Towards Zero Shot Learning in Restless Multi-armed Bandits

Extended Abstract

Yunfan Zhao*
Harvard University
yunfanzhao@fas.harvard.edu

Nikhil Behari*
Harvard University
nikhilbehari@g.harvard.edu

Edward Hughes
Google
edwardhughes@google.com

Edwin Zhang
Harvard University
ezhang@g.harvard.edu

Dheeraj Nagaraj
Google
dheerajnagaraj@google.com

Karl Tuyls
Google
karltuyls@google.com

Aparna Taneja
Google
aparnataneja@google.com

Milind Tambe
Harvard University, Google
milind_tambe@harvard.edu

ABSTRACT

Restless multi-arm bandits (RMABs), a class of resource allocation problems with broad application in areas such as healthcare, online advertising, and anti-poaching, have recently been studied from a multi-agent reinforcement learning perspective. Prior RMAB research suffers from several limitations, e.g., it fails to adequately address continuous states, and requires retraining from scratch when arms opt-in and opt-out over time, a common challenge in many real world applications. We propose a neural network-based pre-trained model that has general zero-shot ability on a wide range of previously unseen RMABs.

KEYWORDS

Restless multi-arm bandits; zero shot; pre-trained model

ACM Reference Format:
Yunfan Zhao*, Nikhil Behari*, Edward Hughes, Edwin Zhang, Dheeraj Nagaraj, Karl Tuyls, Aparna Taneja, and Milind Tambe. 2024. Towards Zero Shot Learning in Restless Multi-armed Bandits: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Restless multi-arm bandits (RMABs), a class of resource allocation problems involving multiple agents with a global resource constraint, have recently been studied from a multi-agent reinforcement learning perspective. This has found applications in various scenarios, including resource allocation in multi-channel communication, machine maintenance, and healthcare [1, 8, 11, 14, 17, 21, 22, 24, 28, 29, 33].

The usual RMAB setting considers a fixed number of arms, each associated with a known, fixed MDP with finite state and action spaces; the RMAB chooses K of N arms every round to optimize some long term objective. Even in this setting, the problem has been shown to be PSPACE hard [20]. Several approximation algorithms have been proposed in this setting [7, 27], *particularly when MDP transition probabilities are fully specified,* which are successful in practice. State-of-the-art approaches for *binary action* RMABs commonly provide policies based on the Whittle index [27], an approach that has also been generalized to *multi-action* RMABs[7, 9]. There are also linear programming-based approaches to both *binary and multi-action* RMABs [6, 30–32]. Reinforcement learning (RL) based techniques have also been proposed as state-of-the-art solutions for general *multi-action RMABs* [10].

In this work, we focus on RL-based methods that provide general solutions to binary and multi-action RMABs, without requiring ground truth transition dynamics, or special properties such as indexability as required by other approaches [16, 26]. Unfortunately, several limitations exist in current RMAB solutions, especially for state of the art RL-based solutions, making them challenging or inefficient to deploy in real-world resource allocation problems.

The first limitation arises when dealing with arms that constantly opt-in (also known as streaming RMABs [16]). Existing solutions either require ground truth transition probabilities, which are often unknown in practice, or else require an entirely new model to be trained repeatedly, which can be extremely computationally costly and sample inefficient. For instance, public health programs may model patient intervention deployment as an RMAB problem[3, 4, 13, 18, 19, 25], where new patients (arms in RMABs) arrive asynchronously during intervention deployment[16]. Frequently training models from scratch to account for new patients with unknown transition dynamics may be infeasible, or prohibitively expensive over long time periods, particularly for public health programs that operate with limited resources.

A second limitation occurs for new programs, or existing programs experiencing a slight change in the user base. In these situations, existing approaches do not provide a pretrained RMAB model that can be immediately deployed. In deep learning, pretrained models are the foundation for contemporary, large-scale image and text networks that generalize well across a variety of tasks [2]. For real-world problems modeled with RMABs, establishing a similar pretrained model is essential to reduce the burden of

*Equal Contribution.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

training new RMAB policies from scratch, as well as for transferring knowledge across domains when data is scarce.

