Restless Bandits for Dynamic Sourcing of Resources in Social and Information Networks

Varun Mehta, Rahul Meshram, Kesav Kaza and S. N. Merchant
Department of Electrical Engineering
IIT Bombay, Mumbai INDIA.

***Abstract*— We study a problem of information gathering in a social network with dynamically available sources and time varying quality of information. We formulate this problem as a restless multi-armed bandit (RMAB). In this problem, information quality of a source corresponds to the state of an arm in RMAB. The decision making agent does not know the quality of information from sources a priori. But the agent maintains a belief about the quality of information from each source. This is a problem of RMAB with partially observable states. The objective of the agent is to gather relevant information efficiently from sources by contacting them. We formulate this as a infinite horizon discounted reward problem, where reward depends on quality of information. We study Whittle index policy which determines the sequence of play of arms that maximizes long term cumulative reward. We illustrate the performance of index policy compare it myopic and uniform random policies through numerical examples.**

I. Introduction

Recent developments in the field of information and communication technology have made possible, a large number of "smart" hand-held devices and pervasive internet connectivity. These technologies lead to easier access and delivery of information to users who may use it for various applications. At a social level, this information and communication technology has increased the connectivity and interaction among people across the world through many vibrant media and networking platforms including Facebook, Twitter, Linkedin etc. serving different purposes.

These networks are increasingly being used by people as a source of information that help to make personal and professional decisions. In recent times there have been instances of large news agencies using social media and crowd sourcing to gather and disseminate information [1]. Individuals and organizations who source information from social networking platforms are faced with the problem of not only efficient sourcing, but also of accuracy. This may often be due to the dynamic nature of such sources and a limited capacity to "fact check", [2]–[4]. In this work we define and solve the problem of such an agent in a social or information network who tries gather information efficiently and maximize benefit from it.

Let us look at a motivating example, one from a social network scenario. Suppose there is an agent in a social or information network. The agent has connections to N neighbors which are its information sources. The agent needs information for its use and it gathers this information through its sources at regular intervals. In each interval, the agent can contact only $M < N$ of its sources for information. The information provided by a source may be either relevant (1) or non-relevant (0) to the agent. So, there are two states $\{0, 1\}$ corresponding to the information quality. However, the agent does not know a priori whether a certain source has relevant information. Relevance of the information received becomes apparent to the agent at end of the interval after processing it. The agent however knows that the information quality of a source varies in a Markovian manner and also knows its Markov matrix. The reward from relevant information is high and non-relevant information is low.

Further, in a given interval each source may or may not be available. However, an unavailable source may be leveraged through an additional cost. Hence the immediate reward that such information gives may be lower. This is a situation where the choice of sources might effect the their future availability and information quality of the sources. The agent here needs a policy to choose which sources it must contact in each interval along the time line so that its cumulative reward is maximized. The agent above and its neighbors can be imagined to be - an investigative journalist [1] and his sources, a newspaper editor and her staff reporters, a manufacturing business and its suppliers etc.

A dynamic information sourcing problem such as above is sequential decision problem where a current decision impacts future rewards. Such sequential decision problems are often modeled using Multi-armed bandits (MAB) (see [5]). In this work we model the dynamic sourcing problem as a restless multi-armed bandit (RMAB) problem.

A MAB is an agent with N arms and each arm can be in one of the finite state. The states of arms evolve along Markov chains whose transition probabilities are known to the agent. The agent can pull $M < N$ arms at a time. Reward of pulling an arm depends on its state. The agent's needs a policy to make these arm choices each slot to maximize its long term cumulative reward. In a restless multi-armed bandit (RMAB), state of each arm evolves in each time slot and that evolution can be action dependent, [6].

Our proposed RMAB model will be applicable not only to dynamic sourcing of resources problem but also to crowd-sourcing of computational resources in a communication network such as below. Consider a communication system with a coordinating entity (CE), a group of N user equipment

The work of Varun Mehta and Kesav Kaza was done in SPANN Lab at IIT Bombay. The work of Rahul Meshram was carried out in the Bharti Centre for Communications at IIT Bombay.

(UEs) and N independent links between CE and UEs. The CE always has set of tasks, that need to process at UEs. The UE can process one task at a time. The CE send M tasks to M different UEs in a time slot, $M < N$. The communication link quality between CE and UEs is time varying in nature due to fading, path loss and interference. This link quality may be one of two states good (1) or bad(0). CE sends a task to UE, and feedback such as ACK or NACK is obtained at end of time slot. Clearly the reward for each slot here depends on link state (quality). Here link states are not directly observable by the CE but are inferred using feedback signals, ACK or NACK which are obtained when a link is used for transmission. Further, the links are dynamically available i.e. in each slot a UE may or may not be available due to battery consideration or mobility issues. In such scenario, reward from using that UE is less than that of fully available UE. This may be due to slow processing of task to conserve the battery life. This dynamic availability is represented using a set of probabilities. We now briefly review some related work.

A. Literature overview

RMAB problems have been extensively studied in opportunistic communication systems, [7]–[9]. Here, a wireless channel is modeled using two state Markov chain (Gilbert-Elliot channel model). In general, the transmitter may not observe the state of channel, but it can observe the feedback signals ACK or NACK. This is referred to, as a partially observable model. In all of the RMAB, it is assumed that the model parameters are known. The goal of the transmitter is to determine the optimal sequence of play of the arms (channels) such that it maximizes long term reward. RMAB problems are known to be PSAPCE-hard, [10]. But in seminal paper [6], a heuristic index based policy was proposed, and it is now referred to as Whittle index policy. The Whittle index based policies for opportunistic communication systems were studied in [8], [9], where in [8], index policy is shown to be optimal for identical channel model and in [9], authors considered hidden Markov model which generalizes the work of [8]. The myopic policy was also investigated for restless bandits in [7] and it is shown to be optimal for identical channel model and reward conditions.

Recently, RMAB was used to model recommendation systems in [11]–[14], where Whittle index policy is used to recommend ads to the user.

All the above models of RMAB assume that each arm (communication channel or ad) is always available to play, and the decision in each slot is whether to play or not play an arm. However, availability of communication channels or ads to display can be dynamic. This problem is not analyzed with RMAB. In this paper, we study such problem. The multi-armed bandit problems with dynamic availability constraints have been studied for machine-repair problem in [15], where, if the machine breaks down, then it will be available in next time slot with some probability after getting repaired. This model assumes that the state is observable and the authors analyze index-type policy for rested bandits.

In literature [16]–[20], MAB problems which are different from RMAB, have been studied extensively for social networks and recommendation systems. In these MAB problems, each arm has a single state and the agent does not know the reward distribution for each arm. The objective is to play and learn the arm with the highest mean as quickly as possible. Alternatively, this is formulated as a regret minimization problem. The regret is the difference between expected reward obtained if the distribution were known and expected reward obtained under the policy with unknown distribution. It is shown that the regret grows sub-linear in time and it is also a function of network structures. In [18]–[20], MAB problem with side information is considered. All the above work assumes that arms are always available for play. In [21], authors studied sleeping MAB problem with dynamic availability of arms and regret bounds are derived. They employ the upper confidence bound (UCB) algorithm.

