J(p) = \frac{1-\alpha^{l+1}}{1-\alpha}\lambda + \alpha^{l+2}J(P_{21}) + \alpha^{l+1}(f^{l+1}p)[R + \alpha(J(P_{11}) - J(P_{21}))],$$

see (19), (20).

5.4 Case $P_{21} > P_{11}$

5.4.1 Case $P_{11} = 0,\ P_{21} = 1$

As in section 5.3, we start the study of the remaining case $P_{21} > P_{11}$ with $P_{11} = 0,\ P_{21} = 1$. Then, because the active and the passive actions are necessarily optimal at $p = 1$ and $p = 0$ respectively by lemma 5.4, we have

$$J(1) = R + \alpha J(0), \quad J(0) = \lambda + \alpha J(1),$$

which gives

$$J(1) = \frac{R + \alpha\lambda}{1 - \alpha^2}, \quad J(0) = \frac{\lambda + \alpha R}{1 - \alpha^2}, \tag{24}$$

and from which we get

$$R + \alpha(J(0) - J(1)) = \frac{R + \alpha\lambda}{1 + \alpha}.$$

Now by continuity of J,

$$J(p^*) = \lambda + \alpha J(1 - p^*) = p^* R + \alpha p^* J(0) + \alpha(1 - p^*)J(1), \tag{25}$$

and using the preceding relations in the right-hand side

$$J(p^*) = \alpha\frac{R + \alpha\lambda}{1 - \alpha^2} + p^*\frac{R + \alpha\lambda}{1 + \alpha} = \frac{R + \alpha\lambda}{1 + \alpha}\left(\frac{\alpha}{1 - \alpha} + p^*\right). \tag{26}$$

Now if we have $p^* \geq 1/2$, then $1 - p^* \leq 1/2 \leq p^*$ is in the passive region, and so

$$J(1 - p^*) = \lambda + \alpha J(p^*).$$

Reporting in the first part of (25), we obtain

$$J(p^*) = \frac{\lambda}{1 - \alpha},$$

which, with (26), gives

$$p^* = \frac{\lambda(1 + \alpha - \alpha^2) - \alpha R}{(1 - \alpha)(R + \alpha\lambda)}.$$

p^* is an increasing function of λ since one can verify that

$$\frac{dp^*}{d\lambda} = (1 - \alpha^2)\frac{R}{(1 - \alpha)^2(R + \alpha\lambda)^2} \geq 0.$$

Then the condition $p^* \geq 1/2$ translates to

$$\frac{\lambda}{R} \geq \frac{1 + \alpha}{2 + \alpha - \alpha^2} = \frac{1}{2 - \alpha}.$$

In the case $p^* < 1/2$, $(1 - p^*) > 1/2 > p^*$ is in the active region, so we obtain

$$\begin{aligned} J(1 - p^*) &= (1 - p^*)R + \alpha(1 - p^*)J(0) + \alpha p^* J(1) \\ &= R + \alpha J(0) - p^*[R + \alpha(J(0) - J(1))] \end{aligned}$$

Reporting in (25), we get

$$\lambda + \alpha(R + \alpha J(0) - J(1)) = p^*(1+\alpha)[R + \alpha(J(0) - J(1))]$$

which gives after easy calculation

$$p^* = \frac{\lambda}{R + \alpha\lambda}.$$

This is again an increasing function of λ and the condition $p^* < 1/2$ can be written

$$\frac{\lambda}{R} < \frac{1}{2-\alpha},$$

which is coherent (the junction between the two cases happens when $p^*(\lambda)$ hits $1/2$).

