Learning Restless Bandits in Application to Call-based Preventive Care Programs for Maternal Healthcare

Arpita Biswas\textsuperscript{1}, Gaurav Aggarwal\textsuperscript{1}, Pradeep Varakantham\textsuperscript{2} and Milind Tambe\textsuperscript{1}

\textsuperscript{1} Google Research
\textsuperscript{2} Singapore Management University
\{arpita, gauravaggarwal, milindtambe\}@google.com, pradeepv@smu.edu.sg

1 Introduction

This paper focuses on learning index-based policies in restless multi-armed bandits (RMAB) with applications to public health concerns such as maternal health. Maternal health is a very important public health concern. It refers to the health of women during their pregnancy, childbirth, and the postnatal period. Although maternal health has received significant attention [World Health Organization, 2015], the number of maternal deaths remains unacceptably high, mainly because of the delay in obtaining adequate care [Thaddeus and Maine, 1994]. Most maternal deaths can be prevented by providing timely preventive care information. However, such information is not easily accessible by underprivileged and low-income communities. For ensuring timely information, a non-profit organization, called ARMPAN [2015], carries out a free call-based program called mMitra for spreading preventive care information among pregnant women. Enrolment in this program happens through hospitals and non-government organizations. Each enrolled woman receives around 140 automated voice calls, throughout their pregnancy period and up to 12 months after childbirth. Each call equips women with critical life-saving healthcare information. This program provides support for around 80 weeks. To achieve the vision of improving the well-being of the enrolled women, it is important to ensure that they listen to most of the information sent to them via the automated calls. However, the organization observed that, for many women, their engagement (i.e., the overall time they spend listening to the automated calls) gradually decreases. One way to improve their engagement is by providing an intervention (that would involve a personal visit by healthcare worker). These interventions require the dedicated time of the health workers, which is often limited. Thus, only a small fraction of the overall enrolled women can be provided with interventions during a time period. Moreover, the extent to which the engagement improves upon intervention varies among individuals. Hence, it is important to carefully choose the beneficiaries who should be provided interventions at a particular time period.

This is a challenging problem owing to multiple key reasons: (i) Engagement of the individual beneficiaries is uncertain and changes organically over time; (ii) Improvement in the engagement of a beneficiary post-intervention is uncertain; (iii) Decision making with respect to interventions (which beneficiaries should have intervention) is sequential, i.e., decisions at a step have an impact on the state of beneficiaries and decisions to be taken at the next step; (iv) Number of interventions are budgeted and are significantly smaller than the total number of beneficiaries.

Due to the uncertainty, sequential nature of decision making, and weak dependency amongst patients through a budget, existing research [Lee et al., 2019; Mate et al., 2020; Bhattacharya, 2018] in health interventions has justifiably employed RMABs. However, existing research focuses on the planning problem assuming a priori knowledge of the underlying uncertainty model, which can be quite challenging to obtain. Thus, we focus on learning intervention decisions in absence of the knowledge of underlying uncertainty.

Contributions: First, we represent the maternal health intervention problem as an RMAB with unknown uncertainty model. If the underlying transition probabilities were known, one could compute the Whittle Index [Whittle, 1988] and selection would depend on those indices. Our second contribution leverages Q-Learning to dynamically learn the Whittle index for (indexable) RMABs, when the transition probabilities are unknown. We show that, under mild assumptions, our method WIQL converges to the optimal solution asymptotically. Finally, we evaluate WIQL on (1) a well-studied RMAB example and (2) the RMAB problem in the context of maternal healthcare. We observe that WIQL outperforms other benchmark algorithms.

2 Related Work

RMAB problem was introduced by Whittle [1988]. He showed that the relaxation of the RMAB problem can be solved in polynomial time using a heuristic called Whittle Index policy, and is optimal under indexability. There is a vast literature that focuses on specific classes of RMABs that have indexability and hence allow for optimal Whittle index policies [Akbarzadeh and Mahajan, 2019; Mate et al., 2020; Bhattacharya, 2018; Lee et al., 2019; Glazebrook et al., 2006; Hsu, 2018; Sombabu et al., 2020; Liu and Zhao, 2010]. These papers focus on problems where transition probabilities are known beforehand. Instead, our focus is on providing learning methods based on the Whittle index when the transition are unknown a priori.