In this paper, we analyze RMAB problem with partially observable states and dynamic availability arms. One can also study UCB based learning algorithm for this problem. But our goal here is to propose and analyze an index policy.

B. Our contributions

We formulate the problem of an agent in a social network, gathering information from neighbors, as RMAB problem with dynamic availability of sources. We also consider the problem of using UEs in a wireless network, under time varying channel conditions and processing capabilities.

The problem formulation is given in Section II. We study Whittle index based policy. To use this policy, we first analyze a single-armed bandit problem in Section III. We further show that the optimal policy is of a threshold-type, and the arm is indexable. We next devise the algorithm to compute Whittle index for each arm. The algorithm is based on two timescales stochastic approximations. We also present simulation results in Section IV and compare the performance of Whittle index policy with that of myopic policy and uniform random policy.

II. Preliminaries and Model Description

There is an agent in a social or information network. The agent has connections to N neighbors which are its information sources. The agent can contact only $(M < N)$ neighbors for information. The system is assumed to be time slotted and it is indexed by t. The quality of information available at the source represented by a Markov chain with state space $\{0, 1\}$. Let $X_n(t)$ denote the state corresponding to information quality of source n at beginning of time slot t, $X_n(t) \in \{0, 1\}$. We suppose that each source has dynamic availability i.e. in a given slot it may or may not be available. When a source is not available, it may be leveraged to provide information by incurring an additional cost. Let $Y_n(t) \in \{0, 1\}$ represent the availability of the source n in time slot t and

$$Y_n(t) = \begin{cases} 1 & \text{if source } n \text{ is available,} \\ 0 & \text{if source } n \text{ is not available.} \end{cases}$$

Since the agent contacts M sources out of N in each time slot to gather information, we define $A_n(t) \in \{0,1\}$ as the action in slot t with the following interpretation.

$$A_n(t) = \begin{cases} 1 & \text{if source } n \text{ is contacted in slot } t, \\ 0 & \text{otherwise.} \end{cases}$$

Even when a source is unavailable, the agent can still contact it by incurring additional cost. Thus we can have $A_n(t) = 1$ in both available and unavailable scenarios.

We further assume that the information quality of a source varies in a Markovian manner. Hence $X_n(t)$ changes state at the end of time slot t according to transition probabilities that depend on $A_n(t)$, $Y_n(t)$ and it is defined as follows.

$$\Pr\{X_n(t+1) = j \mid X_n(t) = i, Y_n(t) = y, A_n(t) = a\} = P_{ij}^n(y,a).$$

If source n is contacted in slot t, then quality of information from source n is known exactly at the end of slot, i.e., state of source is known exactly. At the end of slot t, the agent makes a binary observation $Z_n^y(t)$ about source n that has been contacted about the relevance of the information received in the slot. Now,

$$Z_n^y(t) = \begin{cases} 1 & \text{if information from source } n \text{ is relevant,} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\rho_n(i,y)$ be the probability of $Z_n^y(t) = 1$ given that $X_n(t) = i$, $Y_n(t) = y$ and $A_n(t) = 1$.

$$\Pr\left(Z_n^y(t) = 1 \mid X_n(t) = i, Y_n(t) = y, A_n(t) = 1\right) = \rho_n(i,y).$$

We assume that $\rho_n(0,y) = 0$ and $\rho_n(1,y) = 1$ for all $y \in \{0,1\}$. When source n is not used, the agent do not know the quality of information, hence state of source n is unobservable. Hence, the agent maintains a belief $\pi_n(t)$ about the state of source n. Here, belief is the probability that the source is in state 0 given all past availability, actions, observations and given as

$$\pi_n(t) = \Pr\left(X_n(t) = 0 \mid \ (Y_n(s) = y_s, A_n(s), Z_n^{y_s}(s))_{s=1}^{t-1}\right).$$

We now define the reward as measure of the quality of information from different sources. When the agent uses source n, it obtains reward from the information it receives. This reward depends on current state of that source and availability of that source. Let $R_n^a(i,y)$ be the reward obtained from using source n given that $X_n(t) = i$, $Y_n(t) = y$, $A_n(t) = a$, and it is as follows.

$$\begin{aligned} R_n^1(i,1) &= r_{n,i}, \qquad & R_n^1(i,0) &= \eta_{n,i}, \\ R_n^0(i,1) &= 0, & R_n^0(i,0) &= 0. \end{aligned}$$

We further assume that $r_{n,0} = \eta_{n,0} = 0$, no reward from source n if it has $X_n(t) = 0$. Also, we suppose $r_{n,1} > \eta_{n,1}$, for all n. This implies that an unavailable source may be leveraged through an additional cost. Hence, the immediate reward is lower than when source is available. However, agent knows that the availability of sources is dynamic. This dynamic availability of each source n is modeled stochastically as probability of availability $\theta_n^a = \Pr\left(Y_n(t+1) = 1 | A_n(t) = a\right)$. Thus availability of a source varies according to Bernoulli distribution with parameter θ_n^a. This is known to the agent. Let H_t denote the history up to time t,

$$H_t := \left(Y_n(s) = y_s, A_n(s), Z_n^{y_s}(s)\right)_{1 \le n \le N, 1 \le s < t}.$$

We can describe the state of source n at time t by $S_n(t) = (\pi_n(t), Y_n(t)) \in [0,1] \times \{0,1\}$. $(S_1(t), \cdots S_N(t))$ is the state information of all the sources at the beginning of time slot t. The expected reward from using source n at time t given that $Y_n(t) = y$ is

$$\widetilde{R}_n^1(\pi_n(t), y) = \pi_n(t) R_n^1(0,y) + (1 - \pi_n(t) R_n^1(1,y).$$

In each slot, agent uses exactly M sources. Let $\phi(t)$ is the policy of agent such that $\phi(t) : H_t \to \{1, \cdots, N\}$ maps the history to M sources at each slot t. Let

$$A_n^\phi(t) = \begin{cases} 1 & \text{if } n \in \phi(t), \\ 0 & \text{if } n \notin \phi(t), \end{cases}$$

and $\sum_{n=1}^N A_n^\phi(t) = M$.