Now for a given value of p and λ, we also want to compute $J(p)$. Comparing λ/R to $1/(2-\alpha)$, we can deduce the correct formula for p^*. Then if $p \geq p^*$, we are done, using the values of $J(0)$ and $J(1)$ in (24). We obtain in this case:

$$J(p) = \frac{R + \alpha\lambda}{1-\alpha^2}(\alpha + p(1-\alpha)).$$

If $p < p^*$, $J(p) = \lambda + \alpha J(1-p)$, and we distinguish between two subcases. If $(1-p) \leq p^*$ (which is only possible if $p^* > 1/2$), i.e., $1 - p^* \leq p < p^*$, then

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

Otherwise, if $(1-p) > p^*$, i.e., $p < 1 - p^*$ and $p < p^*$, then

$$J(p) = \lambda + \alpha[(1-p)R + \alpha(1-p)J(0) + \alpha p J(1)].$$

and this gives after simplifications

$$J(p) = \frac{\lambda(1 - \alpha^2 p(1-\alpha)) + \alpha R(1 - p(1-\alpha))}{1-\alpha^2}.$$

5.4.2 Case $0 < P_{21} - P_{11} < 1$

The discussion related to the fixed point I at the beginning of section 5.3.2 is still valid here, including equation (11) showing the convergence of the iterates $f^n p$ to I since we assume in this paragraph that $0 < P_{21} - P_{11} < 1$. These iterations land now alternatively on each side of I.

Case $p^* \geq I$.

In the case $p^* \geq I$, the iterates $f^n p^*$ converge to I while remaining in the passive region. So we immediately get

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

By continuity of J at p^*, where the active action is also optimal, we have

$$\frac{\lambda}{1-\alpha} = p^* R + \alpha p^* J(P_{11}) + \alpha(1-p^*)J(P_{21}),$$

and this gives us

$$p^* = \frac{\frac{\lambda}{1-\alpha} - \alpha J(P_{21})}{R + \alpha(J(P_{11}) - J(P_{21}))}. \tag{27}$$

We now compute $J(P_{11})$ and $J(P_{21})$ in the different cases, depending on the position of p^*. Note first that $P_{11} < I$ necessarily, so P_{11} is in the passive region by our assumption $p^* \geq I$. Hence

$$J(P_{11}) = \lambda + \alpha J(fP_{11}).$$

fP_{11} is greater than I however and can fall in the passive or the active region. If $fP_{11} \leq p^*$, it is easy to see that the iterates $f^n P_{11}$ will remain in the passive region, and so we obtain

$$J(P_{11}) = \frac{\lambda}{1-\alpha}.$$

If $fP_{11} > p^*$, we get

$$J(fP_{11}) = (fP_{11})R + \alpha(fP_{11})J(P_{11}) + \alpha(1 - fP_{11})J(P_{21}).$$

We can now have p^* greater or smaller than P_{21}. However note that necessarily $I < P_{21}$ and so if $P_{21} \leq p^*$, the iterates $f^n P_{21}$ remain in the passive region and

$$J(P_{21}) = \frac{\lambda}{1-\alpha}.$$

Otherwise if $p^* < P_{21}$, then

$$J(P_{21}) = P_{21}R + \alpha P_{21} J(P_{11}) + \alpha(1 - P_{21})J(P_{21}),$$

and so

$$\begin{aligned} J(P_{21}) &= P_{21} \frac{R + \alpha J(P_{11})}{1 - \alpha + \alpha P_{21}}, \\ R + \alpha(J(P_{11}) - J(P_{21})) &= (1-\alpha)\frac{R + \alpha J(P_{11})}{1 - \alpha + \alpha P_{21}}. \end{aligned}$$

We can now finish the computation of p^* using (27). Note first that

$$fP_{11} = P_{21} - P_{11}(P_{21} - P_{11}) < P_{21}.$$

Hence we will subdivide the interval $[I, 1]$ into the union $[I, fP_{11}] \cup [fP_{11}, P_{21}] \cup [P_{21}, 1]$. For $p^* \in [P_{21}, 1]$, $J(P_{11}) = J(P_{21}) = \lambda/(1-\alpha)$ and so

$$p^* = \frac{\lambda}{R}.$$

Clearly then $p^*(\lambda) = P_{21}$ if and only if $\lambda = P_{21}R$.

