Learning Index Policies for Restless Bandits with Application to Maternal Healthcare

Extended Abstract

Arpita Biswas
Google Research
barpita@google.com

Gaurav Aggarwal
Google Research
gauravaggarwal@google.com

Pradeep Varakantham
Google Research
pvarakantham@google.com

Milind Tambe
Google Research
milindtambe@google.com

ABSTRACT

In many community health settings, it is crucial to have a systematic monitoring and intervention process to ensure that the patients adhere to healthcare programs, such as periodic health checks or taking medications. When these interventions are expensive, they can be provided to only a fixed small fraction of the patients at any period of time. Hence, it is important to carefully choose the beneficiaries who should be provided with interventions and when. We model this scenario as a *restless multi-armed bandit* (RMAB) problem, where each beneficiary is assumed to transition from one state to another depending on the intervention provided to them. In practice, the transition probabilities are unknown a priori, and hence, we propose a mechanism for the problem of balancing the explore-exploit trade-off. Empirically, we find that our proposed mechanism outperforms the baseline intervention scheme maternal healthcare dataset.

KEYWORDS

Reinforcement learning; Multi-armed bandits; Unknown Transition Probabilities

ACM Reference Format:
Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. 2021. Learning Index Policies for Restless Bandits with Application to Maternal Healthcare: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 2 pages.

1 INTRODUCTION

This paper focuses on the learning problem in the restless multi-armed bandit (RMAB) setting [17] with applications to maternal healthcare. Maternal health refers to the health of women during their pregnancy, childbirth, and postnatal period. Although maternal health has received significant attention [13], the number of maternal deaths remains unacceptably high, mainly because of the delay in obtaining adequate care [16]. 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 based in India, called ARMMAN [2], carries out a free call-based program, called mMitra, for spreading preventive care information among pregnant women via automated calls. 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 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 a health-care 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 at a particular time period. Moreover, the extent to which the engagement improves upon intervention varies among individuals and needs to be estimated. Hence, it is important to carefully choose the beneficiaries who should be provided with interventions at a particular time. This is a challenging problem owing to multiple key reasons:

(1) Engagement of the individual beneficiaries is uncertain and changes organically over time
(2) The improvement in the engagement of a beneficiary post-intervention is uncertain
(3) 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
(4) Number of interventions are budgeted and are significantly smaller than the total number of beneficiaries.

2 BACKGROUND

A *restless multi-armed bandit* (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 $\mathcal{S}$, a set of actions $\mathcal{A}$, transition probabilities $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$, and reward function $R : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$. The action set $\mathcal{A}$ of each MDP consists of two

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

actions: an *active* action (1) and a *passive* action (0). At each time step t, an action $A_i(t) \in \mathcal{A}$ is taken on an 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 $X_i(t) \in \mathcal{S}$ 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 beneficiaries to the actions to be taken on each arm. Thus, given a policy π, the action on an arm i is denoted as $A_i^\pi(t)$ where $A_i^\pi(t) = 1$ if i is selected by policy π at time t and 0 otherwise. In a restless bandits problem, the goal is to find the best policy π^* that maximizes the total expected average reward V^π subject to the budget constraint $|A_i^\pi(t)| = M$ for all t.

$$\begin{aligned} &\max_{\pi} \quad \liminf_{t \to \infty} \quad \frac{1}{t} \mathbb{E}\left[\sum_{i \in N} \sum_{h=0}^{t-1} R_i^{X_i(h)}(A_i^\pi(h))\right] \\ &\text{s.t.} \quad \sum_{i \in N} A_i^\pi(t) = M \qquad \text{for all } t = \{1, 2, \ldots\} \end{aligned} \tag{1}$$

This problem is shown to be PSPACE-hard [14]. To deal with the computational hardness, an index-based heuristic policy based on the Lagrangian relaxation of the RMAB problem (1) is proposed by Whittle [17]. According to this method, at each time step t, an index, called *Whittle index*, is computed for each arm using the current state of the arm, the transition probabilities and the reward function of its MDP. Then, top M arms with highest index values are selected for taking active actions.

Due to the uncertainty, sequential nature of decision making, and weak dependency amongst patients through a budget, existing research [1, 4, 9–12] in health interventions has justifiably employed restless multi-armed bandits (RMAB). However, existing research on RMAB problems assumes *a priori* knowledge of the underlying uncertainty model. Note that, it is important to have the knowledge of transition probabilities $\mathcal{P}_i(Z, a, Z')$ to compute the Whittle Indices. Since the transition probabilities are often unknown in most practical scenarios, we focus on the problem of learning the Whittle index while simultaneously selecting a set of best arms depending on the estimated Whittle Indices. There are very few papers that focus on learning Whittle Indices. Fu *et al.* [5] provide a Q-learning method where the Q value $Q(\lambda, s, a)$ is defined based on the Whittle index λ, states, and action. However, they do not provide proof of convergence to optimal. Along similar lines, Avrachenkov and Borkar [3] provide a fundamental change to the Q-value definition with the aim of computing optimal whittle index policy and provides a convergence proof of their algorithm under the assumption that all the arms are homogeneous, that is, they have the same underlying MDP. However, often in real-world problems, the arms (say, patients) have different transition probabilities. Thus, it is important to provide algorithms that learn even when each arm has different transition probabilities.