We are now ready to define the infinite horizon discounted reward under policy ϕ for initial state information $(\underline{\pi}, \underline{y})$, $\underline{\pi} = (\pi_1(1), \cdots, \pi_N(1))$ and $\underline{y} = (y_1(1), \cdots, y_N(1))$. It is given by

$$V_\phi(\underline{\pi}, \underline{y}) = \mathrm{E}^\phi\left(\sum_{t=1}^\infty \beta^{t-1} \left[\sum_{n=1}^N A_n^\phi(t) \widetilde{R}_n^1(\pi_n(t), Y_n(t))\right]\right).$$

Here, β is discount parameter, $0 < \beta < 1$. Then

$$\begin{aligned} & \phi^* = \arg\max_\phi V_\phi(\underline{\pi}, \underline{y}) \\ s.t. \quad & \textstyle\sum_{n=1}^N A_n^\phi(t) = M, \\ & \underline{\pi} \in [0,1]^N, \underline{y} \in \{0,1\}^N. \end{aligned} \tag{1}$$

The optimization problem (1) is a restless multi-arm bandit problem with availability constraints. Here, each source will correspond to an arm. The state of information quality of source n and its availability represent the state $S_n(t) = (\pi_n(t), Y_n(t))$ of an arm n. This is a generalized version of restless multi-arm bandits with partially observable states and availability constraints. Recall that this problem is known to be PSPACE-hard, [10]. A heuristic index based policies are developed for restless bandits in [6]. In this paper we consider index based policies. In such index polices, the dimensionality of the problem is reduced by calculating the index for each arm separately. The M arms with highest indices are played at each time slot. That is, the agent uses M sources with highest indices.

To use index policies, one requires to study relaxed version of optimization problem (1), where a subsidy w is introduced for not playing arm (not using source by agent), see [5], [6]. We first analyze agent with a single-armed bandit (a single source scenario) in next section.

III. A Single-armed bandit problem

For notation convenience, we will drop the subscript n. In the view of subsidy w, we can rewrite optimization problem (1) for a single-armed bandit as follows.

$$\phi^* = \arg\max_{\phi} V_\phi(\pi, y) = \mathrm{E}^\phi \left(\sum_{t=1}^{\infty} \beta^{t-1} \left[A^\phi(t) \widetilde{R}^1(\pi(t), Y(t)) + w(1 - A^\phi(t)) \right] \right)$$

for initial belief $\pi \in [0,1]$ and availability $y \in \{0,1\}$. Here, action $A(t)$ under policy ϕ is

$$A^\phi(t) = \begin{cases} 1 & \text{if } \phi(t) = 1, \\ 0 & \text{if } \phi(t) = 0. \end{cases}$$

We further simplify the model and assume that $P_{00}(y,a) = \mu_0$ and $P_{10}(y,a) = \mu_1$ for $a, y \in \{0,1\}$.[1]

Recall that $\pi(t) = \Pr(X(t) = 0 | H_t)$ and the using Bayes rule, we update the belief $\pi(t+1)$ in following manner.

$$\pi(t+1) = \begin{cases} \mu_1 & \text{if } A(t) = 1,\ Y(t) = y, \text{ and } Z^y(t) = 1, \\ \mu_0 & \text{if } A(t) = 1,\ Y(t) = y, \text{ and } Z^y(t) = 0, \\ \Gamma(\pi(t)) & \text{if } A(t) = 0, \text{ and } Y(t) = y, \end{cases}$$

for $y \in \{0,1\}$. Here, $\Gamma(\pi(t)) = \pi(t)\mu_0 + (1 - \pi(t))\mu_1$. If the agent uses a source in slot t, and it observed that information is relevant, i.e., $A(t) = 1$, and $Z^y(t) = 1$ for any $y \in \{0,1\}$, then state is known exactly and $X(t) = 1$, thus belief $\pi(t+1) = \mu_1$. Whereas if agent uses a source, $A(t) = 1$ but $Z^y(t) = 0$ then state is known exactly and $X(t) = 0$, thus belief $\pi(t+1) = \mu_0$. If source is not used, state is not observed but belief is updated.

From [22], we know that the $\pi(t)$ captures the information about the history H_t, and it is a sufficient statistic. It suggests that the optimal policies can be restricted to stationary Markov policies. In this, one can obtain the optimum value function by solving dynamic program. We first define the value function under initial action A_1 and availability Y_1.

$$\begin{aligned} V_T &:= \text{value function under } A_1 = 1, Y_1 = 1, \\ \widetilde{V}_T &:= \text{value function under } A_1 = 1, Y_1 = 0, \\ V_{NT} &:= \text{value function under } A_1 = 0, Y_1 = 1, \\ \widetilde{V}_{NT} &:= \text{value function under } A_1 = 0, Y_1 = 0. \end{aligned}$$

We can write the following.

$$V_T(\pi) = \rho(\pi) + \beta[(1-\pi)\{\theta^1 V(\mu_1) + (1-\theta^1)\widetilde{V}(\mu_1)\} + \pi\{\theta^1 V(\mu_0) + (1-\theta^1)\widetilde{V}(\mu_0)\}]$$

$$V_{NT}(\pi) = w + \beta[\theta^0 V(\Gamma_1(\pi)) + (1-\theta^0)\widetilde{V}(\Gamma_1(\pi))]$$

$$\widetilde{V}_T(\pi) = \xi(\pi) + \beta[(1-\pi)\{\theta^1 V(\mu_1) + (1-\theta^1)\widetilde{V}(\mu_1)\} + \pi\{\theta^1 V(\mu_0) + (1-\theta^1)\widetilde{V}(\mu_0)\}]$$

[1] In general, Markov model for source availability and unavailability could be different.

$$\widetilde{V}_{NT}(\pi) = w + \beta[\theta^0 V(\Gamma_0(\pi)) + (1-\theta^0)\widetilde{V}(\Gamma_0(\pi))]$$

Here $r(\pi) = (1-\pi)r_1, \eta(\pi) = (1-\pi)\eta_1,$. The optimal value function $V(\pi, y)$, is determined by solving the following dynamic program

$$\begin{aligned} V(\pi) &= \max\{V_T(\pi), V_{NT}(\pi)\}, \\ \widetilde{V}(\pi) &= \max\{\widetilde{V}_T(\pi), \widetilde{V}_{NT}(\pi)\}. \end{aligned} \tag{2}$$

A. Structural Results

We now derive structural results for value functions, convexity of value functions and a threshold type policy. We will derive all result for $\mu_0 > \mu_1$. This means that source is positively correlated, where a source that provides relevant information is more likely to it will provide relevant information in future also.

Lemma 1:

1) For fixed w, $V(\pi), V_T(\pi), V_{NT}(\pi), \widetilde{V}(\pi), \widetilde{V}_T(\pi)$ and $\widetilde{V}_{NT}(\pi)$ are convex functions of π.
2) For a fixed π, $V(\pi), V_T(\pi), V_{NT}(\pi), \widetilde{V}(\pi), \widetilde{V}_T(\pi)$ and $\widetilde{V}_{NT}(\pi)$ are non decreasing and convex in w.
3) For fixed subsidy w, β, and $\mu_0 > \mu_1$, the value functions $V(\pi), V_T(\pi)$ and $V_{NT}(\pi)$ are decreasing in π.
4) For fixed subsidy w, β, and $\mu_0 > \mu_1$, the value functions $\widetilde{V}(\pi), \widetilde{V}_T(\pi)$ and $\widetilde{V}_{NT}(\pi)$ are decreasing in π.
5) $(V_T(\pi) - V_{NT}(\pi))$ is decreasing in π.
6) $(\widetilde{V}_T(\pi) - \widetilde{V}_{NT}(\pi))$ is decreasing in π.