For $p^* \in [fP_{11}, P_{21}]$,

$$p^* = \frac{\frac{\lambda}{1-\alpha} - \alpha P_{21} \frac{R+\alpha J(P_{11})}{1-\alpha+\alpha P_{21}}}{(1-\alpha)\frac{R+\alpha J(P_{11})}{1-\alpha+\alpha P_{21}}}, \quad J(P_{11}) = \frac{\lambda}{1-\alpha}.$$

This gives after substitution

$$p^* = \frac{\lambda(1 + \alpha P_{21}) - \alpha P_{21} R}{R(1-\alpha) + \alpha\lambda}.$$

We verify easily that $\frac{dp^*}{d\lambda}$ is an increasing function of λ and we check that this expression gives again $p^*(\lambda) = P_{21}$ if and only if $\lambda = P_{21}R$. As for the other side of the interval, we have $p^*(\lambda) = fP_{11}$ if and only if

$$\lambda = \lambda_{fP_{11}} := R\frac{P_{21} - (1-\alpha)P_{11}(P_{21} - P_{11})}{1 + \alpha P_{11}(P_{21} - P_{11})}. \tag{28}$$

Finally we consider the case $p^* \in [I, fP_{11}]$. There is a bit more work to get an expression for $J(P_{11})$. We have

$$\begin{aligned}
J(P_{11}) &= \lambda + \alpha J(fP_{11}) \\
J(P_{11}) &= \lambda + \alpha^2 J(P_{21}) + \alpha(fP_{11})[R + \alpha(J(P_{11}) - J(P_{21}))] \\
J(P_{11}) &= \lambda + \frac{R + \alpha J(P_{11})}{1 - \alpha + \alpha P_{21}}[\alpha^2 P_{21} + \alpha(1-\alpha)(fP_{11})].
\end{aligned}$$

This implies immediately

$$\begin{aligned}
R + \alpha J(P_{11}) &= (R + \alpha\lambda) + \frac{R + \alpha J(P_{11})}{1 - \alpha + \alpha P_{21}}[\alpha^3 P_{21} + \alpha^2(1-\alpha)(fP_{11})] \\
R + \alpha J(P_{11}) &= \frac{(R + \alpha\lambda)(1 - \alpha + \alpha P_{21})}{(1-\alpha)(1 + \alpha P_{21} + \alpha^2 P_{11}(P_{21} - P_{11}))}.
\end{aligned}$$

Finally this gives

$$\begin{aligned}
p^* &= \frac{\lambda(1 + \alpha(1+\alpha)P_{21} - \alpha^2(fP_{11})) - \alpha P_{21}(R + \alpha\lambda)}{(1-\alpha)(R + \alpha\lambda)} \\
p^* &= \frac{\lambda(1 + \alpha(1-\alpha)P_{21} + \alpha^2 P_{11}(P_{21} - P_{11})) - \alpha P_{21}R}{(1-\alpha)(R + \alpha\lambda)}.
\end{aligned}$$

It is again straightforward to verify that this is an increasing function of λ by computing the derivative. For the boundary points, we get by direct calculations that $p^*(\lambda) = fP_{11}$ if and only if λ is given by (28), verifying the continuity at this point, and $p^*(\lambda) = I$ if and only if

$$\lambda = \lambda_I := \frac{P_{21}R}{1 + (P_{21} - P_{11})(1 - \alpha + \alpha P_{11})}. \tag{29}$$

This expression will also be obtained from the analysis below for $P_{11} < p^* < I$.

Case $p^* < I$.

Last, we consider the case $p^* < I$. It is clear graphically or by inspection of the expression for I that $P_{21} > I$, so the active action is optimal at P_{21}, and

$$J(P_{21}) = P_{21}\frac{R + \alpha J(P_{11})}{1 - \alpha(1 - P_{21})}.$$

This gives

$$R + \alpha(J(P_{11}) - J(P_{21})) = (1-\alpha)\frac{R + \alpha J(P_{11})}{1 - \alpha(1 - P_{21})}.$$