3 APPLICATION: MATERNAL HEALTHCARE

We consider a *maternal healthcare problem* where a subset of beneficiaries is required to be selected for providing interventions, in a particular week. We model this problem as an RMAB assuming each arm transitions within a three-state MDP–the beneficiaries listening to more than 50% of the information (sent via automated calls) are at state S (self-motivated), those listening to $5 - 50\%$ of the information are at state P (persuadable), and those listening to only $0 - 5\%$ of the information are at state L (lost cause). We assume that a reward of 2 is obtained when a beneficiary is in state S, a reward of 1 is obtained at state P, and a reward of 0 is obtained at state L. Thus, a high total reward accumulated per week implies the desirable outcome that a large number of beneficiaries are at either state S or P. We study the data obtained from the call-based preventive care program carried out by ARMMAN. The data contains call-records of enrolled beneficiaries—how long they listened to the information sent to them, whether a personal intervention was given, and when. Interventions were given to those who were more likely to drop out of the program. This the **Myopic** intervention scheme serves as a benchmark. Another way is to learn the patients' behavior and the effect of interventions over time, which would help taking better future decisions. We believe that such an approach can be used to learn intervention policies in several other domains such as sensor monitoring [6, 8], anti-poaching patrols [15], uplift modeling [7] and many more.

REFERENCES

[1] N. Akbarzadeh and A. Mahajan. 2019. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *2019 IEEE Conference on Decision and Control*. IEEE.

[2] ARMMAN. 2015. Leveraging technology to create scalable solutions empowering mothers and enabling healthy children. https://armman.org/.

[3] Konstantin Avrachenkov and Vivek S Borkar. 2020. Whittle index based Q-learning for restless bandits with average reward. *arXiv preprint arXiv:2004.14427* (2020).

[4] Biswarup Bhattacharya. 2018. Restless bandits visiting villages: A preliminary study on distributing public health services. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–8.

[5] Jing Fu, Yoni Nazarathy, Sarat Moka, and Peter G Taylor. 2019. Towards Q-learning the Whittle Index for Restless Bandits. In *2019 Australian & New Zealand Control Conference (ANZCC)*. IEEE, 249–254.

[6] K.D. Glazebrook, D. Ruiz-Hernandex, and C. Kirkbride. 2006. Some indexable families of restless bandit problems. *Adv. Appl. Probab* (2006), 643–672.

[7] Robin Gubela, Artem Bequé, Stefan Lessmann, and Fabian Gebert. 2019. Conversion uplift in e-commerce: A systematic benchmark of modeling strategies. *International Journal of Information Technology & Decision Making* 18, 03 (2019), 747–791.

[8] F. Iannello, O. Simeone, and U. Spagnolini. 2012. Optimality of myopic scheduling and Whittle indexability for energy harvesting sensors. In *Conference on Information Sciences and Systems (CISS)*. IEEE.

[9] Jackson A Killian, Andrew Perrault, and Milind Tambe. 2021. Beyond "To Act or Not to Act": Fast Lagrangian Approaches to General Multi-Action Restless Bandits. (2021).

[10] Elliot Lee, Mariel S Lavieri, and Michael Volk. 2019. Optimal screening for hepatocellular carcinoma: A restless bandit model. *Manufacturing & Service Operations Management* 21, 1 (2019), 198–212.

[11] Aditya Mate, Jackson Killian, Haifend Xu, Andrew Perrault, and Milind Tambe. 2020. Collapsing Bandits and Their Application to Public Health Interventions. In *Neural Information Processing Systems, NeurIPS*.

[12] Aditya Mate, Andrew Perrault, and Milind Tambe. 2021. Risk-Aware Interventions in Public Health: Planning with Restless Multi-Armed Bandits. (2021).

[13] World Health Organization. 2015. Maternal and Newborn Health. https://www.euro.who.int/en/health-topics/Life-stages/maternal-and-newborn-health/maternal-and-newborn-health.

[14] Christos H Papadimitriou and John N Tsitsiklis. 1994. The complexity of optimal queueing network control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*. IEEE, 318–322.

[15] Y. Qian, C. Zhang, B. Krishnamachari, and B. Tambe. 2016. Restless poachers: Handling exploration- exploitation tradeoffs in security domains. In *International Joint Conference on Autonomous Agents and Multi-Agent Systems, AAMAS*. IFAAMAS.

[16] Sereen Thaddeus and Deborah Maine. 1994. Too far to walk: maternal mortality in context. *Social science & medicine* 38, 8 (1994), 1091–1110.

[17] Peter Whittle. 1988. Restless bandits: Activity allocation in a changing world. *Journal of applied probability* (1988), 287–298.