A few papers have focused on learning methods for RMABs. Fu et al. [2019] provide a Q-learning method where
the value $Q(\lambda, s, a)$ is defined based on the Whittle index $\lambda$. However, they do not provide proof of convergence to optimal and experimentally, do not learn (near-) optimal policies. Avrachenkov and Borkar [2020] provide a fundamental change to the Q-value definition with the aim of estimating optimal whittle index policy. However, their convergence proof requires all the arms to be homogeneous having the same underlying uncertainties. We propose a Q-learning method that is not only shown to theoretically converge but also outperforms the above mentioned methods, empirically.

Another relevant line of work predicts adherence to a health program and effects of interventions, by formulating these as supervised learning problems [Killian et al., 2019; Cross et al., 2019; Nishitaka et al., 2020; Son et al., 2010; Howes et al., 2012; Lauffenburger et al., 2018]. These papers assume the training data to be available beforehand and do not consider the dynamic nature of the problem, where feedback is received after an intervention is provided which are then used in making future decisions.

3 Preliminaries

RMABs have been used to model various stochastic scheduling problems. An RMAB problem instance is a 3-tuple $(N, M, \{MDP_i\}_{i \in N})$, where $N$ is the set of arms, $M$ is the budget restriction denoting how many arms can be pulled at a given time, and $MDP_i$ an associated Markov Decision Process for each arm $i$. An MDP for an arm $i$, consists of a set of states $S$, a set of actions $A$, transition probabilities $P : S \times A \times S \mapsto [0, 1]$, and reward function $R : S \times A \mapsto \mathbb{R}$. The action set $A$ consists of two actions: an active action (1) and a passive action (0). At each time step $t$, an action $A_i(t) \in A$ is chosen for arm $i$, such that $\sum_i A_i(t) = M$. Then, each arm $i$ transitions to a new state and observes a reward, according to the state transition process of MDP$_i$. Let $X_i(t) \in S$ and $R_i^{X_i(t)}(A_i(t))$ denote the current state and reward obtained at time $t$ respectively. Now, a policy $\pi : X_1(t) \times \ldots \times X_N(t) \mapsto \{A_i(t)\}_{i \in N}$ can be defined as a mapping from the current states of all arms to the actions to be taken on each arm. Thus, given a policy $\pi$, the action on an arm $i$ is denoted as $A_i^\pi(t)$. The goal is to find a policy $\pi^*$ that maximizes the total expected reward under budget constraint.

$$\max_{\pi} \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \left[ \sum_{h=0}^{t-1} \sum_{i \in N} R_i^{X_i(h)}(A_i^\pi(h)) \right]$$

(1)

subject to

$$\sum_{i \in N} A_i^\pi(t) = M \quad \text{for all } t \in \{1, 2, \ldots\}$$

This problem is PSPACE-hard [Papadimitriou and Tsitsiklis, 1994]. To deal with the computational hardness, Whittle [1988] proposed an index-based heuristic policy based on the Lagrangian relaxation of the RMAB problem (1):

$$\max_{\pi} \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \left[ \sum_{h=0}^{t-1} \left( R_i^{X_i(h)}(A_i^\pi(h)) + \lambda \cdot (1 - A_i^\pi(h)) \right) \right]$$

(2)

Whittle showed that this problem can be decoupled and solved for each bandit by computing the index $\lambda(Z)$ which acts like a subsidy that needs to be given to an arm $i$ at state $Z$, so that taking the action 0 becomes as beneficial as taking action 1. Assuming indexability, choosing arms with highest Whittle Indices leads to an optimal solution for Equation (2).

Definition 1 (Indexability) Let $\Phi(\lambda)$ be the set of all states for which it is optimal to take action a passive action when taking an active action costs $\lambda$. An arm is called indexable if $\Phi(\lambda)$ monotonically increases from $\emptyset$ to $S$ when $\lambda$ increases from $-\infty$ to $+\infty$. An RMAB problem is indexable if all the arms are indexable.

We now use these concepts to formalize the selection problem for maternal healthcare.

4 The Model

We consider the problem of selecting a subset of $M$ beneficiaries at each time step, out of total $N$ enrolled women, for providing intervention so as to maximize the total benefit of the intervention. We formulate this problem as an RMAB where the beneficiaries are the arms. We assume that their engagement patterns (how much information they listen to which are sent to them via automated calls) are MDPs with three “behavioral” states \{S, P, Z\}.