For completeness, we detailed sketch of the proof in Appendix.

We first define the threshold type policy and later we prove this results.

Definition 1: (Threshold type policy) A policy is said to be threshold type if one of the following is true.

1) The optimal action is to play an arm for all π.
2) The optimal action is to not play an arm for all π.
3) There exists a threshold π^* such that for all $\pi \leq \pi^*$ the optimal action is to play an arm and not to play an arm otherwise.

Theorem 1: For fixed w and β,

1) The optimal policy is threshold type for $V_T(\pi)$ and $V_{NT}(\pi)$. That is, either $V(\pi) = V_T(\pi)$ for all $\pi \in [0,1]$ or $V(\pi) = V_{NT}(\pi)$ for all $\pi \in [0,1]$ or there exists π^* such that
$$V(\pi) = \begin{cases} V_T(\pi) & \text{for } \pi \leq \pi^* \\ V_{NT}(\pi) & \text{for } \pi \geq \pi^*. \end{cases}$$
2) The optimal policy is threshold type for $\widetilde{V}_T(\pi)$ and $\widetilde{V}_{NT}(\pi)$. That is, either $\widetilde{V}(\pi) = \widetilde{V}_T(\pi)$ for all $\pi \in [0,1]$ or $\widetilde{V}(\pi) = \widetilde{V}_{NT}(\pi)$ for all $\pi \in [0,1]$ or there exists $\widetilde{\pi}$ such that
$$\widetilde{V}(\pi) = \begin{cases} \widetilde{V}_T(\pi) & \text{for } \pi \leq \widetilde{\pi} \\ \widetilde{V}_{NT}(\pi) & \text{for } \pi \geq \widetilde{\pi}. \end{cases}$$

Proof: The difference $(V_T(\pi) - V_{NT}(\pi))$ and $(\widetilde{V}(\pi) - \widetilde{V}_{NT}(\pi))$ is decreasing in π as per Lemma 1(4-5). Also the value functions $V_T(\pi), V_{NT}(\pi), \widetilde{V}_T(\pi)$ and $\widetilde{V}_{NT}(\pi)$ are convex in π. Which implies that there exists $\pi^* \in [0,1]$ and $\widetilde{\pi} \in [0,1]$ such that $V_T(\pi*) = V_{NT}(\pi^*), \widetilde{V}_T(\widetilde{\pi}) = \widetilde{V}_{NT}(\widetilde{\pi})$ or $V_T(\pi) > V_{NT}(\pi), \widetilde{V}_T(\pi) > \widetilde{V}_{NT}(\pi)$ or $V_T(\pi) < V_{NT}(\pi), \widetilde{V}_T(\pi) < \widetilde{V}_{NT}(\pi)$ for all π. Hence its proved. ■

Remark 1: In Lemma 1 and Theorem 1, we assumed that $0 < \theta^a < 1$ for $a \in \{0,1\}$. This suggest that there is interaction between value function from available to not available and vice-versa. If either $\theta^a = 0$ or $\theta^a = 1$ for all $a \in \{0,1\}$, results similar to Lemma 1 and Theorem 1 are studied in [8].

B. Indexability and Whittle index computation

Recall that our interest is to seek the index type policy. We use threshold policy result to show indexability and later provide index computation algorithm.

We now define indexability and index. Let $\mathcal{G}(w)$ be the subset of state vector S in which it is optimal to not play the arm with subsidy w, it is given as follows.

$$\begin{aligned}\mathcal{G}(w) :=&\{(\pi, y) \in [0,1] \times \{0,1\} : \\ &V_T(\pi, w) \leq V_{NT}(\pi, w), \widetilde{V}_T(\pi, w) \leq \widetilde{V}_{NT}(\pi, w)\}.\end{aligned} \tag{3}$$

For clarity, we have explicitly mentioned dependence of value function on w. Using set $\mathcal{G}(w)$, the indexability and index is defined as follows.

Definition 2: An arm is indexable if $\mathcal{G}(w)$ is increasing in subsidy w, i.e.,

$$w_2 \leq w_1 \Rightarrow \mathcal{G}(w_2) \subseteq \mathcal{G}(w_1).$$

Definition 3: The index of an indexable arm is defined as

$$w(\pi, y) := \inf\{w \in \mathbb{R} : (\pi, y) \in \mathcal{G}(w), \forall (\pi, y) \in S\}. \tag{4}$$

Remark 2:

1) Note that we can rewrite the definition of set $\mathcal{G}(w)$ in following way.

$$\mathcal{G}(w) = \{[\pi_L, 1] \times \{1\}, [\widetilde{\pi}_L, 1] \times \{0\}\},$$

where $\pi_L := \min\{\pi \in [0,1] : V_T(\pi, w) \leq V_{NT}(\pi, w)\}$, and $\widetilde{\pi}_L := \min\{\pi \in [0,1] : \widetilde{V}_T(\pi, w) \leq \widetilde{V}_{NT}(\pi, w)\}$. If the optimal policy is of threshold type, then π_L and $\widetilde{\pi}_L$ are singleton.

2) Here, the definition of indexability and index is motivated from work of [6] on restless bandits. In standard restless bandits, arms are assumed to be always available and $y = 0$ is not feasible option.
3) When $\theta^a = 0$ or $\theta^a = 1$ for all $a \in \{0,1\}$, our definitions of indexability and index are still valid.

To claim the indexability, we will require to show that $\pi_L(w)$ and $\widetilde{\pi}_L(w)$ is non-increasing in w. Now, we state the following lemma

Lemma 2: If

$$\begin{aligned}\left.\frac{\partial V_T(\pi, w)}{\partial w}\right|_{\pi=\pi_L(w)} &< \left.\frac{\partial V_{NT}(\pi, w)}{\partial w}\right|_{\pi=\pi_L(w)}, \\ \left.\frac{\partial \widetilde{V}_T(\pi, w)}{\partial w}\right|_{\pi=\widetilde{\pi}_L(w)} &< \left.\frac{\partial \widetilde{V}_{NT}(\pi, w)}{\partial w}\right|_{\pi=\widetilde{\pi}_L(w)},\end{aligned}$$

then $\pi_L(w)$ and $\widetilde{\pi}_L(w)$ is monotonically decreasing function of w.

Proof is along the same lines of Lemma 4 of [9]. Now, using Lemma 2 and Definition 2, we can show that single-armed restless bandit is indexable.

Theorem 2: If $\mu_0 > \mu_1$, then a single-armed restless bandit is indexable.

Proof of indexability for $0 < \theta^a < 1$ for all $a \in \{0,1\}$ is non-trivial and it is given in Appendix V-G. Whereas for $\theta^a = 1$ or 0, indexability can be shown easily by obtaining value function expression and then differentiating w.r.t. subsidy w, such result is studied in [8, Theorem 1].

We now use Definition 3 and restate the Whittle index definition as follows.