The third limitation occurs in handling continuous state *multi-action* RMABs. Continuous state restless bandits have several important applications [5, 12, 23]. However, in field studies, naturally continuous domain state-spaces, such as patient adherence, are often binned into manually crafted discrete state spaces to improve model tractability and scalability [15]. In this process, we may lose crucial information about raw observations, and spend substantial time crafting these discrete state spaces manually.

We propose a pretrained model that enables zero-shot deployment on unseen arms as well as rapid fine-tuning for specific RMAB instances.

2 BACKGROUND

We consider multi-action RMABs with system capacity N, where existing arms have the option to opt-out (that is, the state-action-rewards corresponding to them are disregarded by the model post opt-out), and new, unseen arms can request to opt-in (that is, these arms are considered only post the opt-in time). Such requests will be accepted if and only if the system capacity permits. A vector $\xi_t \in \{0,1\}^N$ represents the opt-in decisions:

$$\xi_{i,t} = \begin{cases} 1 & \text{if arm } i \text{ opts-in at round } t, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that existing arms must opt-in in each round t to remain in the system. For each arm $i \in [N]$, the state space $\mathcal{S}_i$ can be either discrete or continuous, and the action space $\mathcal{A}_i$ is a finite set of discrete actions. Each action $a \in \mathcal{A}_i$ has an associated cost $C_i(a)$, with $C_i(0)$ denoting a no-cost passive action. The reward at a state is given by a function $R_i : \mathcal{S}_i \to \mathbb{R}$. We let $\beta \in [0,1)$ denote a discount factor. Each arm has a unique feature vector $z_i \in \mathbb{R}^m$ that provides useful information about the arm. Notice our model directly utilizes feature information in its policy network, without requiring intermediate steps to extract transition dynamics information from features.

When the state space is discrete, each arm $i \in [N]$ follows a Markov Decision Process $(\mathcal{S}_i, \mathcal{A}_i, C_i, T_i, R_i, \beta, z_i)$, where $T_i : \mathcal{S}_i \times \mathcal{A}_i \times \mathcal{S}_i \to [0,1]$ is a transition matrix representing the probability of transitioning from the current state to the next state given an action. In contrast, when the state space is continuous, each arm $i \in [N]$ follows a Markov Decision Process $(\mathcal{S}_i, \mathcal{A}_i, C_i, \boldsymbol{\Gamma}_i, R_i, \beta, z_i)$, where $\boldsymbol{\Gamma}_i$ is a set of parameters encoding the transition dynamics. For example, in the case that the next state moves according to a Gaussian distribution, $\boldsymbol{\Gamma}_i$ may denote the mean and variance of the Gaussian.

For simplicity, we assume that $\mathcal{S}_i, \mathcal{A}_i, C_i$, and R_i are the same for all arms $i \in [N]$ and omit the subscript i. Note that our algorithms can also be used in the general case where rewards and action costs are different across arms. For ease of notation, we let $\boldsymbol{s} \in \mathbb{R}^N$ denote the state over all arms, and we let $\boldsymbol{A} \in \{0,1\}^{N \times |\mathcal{A}|}$ denote one-hot-encoding of the actions taken over all arms. The agent learns a policy π that maps states $\boldsymbol{s}$ and features $\boldsymbol{z}$ to actions $\boldsymbol{A}$, while satisfying a constraint that the sum cost of actions taken is no greater than a given budget B in every timestep $t \in [H]$, where H is the length of the horizon.