1. Self Motivated (S): In this state, the beneficiary picks up most of the calls and listens to them almost completely, thus, showing high engagement. Naturally, there is not much necessity of providing interventions in this state.

2. Persuadable (P): In this state, the beneficiary engages less frequently but, there is a possibility of increasing engagement when an intervention is given to them. It is important to identify beneficiaries who have reached this state, so that proper interventions can be provided to them.

3. Lost Cause (L): Typically, the extent of engagement is very low in this state, and there is a high chance that the engagement remains low regardless of any intervention. Once a beneficiary reaches this state, they are likely to remain disengaged in near future.

Each behavioral state is an abstraction of their engagement values—for example, listening to more than 50% of the information can be considered as being in state $S$, 5–50% is considered as being in state $P$, and 0–5% is equivalent to being in state $L$. Let $A_i(t) \in \{0, 1\} = 1$ denotes that beneficiary $i$ is being selected for an intervention and 0 denotes she is not selected. At most $M$ beneficiaries can be selected at each time $t$, $\sum_i A_i(t) = M$. Depending on the action taken, each beneficiary changes its state according to the transition probabilities, as illustrated in Figure 1.
When a beneficiary is in the persuadable state and intervention is provided \((A_i(t) = 1)\), then she is likely to transition to \(S\) whereas, when no intervention is provided \((A_i(t) = 0)\), the beneficiary is more likely to transition to \(L\). These transitions are denoted as \(p_{i,1}^{S,L}\) and \(p_{i,1}^{L,S}\) for action 1 and 0, respectively. Note that the outgoing transition probabilities from states \(S\) and \(L\) remain the same irrespective of the action, and differs only for the state \(P\). Thus, the benefit of an intervention is maximum when an intervention is provided at state \(P\). In absence of a priori knowledge of the transition probabilities we cannot directly computing the Whittle Indices and hence, we propose Q-learning to learn these indices.

5 Whittle Index for Q-Learning

Q-Learning [Watkins, 1989; Watkins and Dayan, 1992] is a well-studied reinforcement learning algorithm used for estimating the optimal state-action value \(Q^*(Z, a)\).

\[
Q^*(Z, a) := R^Z + \sum_{Z' \in \{S,P,L\}} P(Z, a, Z') \cdot V^*(Z'),
\]

where \(V^*(Z)\) is the optimal expected value of a state,

\[
V^*(Z) := \max_{a \in [0,1]} \left( R^Z + \sum_{Z' \in \{S,P,L\}} P(Z, a, Z') \cdot V^*(Z') \right).
\]

Q-Learning estimates \(Q^*\) using point samples—at each time \(t\), an agent (policy maker) takes an action \(a\) using estimated \(Q\) values at the current state \(Z\), a reward \(R\) is observed, a new state \(Z'\) is reached, and \(Q\) values are updated according to the following update rule:

\[
Q^{t+1}(Z, a) \leftarrow (1 - \alpha_t(Z, a)) \cdot Q^t(Z, a) + \alpha_t(Z, a) \cdot \left( R + \max_{a' \in [0,1]} Q^t(Z', a') \right) \quad (3)
\]

Here, \(\alpha_t(\cdot) \in [0,1]\) is called the learning parameter. Setting \(0 < \alpha < 1\), helps to strike a balance between the new observations and the old ones. The convergence of Q-Learning to the optimal \(Q^*\) values has been well established in the literature [Watkins and Dayan, 1992; Jaakkola et al., 1994; Borkar and Meyn, 2000] under some mild assumptions on the value of \(\alpha(\cdot)\). We build on the results to provide a Q-Learning approach for learning Whittle-index policy in RMABs.