Definition 4 (Whittle's index): For a given belief $\pi \in [0,1]$ and availability $y \in \{0,1\}$, Whittle index $w(\pi, y)$ is the minimum subsidy w for which not playing the arm is the optimal action.

$$\begin{aligned}w(\pi, 1) &= \inf\{w \in \mathbb{R} : V_{NT}(\pi) = V_T(\pi)\}, \\ w(\pi, 0) &= \inf\{w \in \mathbb{R} : \widetilde{V}_{NT}(\pi) = \widetilde{V}_T(\pi)\}.\end{aligned} \tag{5}$$

When $\theta^a = 0, 1$ for all $a \in \{0,1\}$, the expression for index can be computed and this is given in [8, Section IV]. But for $\theta^a \in (0,1)$, it is very difficult to obtain closed form expression for value functions because there is coupling between value functions from when arm is available and arm is not available. Thus it is difficult to derive closed form expression for Whittle index formula.

Hence, we study numerical scheme for Whittle index computation. This scheme uses the threshold result of value functions and two-timescales stochastic approximations. In two-timescales stochastic approximations, we update w_t at slower timescales or natural timescales, and the value functions are updated using value iteration algorithm at faster timescales. This scheme here is inspired from stochastic approximation algorithms, see [23], [24].

In this scheme for fixed w, $y = 1$ and a threshold π, we know that $V_T(\pi, w) = V_{NT}(\pi, w)$. Using value iteration algorithm, we compute $V_T(\pi, w)$ and $V_{NT,w}(\pi, w)$ on faster time scales until difference $|V_T(\pi, w) - V_{NT,w}(\pi, w)|$ becomes smaller than tolerance h. To compute the index $w(\pi, 1)$, our algorithm starts with initial subsidy w_0 and it is updated iteratively at slower timescales according to following expression.

$$w_{t+1} = w_t + \alpha(V_T(\pi, w_t) - V_{NT}(\pi, w_t)).$$

These computations are performed till difference $|V_T(\pi, w_t) - V_{NT}(\pi, w_t)|$ is smaller than tolerance h.

Using similar procedure mentioned above, we update w_t with slower timescales and run value iteration for $\widetilde{V}_T(\pi, w_t)$

and $\widetilde{V}_{NT}(\pi, w_t)$ on faster timescales when π is threshold and $y = 0$. Hence this is used to compute the index $w(\pi, 0)$. The details are given in Algorithm 1. The convergence of two timescales stochastic approximation algorithm is presented in [23, Chapter 6].

Algorithm 1: Algorithm that computes Whittle index for the single arm

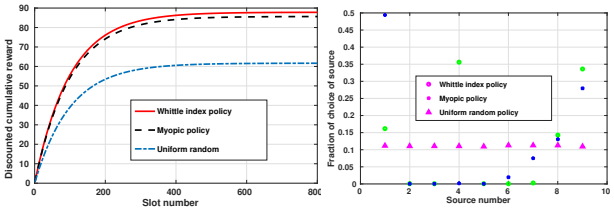
Input: Reward values r_1, η_1; Initial subsidy w_0, tolerance h, step size α.
Output: Whittle index,$w(\pi, y)$
if *(y==1)* **then**
  $w_t \leftarrow w_0$
  while $|V_T(\pi, w_t) - V_{NT}(\pi, w_t)| > h$ **do**
    $w_{t+1} = w_t + \alpha(V_T(\pi, w_t) - V_{NT}(\pi, w_t))$;
    $t = t + 1$;
    compute $V_T(\pi, w_t)$, $V_{NT}(\pi, w_t)$;
  end
else
  $w_t \leftarrow w_0$
  while $|\widetilde{V}_T(\pi, w_t) - \widetilde{V}_{NT}(\pi, w_t)| > h$ **do**
    $w_{t+1} = w_t + \alpha(\widetilde{V}_T(\pi, w_t) - \widetilde{V}_{NT}(\pi, w_t))$;
    $t = t + 1$;
    compute $\widetilde{V}_T(\pi, w_t)$, $\widetilde{V}_{NT}(\pi, w_t)$;
  end
end
return $w(\pi, y) = w_t$

IV. Numerical Results

We now illustrate performance of index based algorithm and compare it with different algorithms. We assume $M = 1$. The algorithms included in the comparative analysis are 1) Whittle index policy (WI)—contacts a source with highest Whittle index, 2) myopic policy (MP)—contacts a source with highest expected immediate reward, 3) uniform random policy(UR)—contacts a source randomly with uniform distribution.

Simulations were performed using MATLAB. In these simulations, the sources start in random states and random initial beliefs. The initial availability of sources are random. In each slot one source is contacted by agent according to the given policy. The reward is accumulated at the end of each slot from the source that is contacted and this reward is stored. These rewards are averaged over L iterations.

We will plot and compare the discounted cumulative reward that is obtained from these policies as function of time slots. We define source choice fraction as follows. Let $1_{m,t,l}$ be the indicator variable if source m is contacted in slot t, and lth iteration. Then $N_{m,l} := \frac{1}{T_{\max}} \sum_{t=1}^{T_{\max}} 1_{m,t,l}$, where $T_{\max}$ number of time slots for which simulations are performed. This is further averaged over L number of iterations. We call this as source m choice fraction. To gain further insight, we will plot this for all the sources for different policies.



a) Discounted cumulative reward b) Source choice fraction

Fig. 1. We plot a) discounted cumulative rewards as function of time slot for different policies and b) source choice fraction for different policies.

We use discount parameter $\beta = 0.99$ and $N = 9, 10$. In our numerical examples, we consider that 5 sources are always available and other sources have dynamically availability. We present three different numerical examples.

In our first example, $N = 9$ and we use non-identical transition probabilities for sources and identical probability of availability and rewards. We use following set of parameters.

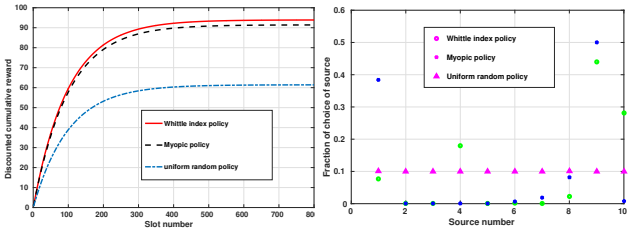
$$\begin{aligned}
\mu_o &= [0.66, 0.69, 0.75, 0.78, 0.63, 0.66, 0.69, 0.75, 0.78], \\
\mu_1 &= [0.28, 0.25, 0.2, 0.15, 0.3, 0.28, 0.25, 0.2, 0.15], \\
r_1 &= [1.3, 1.1, 0.9, 1.1, 1, 1.2, 1.2, 1.2, 1.2], \\
\eta_1 &= [0.7, 0.7, 0.7, 0.7, 0.75, 0.7, 0.7, 0.7, 0.7], \\
\theta^0 &= [1, 1, 1, 1, 1, 0.85, 0.85, 0.85, 0.85], \\
\theta^1 &= [1, 1, 1, 1, 1, 0.85, 0.85, 0.85, 0.85].
\end{aligned}$$

In Fig. 1-a) we plot the discounted cumulative reward as a function of time slots. It can be seen that the discounted cumulative reward under Whittle index policy (WI) is comparable with that of the Myopic policy (MP). We also observe that WI and MP yield higher discounted cumulative reward compared to that of uniform random policy. We also plot source choice fraction in Fig. 1-b). It suggests that myopic policy contacts source 1 most frequently compared to other sources and this is due to source 1 is always available and it has the highest reward. Whereas WI policy contacts from sources $\{4, 9\}$ more frequently even though they have less reward. This behavior of Whittle index policy due to it accounts for future rewards, availability of sources through the action value function. This is also determined by transition probabilities i.e., μ_0, and μ_1, where we observe that for source 4 and 9 difference $(\mu_0 - \mu_1)$ is very large compared to other sources.