We have, again by equation (11), that $fp^* > I > p^*$, so, by continuity of J at p^*,

$$\lambda + \alpha J(fp^*) = \lambda + \alpha[(fp^*)R + \alpha(fp^*)J(P_{11}) + \alpha(1 - fp^*)J(P_{21})] = p^*R + \alpha p^* J(P_{11}) + \alpha(1-p^*)J(P_{21}),$$

identical to (14). We get

$$\begin{aligned}&\lambda(1-\alpha+\alpha P_{21})+(R+\alpha J(P_{11}))(\alpha^2 P_{21}+\alpha(1-\alpha)(P_{21}+p^*(P_{11}-P_{21})))\\&=(R+\alpha J(P_{11}))(\alpha P_{21}+(1-\alpha)p^*),\end{aligned}$$

i.e.,

$$\frac{\lambda(1-\alpha+\alpha P_{21})}{R+\alpha J(P_{11})}+\alpha(1-\alpha)p^*(P_{11}-P_{21})=(1-\alpha)p^*,$$

so

$$p^*=\frac{(1+\alpha P_{21}-\alpha)\lambda}{(1-\alpha)(1+\alpha P_{21}-\alpha P_{11})(R+\alpha J(P_{11}))}.$$

Hence the problem is solved if we have an expression for $J(P_{11})$. It is clear graphically that $P_{11}<I$. From equation (11), we also see that $fP_{11}>I$ since $P_{11}-P_{21}<0$. Hence this time, fP_{11} is in the active region, and so the analysis is simpler than in paragraph 5.3.2. The only cases to consider are $0\le p^*\le P_{11}$ and $P_{11}<p^*<I$.

In the first subcase $0\le p^*\le P_{11}$, P_{11} is in the active region and so

$$\begin{aligned}J(P_{11})&=P_{11}R+\alpha P_{11}J(P_{11})+\alpha(1-P_{11})J(P_{21})\\J(P_{11})&=\alpha\frac{P_{21}R+\alpha P_{21}J(P_{11})}{1-\alpha(1-P_{21})}+P_{11}(1-\alpha)\frac{R+\alpha J(P_{11})}{1-\alpha(1-P_{21})}\\J(P_{11})&=\frac{(P_{11}+\alpha(P_{21}-P_{11}))(R+\alpha J(P_{11}))}{1-\alpha(1-P_{21})},\end{aligned}$$

and so

$$\begin{aligned}J(P_{11})&=\frac{R(P_{11}+\alpha(P_{21}-P_{11}))}{(1-\alpha)(1+\alpha(P_{21}-P_{11}))},\\R+\alpha J(P_{11})&=\frac{R(1-\alpha+\alpha P_{21})}{(1-\alpha)(1+\alpha(P_{21}-P_{11}))}.\end{aligned}$$

This gives

$$p^*=\frac{\lambda}{R},$$

and the condition $p^*\le P_{11}$ is $\lambda\le P_{11}R$.

In the second subcase, $P_{11}<p^*<I$, and fP_{11} is in the active region, so we have

$$\begin{aligned}J(P_{11})&=\lambda+\alpha J(fP_{11})=\lambda+\alpha[fP_{11}R+\alpha fP_{11}J(P_{11})+\alpha(1-fP_{11})J(P_{21})]\\J(P_{11})&=\lambda+\frac{R+\alpha J(P_{11})}{1-\alpha(1-P_{21})}(\alpha^2P_{21}+\alpha(1-\alpha)fP_{11})\end{aligned}$$

$$(R+\alpha J(P_{11}))(1-\alpha+\alpha P_{21})=(R+\alpha\lambda)(1-\alpha+\alpha P_{21})+\alpha^2(\alpha P_{21}+(1-\alpha)fP_{11})(R+\alpha J(P_{11}))$$

and so we get

$$R+\alpha J(P_{11})=\frac{(R+\alpha\lambda)(1-\alpha+\alpha P_{21})}{(1-\alpha)[1+\alpha P_{21}+\alpha^2P_{11}(P_{21}-P_{11})]}.$$