Our goal is to learn an RMAB policy that maximizes the following Bellman equation. The key difficulty in learning such a policy is how to utilize features $\boldsymbol{z}$ and address opt-in decisions $\boldsymbol{\xi}$. These are important research questions not addressed in previous works [10, 16].

$$J(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\xi}) = \max_{\boldsymbol{A}} \left\{ \sum_{i=1}^{N} R(\boldsymbol{s}_i) + \beta \mathbb{E}\left[J\left(\boldsymbol{s}', \boldsymbol{z}, \boldsymbol{\xi}\right) \mid \boldsymbol{s}, \boldsymbol{A} \right] \right\}, \quad (1)$$

$$\text{s.t.} \sum_{i=1}^{N} \sum_{j=1}^{|\mathcal{A}|} \boldsymbol{A}_{ij} c_j \le B \quad \text{and} \quad \sum_{j=1}^{|\mathcal{A}|} \boldsymbol{A}_{ij} = 1 \quad \forall i \in [N],$$

where $c_j \in C$ is the cost of j^{th} action, and $A_{ij} = 1$ if action j is chosen on arm i and $A_{ij} = 0$ otherwise. To learn a policy in multi-action RMAB problems, a scalable approach is to use the Lagrangian relaxation [7, 9, 10]:

$$J\left(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\xi}, \lambda^\star\right) = \min_{\lambda \ge 0} \left(\frac{\lambda B}{1-\beta} + \sum_{i=1}^{N} \max_{j \in |\mathcal{A}|} \left\{ Q_i\left(\boldsymbol{s}_i, a_{ij}, z_i, \xi_i, \lambda\right) \right\} \right), \quad (2)$$

$$\text{s.t. } Q_i\left(\boldsymbol{s}_i, a_{ij}, z_i, \xi_i, \lambda\right) = \xi_i R\left(\boldsymbol{s}_i\right) - \xi_i \lambda c_j + \beta \mathbb{E}\left[Q_i\left(\boldsymbol{s}_i', a_{ij}, z_i, \xi_i, \lambda\right) \mid \pi(\lambda) \right].$$

where Q is the Q-function, a_{ij} is the j^{th} action of arm i, s_i' is the state transitioned to from s_i under action a_{ij}, and $\pi(\lambda)$ is the optimal policy under a given λ. Notice that this relaxation decouples the Q-functions of the arms, and therefore Q_i can be solved independently for a given λ. Computing an appropriate λ is critical in learning a good policy [9, 10].

3 CONTRIBUTION

We propose a pretrained model that has general zero-shot ability on entire sets of unseen arms. The proposed model would allow fine-tuning on specific instances in a more sample-efficient way than training from scratch, and it would accommodate both discrete state setting and challenging continuous state setting with nonlinear reward functions. Additionally, the proposed model would allow agents to learn from each other's experience.

ACKNOWLEDGMENTS

This work was supported by the Harvard Data Science Initiative.

REFERENCES

[1] Saeed Bagheri and Anna Scaglione. 2015. The restless multi-armed bandit formulation of the cognitive compressive sensing problem. *IEEE Transactions on Signal Processing* 63, 5 (2015), 1183–1198.

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[3] J Nell Brownstein, Farah M Chowdhury, Susan L Norris, Tanya Horsley, Leonard Jack Jr, Xuanping Zhang, and Dawn Satterfield. 2007. Effectiveness of community health workers in the care of people with hypertension. *American journal of preventive medicine* 32, 5 (2007), 435–447.

[4] Panayiotis Danassis, Shresth Verma, Jackson A Killian, Aparna Taneja, and Milind Tambe. 2023. Limited Resource Allocation in a Non-Markovian World: The Case of Maternal and Child Healthcare. *arXiv preprint arXiv:2305.12640* (2023).

[5] Fabrice Dusonchet and M-O Hongler. 2003. Continuous-time restless bandit and dynamic scheduling for make-to-stock production. *IEEE Transactions on Robotics and Automation* 19, 6 (2003), 977–990.

[6] Abheek Ghosh, Dheeraj Nagaraj, Manish Jain, and Milind Tambe. 2023. Indexability is Not Enough for Whittle: Improved, Near-Optimal Algorithms for Restless Bandits. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1294–1302.

[7] Jeffrey Thomas Hawkins. 2003. *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[8] David J Hodge and Kevin D Glazebrook. 2015. On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Advances in Applied Probability* 47, 3 (2015), 652–667.

[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. In *AAMAS*. AAMAS, UK, 710–718.

[10] Jackson A Killian, Lily Xu, Arpita Biswas, and Milind Tambe. 2022. Restless and uncertain: Robust policies for restless bandits via deep multi-agent reinforcement learning. In *Uncertainty in Artificial Intelligence*. PMLR, 990–1000.