In second example, $N = 10$ and we consider following parameters. We have used non-identical transition probabilities for sources, probability of availability and rewards.

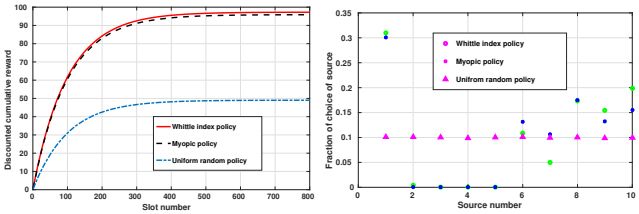
$$\begin{aligned}
\mu_o &= [0.66, 0.69, 0.75, 0.78, 0.63, 0.66, 0.69, 0.75, 0.78, 0.87], \\
\mu_1 &= [0.28, 0.25, 0.2, 0.15, 0.3, 0.28, 0.25, 0.2, 0.15, 0.1], \\
r_1 &= [1.3, 1.1, 0.9, 1.1, 1, 1.4, 1.3, 1.2, 1.35, 1.15], \\
\eta_1 &= [0.7, 0.7, 0.7, 0.7, 0.75, 0.9, 0.8, 0.7, 0.6, 0.7], \\
\theta^0 &= [1, 1, 1, 1, 1, 0.35, 0.45, 0.75, 0.85, 0.9], \\
\theta^1 &= [1, 1, 1, 1, 1, 0.35, 0.45, 0.75, 0.85, 0.9].
\end{aligned}$$

In Fig. 2-a) we plot the discounted cumulative reward verses time slot. The Whittle index policy yields higher discounted cumulative reward compared to UR policy. The myopic



a) Discounted cumulative reward b) Source choice fraction

Fig. 2. a) The discounted cumulative reward verses time slot for different policies and b) source choice fraction with different policies.



a) Discounted cumulative reward b) Source choice fraction

Fig. 3. a) The discounted cumulative reward verses time slot for different policies and b) source choice fraction with different policies.

policy has better cumulative reward than other policy. In Fig. 2-b), we have plotted source choice fraction. It suggests that Whittle index policy has a tendency to contacts the source from a smaller subset of sources, $\{4, 9, 10\}$. Myopic policy contacts the source from subset of sources, $\{1, 9\}$. UR policy contacts all the sources equally. The behavior of Whittle index policy is determined by μ_0 and μ_1. Here, index is dependent on these parameters, availability and it accounts for future rewards through the action value function. The behavior of myopic policy depends on value of r_1 and observe that sources 1 and 9 have highest values of r_1. Thus MP contacts sources 1 and 9 more frequently compared to other sources.

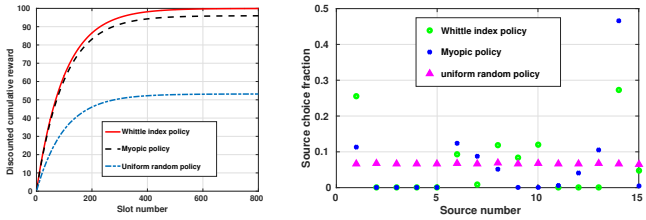
Finally, we illustrate an example with identical transition probabilities for all sources but non-identical probability of availability and rewards. Parameters used are as follows. $N = 10$, $\mu_{i,0} = 0.9$ and $\mu_{i,1} = 0.1$ for $1 \leq i \leq N$.

$$\begin{aligned} r_1 &= [1.25, 0.8, 0.8, 0.8, 0.8, 1.5, 1.3, 1.25, 1.2, 1.2], \\ \eta_1 &= [0.7, 0.7, 0.7, 0.7, 0.75, 0.7, 0.7, 0.7, 0.7, 0.7], \\ \theta^0 &= [1, 1, 1, 1, 1, 0.35, 0.45, 0.75, 0.85, 0.9], \\ \theta^1 &= [1, 1, 1, 1, 1, 0.35, 0.45, 0.75, 0.85, 0.9]. \end{aligned}$$

From Fig. 3-a) we observe that the Whittle index policy and myopic policy yield higher discounted cumulative reward compared to UR policy. WI and Myopic policy performance is very similar. We notice from Fig. 3-b) that both WI policy and myopic policy contacts sources from $\{1, 10\}$, this is due to large reward and better availability.

A. Additional example

We use discount parameter $\beta = 0.99$ and $N = 15$. In our numerical examples, we consider that 5 sources are always available and other sources have dynamically availability.



a) Discounted cumulative reward b) Source choice fraction

Fig. 4. We plot a) discounted cumulative rewards as function of time slot for different policies and b) source choice fraction for different policies.

Parameters used are as follows.

$$\begin{aligned} \mu_o &= [0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.66, 0.69, 0.75, 0.78, 0.87], \\ \mu_1 &= [0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.28, 0.25, 0.2, 0.15, 0.1], \\ r_1 &= [1.25, 0.8, 0.8, 0.8, 0.8, 1.5, 1.3, 1.25, 1.2, 1.2, 1.4, 1.3, 1.2, 1.35, 1.15], \\ \eta_1 &= [0.7, 0.7, 0.7, 0.7, 0.75, 0.7, 0.7, 0.7, 0.7, 0.7, 0.9, 0.8, 0.7, 0.6, 0.7], \\ \theta^0 &= [1, 1, 1, 1, 1, 0.35, 0.45, 0.75, 0.85, 0.9, 0.35, 0.45, 0.75, 0.85, 0.9], \\ \theta^1 &= [1, 1, 1, 1, 1, 0.35, 0.45, 0.75, 0.85, 0.9, 0.35, 0.45, 0.75, 0.85, 0.9]. \end{aligned}$$

From Fig. 4-a) we observe that the Whittle index policy yields higher discounted cumulative reward compare to previous examples. This is due to increase in number of sources available to the agent. The gain over Myopic and uniform random policy can be improved further with more number of sources.

V. Concluding Remarks

We formulated problem of information gathering in a social network with dynamic availability of sources and time varying information quality using RMAB model. We studied numerical scheme for Whittle index computation and compared performance with that of myopic and uniform random policy.