This gives finally

$$p^*=\frac{\lambda[1+\alpha P_{21}+\alpha^2P_{11}(P_{21}-P_{11})]}{(R+\alpha\lambda)(1+\alpha P_{21}-\alpha P_{11})},$$

which simplifies to (add and substract αP_{11} in the numerator)

$$p^*=\frac{\lambda(1+\alpha P_{11})}{R+\alpha\lambda}.$$

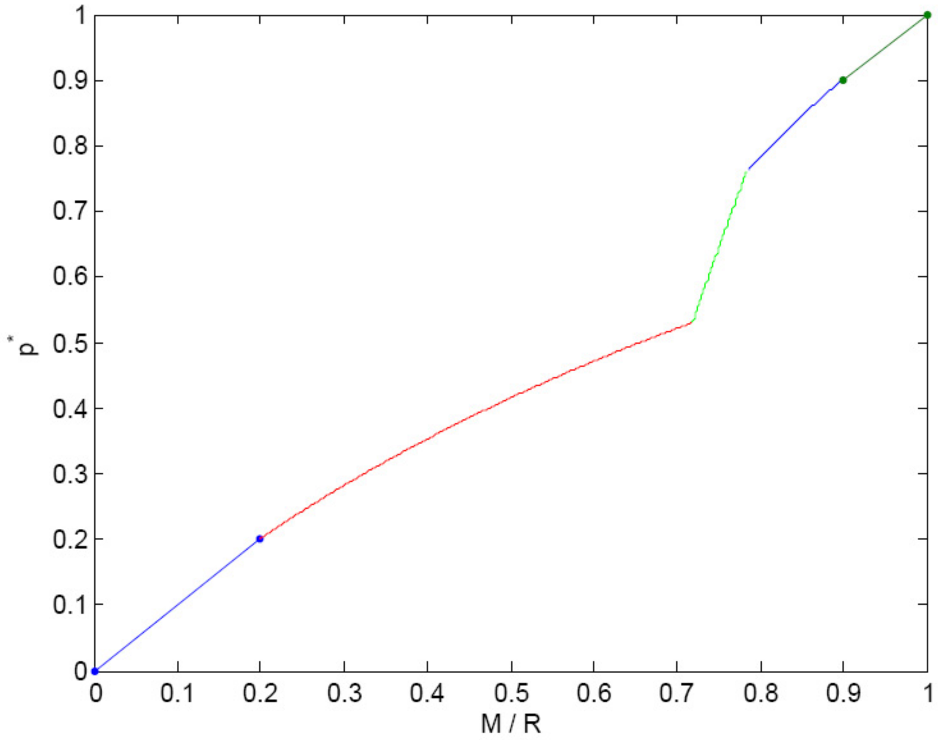

Figure 6: Plot of $p^*(\lambda)$ for $P_{21} = 0.9, P_{11} = 0.2, \alpha = 0.9$.

It is easy to see that it is an increasing function of λ, and that the condition $p^* \geq P_{11}$ translates to $\lambda \geq P_{11}R$. We also verify that the condition $p^* < I$ corresponds to $\lambda < \lambda_I$, where λ_I is given by (29), which implies the continuity of $p^*(\lambda)$ at λ_I.

A plot of $p^*(\lambda)$ is given on Fig. 6. Finally, we summarize the computational procedure for the case $0 < P_{21} - P_{11} < 1$. Given λ, we first check in which subset of the partition $[0, P_{11}R] \cup [P_{11}R, \lambda_I] \cup [\lambda_I, \lambda_{fP_{11}}] \cup [\lambda_{fP_{11}}, P_{21}R] \cup [P_{21}R, R]$ it belongs. We then compute $p^*(\lambda)$ accordingly. That is,

$$p^*(\lambda) = \begin{cases} \frac{\lambda}{R} \;\text{ if } \lambda \in [0, P_{11}R] \cup [P_{21}R, R] \\ \frac{\lambda(1+\alpha P_{11})}{R+\alpha\lambda} \;\text{ if } \lambda \in [P_{11}R, \lambda_I] \\ \frac{\lambda(1+\alpha(1-\alpha)P_{21}+\alpha^2 P_{11}(P_{21}-P_{11}))-\alpha P_{21}R}{(1-\alpha)(R+\alpha\lambda)} \;\text{ if } \lambda \in [\lambda_I, \lambda_{fP_{11}}] \\ \frac{\lambda(1+\alpha P_{21})-\alpha P_{21}R}{R(1-\alpha)+\alpha\lambda} \;\text{ if } \lambda \in [\lambda_{fP_{11}}, P_{21}R]. \end{cases}$$