WIQL (Algorithm 1) uses an \(\epsilon\)-decay policy to select a set of \(M\) arms at each time step \(t\). During early time steps, arms are more likely to be selected uniformly at random and as time proceeds, more priority is given to the arms with a higher value of their estimated \(\lambda_i(X_i(t))\) values. The selected set of \(M\) arms is denoted as \(\Psi\). Active action is taken on the arms in the set \(\Psi\), and passive action is taken on the other arms. Note that each arm is a restless bandit and hence, each arm transitions to a new state and observes a reward. These observations are then used for updating the \(Q\) values (Step 12), as in Equation 3. However, while updating \(Q_i(Z, a)\), the learning parameter \(\alpha(c_i, a)\) decreases with increase in \(c_i^a\) (the number of times the action \(a\) was taken in the arm \(i\) when it was at state \(Z\)); for example, \(\alpha(c_i^a) = 1/(c_i^a + 1)\) satisfies this criteria. The updated \(Q\) values are then used to estimate the Whittle index \(\lambda(X_i(t))\).

### Algorithm 1: Whittle Index for Q-Learning (WIQL)

Input: \(N, M, X_i(0) \in S\), for all \(i \in N, c_i(\cdot)\).
1. Initialize: \(Q_i^0(Z, a) \leftarrow 0\) and \(\lambda_i^0(Z) \leftarrow 0\) for each \(i, Z\) and \(a\).
2. for \(t = 1, \ldots\) do
3. // Select \(M\) arms using \(\epsilon\)-decay policy
4. \(\epsilon \leftarrow \frac{\epsilon}{M^t}\)
5. With probability \(\epsilon\), select \(M\) arms uniformly at random. Otherwise, select top \(M\) arms according to \(\lambda_i^t(Z)\). Let \(\Psi\) be the selected arms.
6. // Take suitable actions on the arms
7. for \(i = 1, \ldots, M\) do
8. if \(i \in \Psi\ then
9. Take action \(A_i(t) = 1\) on arm \(i\).
10. else
11. Take action \(A_i(t) = 0\) on arm \(i\).
12. Observe reward \(r\) and next state \(X_i(t + 1)\).
13. // Update \(Q\) and \(\lambda\) values
14. for \(i = 1, \ldots, M\) do
15. for \(Z, a\) do
16. \(Q_i^{t+1}(Z, a) \leftarrow (1 - \alpha_t(Z, a)) \cdot Q_i^t(Z, a) + \alpha_t(Z, a) \cdot \left( R + \max_{a' \in [0,1]} Q_i^t(Z', a') \right) \quad (3)\)
17. \(\lambda_i^{t+1}(X_i(t) = Z) \leftarrow Q_i^{t+1}(Z, 1) - Q_i^{t+1}(Z, 0)\)
18. \(\sum_k \alpha(c_i^k, Z, a) = \infty\) and \(\sum_k \alpha(c_i^k, Z, a)^2 < \infty\) (5.3)

6 Experimental Evaluation

We compare WIQL with five other algorithms: OPT assumes full knowledge of optimal Whittle Indices, \textbf{AB} [Avrachenkov and Borkar, 2020], \textbf{Fu} [Fu et al., 2019], \textbf{Greedy} chooses the top \(M\) arms with the highest difference in the average observed rewards between actions 1 and 0, \textbf{Random} chooses \(M\) arms uniformly at random.

We first evaluate WIQL on the Circulant Dynamics example, considered in the prior works. We then consider the example in Figure 1, and simulate RMAB instances using the behavioral pattern of the beneficiaries in a call-based program that provides preventive care information for maternal healthcare (detailed in Section 6.2). For each set of experiments, we plot the cumulative reward (averaged over the number of time steps). The results are further averaged over 30 trials, to nullify the effect of randomness in the action-selection policy.

6.1 Numerical Example: Circulant Dynamics

Each arm has four states \(S = \{0, 1, 2, 3\}\) and two actions \(A = \{0, 1\}\). The rewards are \(R^0 = 1, R^1 = R^2 = 0,\) and \(R^3 = -1\) for \(a \in \{0, 1\}\). The transition probabilities for each action \(a \in \{0, 1\}\) are represented as a \(|S| \times |S|\) matrix:

\[
\mathbf{P}^1 = \begin{pmatrix}
0.5 & 0.5 & 0 & 0 \\
0 & 0.5 & 0.5 & 0 \\
0 & 0 & 0.5 & 0.5 \\
0.5 & 0 & 0 & 0.5
\end{pmatrix}, \quad \mathbf{P}^0 = \begin{pmatrix}
0.5 & 0 & 0 & 0.5 \\
0.5 & 0 & 0 & 0 \\
0.5 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.5
\end{pmatrix}
\]