The immediate possible generalisations of our current work are as follows. 1) Learning algorithm when model parameters are unknown. 2) Multi-agent problem in social network, where these agents compete for gathering reliable information from sources.

References

[1] J. Vehkoo, "Crowdsourcing in investigative journalism," *Reuters Institute for the Study of Journalism*, 2013.

[2] Yiangos Papanastasiou, "Fake news propagation and detection: A sequential model," *SSRN*, 2017.

[3] Financial Times, "Facebook to pay fact-checkers to combat fake news," 6th April 2017.

[4] N. Verma, K. R. Fleischmann, and K. S. Koltai, "Human values and trust in scientific journals, the mainstream media and fake news," *Proc. of the Assoc. for Info. and Tech.*, vol. 54, no. 1, pp. 426–435, 2017.

[5] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed Bandit Allocation Indices*, Wiley, 2011.

[6] P. Whittle, "Restless bandits: activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, no. A, pp. 287–298, 1988.

[7] A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Trans. on Info. Theory*, vol. 55, no. 9, pp. 4040–4050, 2009.

[8] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[9] R. Meshram, D. Manjunath, and A. Gopalan, "On the Whittle index for restless multi-armed hidden Markov bandits," *ArXiv e-prints*, March 2016.
[10] Christos H. Papadimitriou and John N. Tsitsiklis, "The complexity of optimal queuing network control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, May 1999.
[11] R. Meshram, A. Gopalan, and D. Manjunath, "Optimal recommendation to users that react: Online learning for a class of POMDPs," in *Proc. of CDC*, Dec 2016, pp. 7210–7215.
[12] R. Meshram, A. Gopalan, and D. Manjunath, "A hidden Markov restless multi-armed bandit model for playout recommendation systems," *ArXiv e-prints*, April 2017.
[13] R. Meshram, A. Gopalan, and D. Manjunath, "Restless bandits that hide their hand and recommendation systems," in *Proc. of COMSNETS*, Jan 2017, pp. 206–213.
[14] K. E. Avrachenkov and V. S. Borkar, "Whittle index policy for crawling ephemeral content," *IEEE Transactions on Control of Network Systems*, , no. 99, 2016.
[15] S. Dayanik, W. Powell, and K. Yamazaki, "Index policies for discounted bandit problems with availability constraints," *Advances in Applied Probability*, vol. 40, no. 02, pp. 377–400, 2002.
[16] S. Bubeck and N. C. Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.
[17] N. Bianchi, C. Gentile, and G. Zappella, "A gang of bandits," in *Proc. of NIPS*, 2013, pp. 737–745.
[18] S. Buccapatnam, A. Eryilmaz, and N. B. Shroff, "Multi-armed bandits in the presence of side observations in social networks," in *Proc. of CDC*, Dec 2013, pp. 7309–7314.
[19] S. Buccapatnam, J. Tan, and L. Zhang, "Information sharing in distributed stochastic bandits," in *Proc. of INFOCOM*, April 2015, pp. 2605–2613.
[20] L. E. Celis and F. Salehi, "Lean from thy neighbor: Stochastic & adversarial bandits in a network," *ArXiv:1704.04470*, 2017.
[21] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma, "Regret bounds for sleeping experts and bandits," *Machine learning*, vol. 80, no. 2-3, pp. 245–272, 2010.
[22] Dimitri P Bertsekas and Bertsekas, *Dynamic programming and optimal control*, vol. 1-2, Athena Scientific Belmont, MA, 2 edition, 1995.
[23] V. S. Borkar, *Stochastic approximation: A dynamical systems viewpoint*, Cambridge Univ. Press, 2008.
[24] V. S. Borkar, G. S. Kasbekar, S. Pattathil, and P. Y. Shetty, "Opportunistic scheduling as restless bandits," *ArXiv:1706.09778*, 2017.
[25] Dimitri P Bertsekas and Bertsekas, *Dynamic programming and optimal control*, vol. 2, Athena Scientific Belmont, MA, 2 edition, 1995.

APPENDIX

A. Proof of Lemma 1.1

The proof is similar to the proof of Lemma 2 in [9].

B. Proof of Lemma 1.2

Using induction technique, We can rewrite $V_T(\pi)$ and $V_{NT}(\pi)$, in form of $V_{n+1,T}(\pi,w)$ and $V_{n+1,NT}(\pi,w)$ as function of w. We can see that $V_1(\pi,w)$ is monotone non decreasing and convex in w. $V_{n+1,T}(\pi,w)$ is constant independent of w. $V_{n+1,NT}(\pi,w)$ is the sum of three non decreasing function of w. The convexity is preserved under max operation so $V_{n+1}(\pi,w)$ is also non decreasing and convex in w and using induction, all $V_n(\pi,w)$ follows the same. As $V_n(\pi,w)\to V(\pi,w)$ and this complete the proof for $V(\pi)$. Similarly, we can show for other value functions.

C. Proof of Lemma 1.3

The proof is done by induction technique. Assume that $V_n(\pi)$ and $\widetilde{V}_n(\pi)$ is non increasing in π. Lets take $\pi'\geq\pi$ and playing an arm is optimal. Then induction step

$$V_{n+1}(\pi)=\rho(\pi)+\beta[(1-\pi)\{\theta^1V_n(\mu_1)+(1-\theta^1)\widetilde{V}_n(\mu_1)\}+\pi\{\theta^1V_n(\mu_0)+(1-\theta^1)\widetilde{V}_n(\mu_0)\}]$$

Here $\rho(\pi)$ is decresing in π, i.e. $\rho(\pi')<\rho(\pi)$ for $\pi'>\pi$. Hence

$$V_{n+1}(\pi)\geq\rho(\pi')+\beta[(1-\pi)\{\theta^1V_n(\mu_1)+(1-\theta^1)\widetilde{V}_n(\mu_1)\}+\pi\{\theta^1V_n(\mu_0)+(1-\theta^1)\widetilde{V}_n(\mu_0)\}]$$

From our assumptions $\mu_0>\mu_1$, we get stochastic ordering on observation probability, i.e., $[\pi,1-pi]^T\leq_s[\pi',1-\pi']^T$. and $V_n(\pi),\widetilde{V}(\pi)$ are decreasing in π, then we have

$$V_{n+1}(\pi)\geq\rho(\pi')+\beta[(1-\pi')\{\theta^1V_n(\mu_1)+(1-\theta^1)\widetilde{V}_n(\mu_1)\}+\pi'\{\theta^1V_n(\mu_0)+(1-\theta^1)\widetilde{V}_n(\mu_0)\}]$$

$$V_{n+1}(\pi)\geq V_{n+1}(\pi').$$

Similarly we can show that $\widetilde{V}_{n+1}(\pi)\geq\widetilde{V}_{n+1}(\pi')$. This is true for every $n\geq 1$. From Chapter 7 of [22] and Proposition 2.1 of Chapter 2 of [25], $V_n(\pi)\to V(\pi)$, uniformly and similarly $\widetilde{V}_n(\pi)\to\widetilde{V}(\pi)$. Hence $V(\pi)\geq V(\pi')$ and $\widetilde{V}(\pi)\geq\widetilde{V}(\pi')$ for $\pi'\geq\pi$.