With the value of $p^*(\lambda)$, we can finish the computation. For a given $p \in [0, 1]$, we can first have $p > p^*$, in which case we are done, using the values of $J(P_{11})$ and $J(P_{21})$ computed in the various cases. If $p < p^*$, as in the previous paragraph, we need to distinguish two subcases. If $fp \leq p^*$ (which can happen only when $p* > I$, i.e., $\lambda > \lambda_I$), then immediately $J(P) = \lambda/(1-\alpha)$ since the iterations remain in the passive region. Otherwise, $fp > p^*$ and we can compute

$$J(p) = \lambda + \alpha(fp)R + \alpha^2(fp)J(P_{11}) + \alpha^2(1 - fp)J(P_{21}),$$

using the values of $J(P_{11})$ and $J(P_{11})$ obtained in the various cases.

5.5 Expression of the Indices

The previous paragraphs establish the indexability property for the two-state Markov chain with the described information structure, for all possible values of the state-transition matrix. Now we obtain Whittle's index by inverting the relation $p^*(\lambda)$ to $\lambda(p)$. We get

Theorem 5.5. *A two-state restless bandit with null/perfect-observations 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 $s = 0$ ($P_{11} = P_{21}$):*
$$\lambda(p) = pR$$
2. *Case $s = 1$ ($P_{11} = 1, P_{21} = 0$):*
$$\lambda(p) = \frac{pR}{1 - \alpha(1-p)}.$$
3. *Case $0 < s < 1$ ($P_{11} > P_{21}, P_{11} - P_{21} < 1$ and note that $P_{21} < I \leq P_{11}$):*
    - *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 $k(p) = \lceil \frac{\ln(1-\frac{p}{I})}{\ln s} \rceil - 2$ (i.e., $k(p)$ is the unique integer such that $\frac{\ln(1-\frac{p}{I})}{\ln s} - 2 \leq k(p) < \frac{\ln(1-\frac{p}{I})}{\ln s} - 1$). Then let*
    $$\begin{aligned} A_{k(p)} &= \frac{(1-\alpha P_{11})(1-\alpha^{k(p)+2}) + \alpha^{k(p)+2}(1-\alpha)(f^{k(p)+1}P_{21})}{1-\alpha s} \\ B_{k(p)} &= 1 - \alpha^{k(p)+2} \\ C_{k(p)} &= \alpha - \alpha^{k(p)+2}. \end{aligned}$$
    We have
    $$\lambda(p) = \frac{A_{k(p)} - (1-p)B_{k(p)}}{A_{k(p)} - (1-p)C_{k(p)}} R.$$
4. *Case $s = -1$ ($P_{11} = 0, P_{21} = 1$):*
    - *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 $-1 < s < 0$ ($P_{11} < P_{21}, P_{21} - P_{11} < 1$ and note that $P_{21} > I > P_{11}$):*
    - *If $p \geq P_{21}$ or $p \leq P_{11}$: $\lambda(p) = pR$.*
    - *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 $P_{11} < p < I$: $\lambda(p) = \frac{p}{1-\alpha(p-P_{11})} R$.*

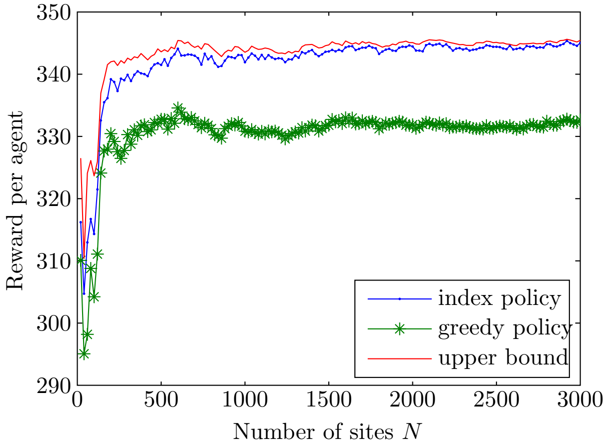

Figure 7: Monte-Carlo Simulation for Whittle's index policy and the greedy policy. The upper bound is computed using the subgradient optimization algorithm. We fixed $\alpha = 0.95$.