The optimal Whittle Indices are as follows: \(\lambda^*(0) = -0.5, \lambda^*(1) = 0.5, \lambda^*(2) = 1,\) and \(\lambda^*(3) = -1.\)
Figure 2 compares the performance of our proposed method with the benchmarks. We observe that WIQL gradually increases towards the OPT policy, for two sets of experiments—(1) when \( N = 5 \) and \( M = 1 \) and (2) when \( N = 100 \) and \( M = 20 \). We observe that AB converges towards an average reward of zero. This happens because it prioritizes arms that are currently at state 1 over other arms at any other state. Since, the expected reward of taking action 1 at state 1 is \( P(1, 1, 1) \cdot R_1^1 + P(1, 1, 2) \cdot R_2^1 = 0 \), the total average reward also tends to zero. The result obtained by the algorithm Fu is the same as what is shown in Figure 2 of their paper [Fu et al., 2019] where the total average reward converges to a value of 0.08. As expected, Greedy and Random, one being too myopic and the other being too exploratory in nature, are unable to converge to the optimal value.

![Figure 2: Evaluating WIQL on Circulant Dynamics example.](image)

Note that, AB and Fu consider specific hyperparameters for this example. However for the real-world application, that we consider next, it is not clear how the best set of hyperparameters for their algorithms. Thus, we do not compare them for the maternal healthcare application. In the next section, we compare the performance of WIQL algorithm with Greedy, Random and a Myopic policy (defined in 6.2).

### 6.2 Real-world Application: Maternal Healthcare

We now focus on the maternal healthcare problem (detailed in Section 4). We assume that \( R^S_1 = 2, R^P_1 = 1 \), and \( R^L_1 = 0 \). Thus, a high total reward accumulated per week implies that a large number of beneficiaries are at either state \( S \) or \( P \).

We study the data obtained from ARMMAN that contains call-records of enrolled beneficiaries. The data was collected for the experimental study towards building up a robust intervention program (which is the aim of our work). During the experimental study, 1559 out of 3031 beneficiaries received only one intervention (personalized visit by a health worker). Interventions were given to those who were more likely to drop-out of the program. We call this a Myopic intervention scheme. For the simulation, we assume 5000 enrolled beneficiaries and 100 health-care workers. Each health-care worker is responsible for providing personalized care to a set of 50 beneficiaries. Every week, a health-care worker would select a subset of beneficiaries (say 5 or 15) and visit them personally to ensure that they listen to the subsequent automated calls. For simulating the beneficiaries, we assume three categories of arms—(A) highly likely to improve their engagement on receiving an intervention, and deteriorate in absence of an intervention when they are at state \( P \), i.e., \( p^{PS} = 0.8 \) and \( p^{PL} = 0.8 \), (B) medium chance of improvement: \( p^{PS} = 0.4 \) and \( p^{PL} = 0.6 \), and (C) low chance: \( p^{PS} = 0.1 \) and \( p^{PL} = 0.6 \). We assume 10 arms belong to category-A, 10 arms belong to category-B and 30 arms belong to category-C. This assumption helps us determine the efficacy of any learning algorithm; in particular, the most efficient algorithm would quickly learn to intervene the 10 arms of category-A whenever they are at state \( P \). We compare WIQL with greedy, random, and Myopic algorithm. As mentioned earlier, the Myopic algorithm had been a standard way of selecting beneficiaries for intervention—select the beneficiaries with the lowest rate of engagement. We run each simulation for 80 weeks (since a beneficiary is enrolled for about 80 weeks during her pregnancy, childbirth, and up to one year after the childbirth). In Figure 3 we provide results obtained by considering various values of \( M \in \{500, 1500\} \), where a value of \( M \) represents the total number of personalized visits made by 50 health-care workers. The average cumulative reward by WIQL is higher than Greedy, Random and Myopic. Observe that, the convergence of WIQL to a total reward of 5000 is quicker when \( M \) is higher. This is because more sample points with active actions are observed per week.

![Figure 3: Evaluating WIQL on maternal healthcare application.](image)

The empirical results show that WIQL is able to learn which arms should be intervened at which state without any prior knowledge about the transition probabilities, and outperforms benchmark algorithms.