Next we prove, $V_T(\pi)$ and $V_{NT}(\pi)$ is non increasing in π.

$$V_T(\pi)=\rho(\pi)+\beta[(1-\pi)\{\theta^1V(\mu_1)+(1-\theta^1)\widetilde{V}(\mu_1)\}+\pi\{\theta^1V(\mu_0)+(1-\theta^1)\widetilde{V}(\mu_0)\}] \quad (6)$$

$$V_{NT}(\pi)=w+\beta[\theta^0V(\Gamma_1(\pi))+(1-\theta^0)\widetilde{V}(\Gamma_1(\pi))] \quad (7)$$

For $\pi_1>\pi_2$,

$$V_T(\pi_1)-V_T(\pi_2)=(\pi_1-\pi_2)\beta\theta^1(V(\mu_0)-V(\mu_1))+(\pi_1-\pi_2)\beta(1-\theta^1)(\widetilde{V}(\mu_0)-\widetilde{V}(\mu_1))$$

Using above result and $\mu_0>\mu_1$, $V_T(\pi)$ is non increasing in π. Similarly, $V_{NT}(\pi)$ is non increasing in π.

D. Proof of Lemma 1.4

The proof is similar to the proof of Lemma 2(3) in Appendix V-C.

E. Proof of Lemma 1.5

Let $D(\pi)=V_T(\pi)-V_{NT}(\pi)$ and $D(\pi)$ is decreasing in π, i.e $D(\pi)<D(\pi')$ for $\pi>\pi'$. This implies that we need to show

$$V_T(\pi)-V_{NT}(\pi)<V_T(\pi')-V_{NT}(\pi') \quad (8)$$

Rearranging 8 we need to show

$$V_T(\pi)-V_T(\pi')<V_{NT}(\pi)-V_{NT}(\pi') \quad (9)$$

Now, the right hand side of the (9),

$$\begin{aligned}V_{NT}(\pi)-V_{NT}(\pi')&=\beta\theta^0\{V(\Gamma_1(\pi))-V(\Gamma_1(\pi'))\}\\&\quad+\beta(1-\theta^0)\{\widetilde{V}(\Gamma_1(\pi))-\widetilde{V}(\Gamma_1(\pi'))\}\\&\geq\beta\theta^0\{V(\mu_0)-V(\Gamma_1(\pi'))\}\\&\quad+\beta(1-\theta^0)\{\widetilde{V}(\mu_0)-\widetilde{V}(\Gamma_1(\pi'))\}\\&\geq\beta\theta^0\{V(\mu_0)-V(\mu_1)\}\\&\quad+\beta(1-\theta^0)\{\widetilde{V}(\mu_0)-\widetilde{V}(\mu_1)\}\end{aligned}$$

The left hand side of the (9),

$$V_T(\pi) - V_T(\pi') = (\rho(\pi) - \rho(\pi')) + \beta(\pi - \pi')\theta^1\{V(\mu_0) - V(\mu_1)\} + \beta(\pi - \pi')(1 - \theta^1)\{\widetilde{V}(\mu_0) - \widetilde{V}(\mu_1)\}$$

Note that $\rho(\pi) - \rho(\pi') = r_1(\pi' - \pi) < 0$ because $\pi > \pi'$. Also note from previous two expressions of difference in value functions that for $\theta^0 = \theta^1$, we can easily see that Eqn (9) is true.

But even for $\theta^0 \neq \theta^1$, Eqn (9) is true because $\rho(\pi) - \rho(\pi') < 0$ and other terms of $V_T(\pi) - V_T(\pi')$ are scaled by $\pi - \pi'$ which is positive and if this difference is small, other terms are going to be small.

F. Proof of Lemma 1.6

The proof is similar to the proof of Lemma 1(5) in Appendix V-E.

G. Proof of Theorem 2

The following inequalities obtain using induction technique, .

$$\left|\frac{\partial V(\pi,w)}{\partial w}\right|, \left|\frac{\partial V_T(\pi,w)}{\partial w}\right|, \left|\frac{\partial V_{NT}(\pi,w)}{\partial w}\right| \leq \frac{1}{1-\beta}$$

and

$$\left|\frac{\partial \widetilde{V}(\pi,w)}{\partial w}\right|, \left|\frac{\partial \widetilde{V}(\pi,w)}{\partial w}\right|, \left|\frac{\partial \widetilde{V}_{NT}(\pi,w)}{\partial w}\right| \leq \frac{1}{1-\beta}$$

Also,

$$\frac{\partial V_T(\pi,w)}{\partial w} = \beta\left[(1-\pi)\{\theta^1\frac{\partial V(\mu_1,w)}{\partial w} + (1-\theta^1)\frac{\partial \widetilde{V}(\mu_1,w)}{\partial w}\} + \pi\{\theta^1\frac{\partial V(\mu_0,w)}{\partial w} + (1-\theta^1)\frac{\partial \widetilde{V}(\mu_0,w)}{\partial w}\}\right]$$

and

$$\frac{\partial V_{NT}(\pi,w)}{\partial w} = 1 + \beta\{\theta^0\frac{\partial V(\Gamma_1(\pi),w)}{\partial w} + (1-\theta^0)\frac{\partial V(\Gamma_1(\pi),w)}{\partial w}\}.$$

Now from Lemma 2, we require the difference $\frac{\partial V_{NT}(\pi,w)}{\partial w} - \frac{\partial V_T(\pi,w)}{\partial w}$ to be nonegative at $\pi_L(w)$ and $\widetilde{\pi}_L(w)$. This reduces to following expression.

$$\begin{aligned}&\left[(1-\pi)\{\theta^1\frac{\partial V(\mu_1,w)}{\partial w} + (1-\theta^1)\frac{\partial \widetilde{V}(\mu_1,w)}{\partial w}\}\right.\\ &\left.+ \pi\{\theta^1\frac{\partial V(\mu_0,w)}{\partial w} + (1-\theta^1)\frac{\partial \widetilde{V}(\mu_0,w)}{\partial w}\}\right]\\ &- \left[\theta^0\frac{\partial V(\Gamma_1(\pi),w)}{\partial w}\right.\\ &\left.+ (1-\theta^0)\frac{\partial V(\Gamma_1(\pi),w)}{\partial w}\right] < \frac{1}{\beta}.\end{aligned} \tag{10}$$

Note that we can provide upper bound on LHS of above expression and it is upper bounded by $2/(1-\beta)$. If $\beta < 1/3$, Eqn. (10) is satisfied. $\pi_L(\eta)$ is decreasing in w. Similarly $\widetilde{\pi}_L(\eta)$ is decreasing in w. And claim follows.