6 Computational Experiments

In this section, we 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. 7 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 consistantly stronger performance of the index policy with respect to the simple greedy policy, and in fact its asymptotic quasi-optimality.

7 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 bandits 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

[Alt99] E. Altman. *Constrained Markov Decision Processes.* Chapman and Hall, 1999.

[Ath72] M. Athans. On the determination of optimal costly measurement strategies. *Automatica*, 8:397–412, 1972.

[Ber99] D.P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.

[Ber01] D.P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1 and 2. Athena Scientific, 2 edition, 2001.

[BTAH02] J.S. Bellingham, M. Tillerson, M. Alighanbari, and J.P. How. Cooperative path planning for multiple uavs in dynamic and uncertain environments. In *Proceedings of the 41st IEEE Conference on Decision and Control*, 2002.

[Cas97] D.A. Castañón. Approximate dynamic programming for sensor management. In *Proceedings of the 36th Conference on Decision and Control*, pages 1202–1207, December 1997.

[Cas05] D.A. Castañón. Stochastic control bounds on sensor network performance. In *Proceedings of the 44th IEEE Conference on Decision and Control*, 2005.

[FO90] E. Feron and C. Olivier. Targets, sensors and infinite-horizon tracking optimality. In *Proceedings of the 29th IEEE Conference on Decision and Control*, 1990.

[GCHR06] V. Gupta, T.H. Chung, B. Hassibi, and R.M.Murray. On a stochastic sensor selection algorithm with applications in sensor scheduling and sensor coverage. *Automatica*, 42(2):251–260, 2006.

[Git89] J. Gittins. *Multi-armed Bandit Allocation Indices.* Wiley-Interscience series in Systems and Optimization. John Wiley and sons, New York, 1989.

[GJ74] J.C. Gittins and D.M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani, editor, *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam, 1974.

[GRHK06] K.D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride. Some indexable families of restless bandit problems. *Advances in Applied Probability*, 38:643–672, 2006.

[KE01] V. Krishnamurthy and R.J. Evans. Hidden markov model multiarm bandits: a methodology for beam scheduling in multitarget tracking. *IEEE Transactions on Signal Processing*, 49(12):2893 – 2908, December 2001.

[KE03] V. Krishnamurthy and R.J. Evans. Correction to "hidden markov model multiarm bandits: a methodology for beam scheduling in multitarget tracking". *IEEE Transactions on Signal Processing*, 51(6):1662–1663, June 2003.

[LDF06] J. Le Ny, M. Dahleh, and E. Feron. Multi-agent task assignment in the bandit framework. In *Proccedings of the 45th IEEE Conference on Decision and Control*, San Diego, CA, December 2006.

[MPD67] L. Meier, J. Perschon, and R.M. Dressler. Optimal control of measurement systems. *IEEE Transactions on Automatic Control*, 12(5):528–536, 1967.

[NM01] J. Niño-Mora. Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33:76–98, 2001.

[PT99] C.H. Papadimitriou and J.N. Tsitsiklis. The complexity of optimal queueing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

[Son78] E.J. Sondik. The optimal control of partially observable markov decision processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2):282–304, March-April 1978.

[UAV05] Unmanned aircraft systems roadmap 2005-2030. Technical report, Office of the Secretary of Defense, 2005.

[Whi88] P. Whittle. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, 25A:287–298, 1988.

[Wil07] J.L. Williams. *Information Theoretic Sensor Management*. PhD thesis, Massachusetts Institute of Technology, February 2007.

[WW90] R.R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27:637–648, 1990.

[YW00] K.A. Yost and A.R. Washburn. The lp/pomdp marriage: Optimization with imperfect information. *Naval Research Logistics*, 47(8):607 – 619, 2000.