[11] 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.

[12] Claude Lefévre. 1981. Optimal control of a birth and death epidemic process. *Operations Research* 29, 5 (1981), 971–982.

[13] Bernd Löwe, Jürgen Unützer, Christopher M Callahan, Anthony J Perkins, and Kurt Kroenke. 2004. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical care* (2004), 1194–1201.

[14] Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. 2020. Collapsing Bandits and Their Application to Public Health Intervention. *Advances in Neural Information Processing Systems* 33 (2020), 15639–15650.

[15] Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. 2022. Field Study in Deploying Restless Multi-Armed Bandits: Assisting Non-profits in Improving Maternal and Child Health. *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (06 2022), 12017–12025.

[16] Aditya S Mate, Arpita Biswas, Christoph Siebenbrunner, Susobhan Ghosh, and Milind Tambe. 2022. Efficient Algorithms for Finite Horizon and Streaming Restless Multi-Armed Bandit Problems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 880–888.

[17] Navikkumar Modi, Philippe Mary, and Christophe Moy. 2019. Transfer restless multi-armed bandit policy for energy-efficient heterogeneous cellular network. *EURASIP Journal on Advances in Signal Processing* 2019, 1 (2019), 1–19.

[18] Patrick M Newman, Molly F Franke, Jafet Arrieta, Hector Carrasco, Patrick Elliott, Hugo Flores, Alexandra Friedman, Sophia Graham, Luis Martinez, Lindsay Palazuelos, et al. 2018. Community health workers improve disease control and medication adherence among patients with diabetes and/or hypertension in Chiapas, Mexico: an observational stepped-wedge study. *BMJ global health* 3, 1 (2018), e000566.

[19] Jane Rahedi Ong'ang'o, Christina Mwachari, Hillary Kipruto, and Simon Karanja. 2014. The effects on tuberculosis treatment adherence from utilising community health workers: a comparison of selected rural and urban settings in Kenya. *PLoS One* 9, 2 (2014), e88937.

[20] Christos H Papadimitriou and John N Tsitsiklis. 1999. The Complexity of Optimal Queuing Network Control. *Mathematics of Operations Research* 24, 2 (1999), 293–305.

[21] Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. 2016. Restless Poachers: Handling Exploration-Exploitation Tradeoffs in Security Domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 123–131.

[22] Diego Ruiz-Hernández, Jesús M Pinar-Pérez, and David Delgado-Gómez. 2020. Multi-machine preventive maintenance scheduling with imperfect interventions: A restless bandit approach. *Computers & Operations Research* 119 (2020), 104927.

[23] Amit Sinha and Aditya Mahajan. 2022. Robustness of Whittle index policy to model approximation. *Available at SSRN 4064507* (2022).

[24] Vishrant Tripathi and Eytan Modiano. 2019. A Whittle Index Approach to Minimizing Functions of Age of Information. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Allerton, Allerton, 1160–1167.

[25] Shresth Verma, Aditya Mate, Kai Wang, Neha Madhiwalla, Aparna Hegde, Aparna Taneja, and Milind Tambe. 2023. Restless Multi-Armed Bandits for Maternal and Child Health: Results from Decision-Focused Learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1312–1320.

[26] Kai Wang, Shresth Verma, Aditya Mate, Sanket Shah, Aparna Taneja, Neha Madhiwalla, Aparna Hegde, and Milind Tambe. 2023. Scalable decision-focused learning in restless multi-armed bandits with application to maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 12138–12146.

[27] Peter Whittle. 1988. Restless bandits: Activity allocation in a changing world. *Journal of applied probability* 25, A (1988), 287–298.

[28] Guojun Xiong and Jian Li. 2023. Finite-Time Analysis of Whittle Index based Q-Learning for Restless Multi-Armed Bandits with Neural Network Function Approximation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[29] Zhe Yu, Yunjian Xu, and Lang Tong. 2018. Deadline scheduling as restless bandits. *IEEE Trans. Automat. Control* 63, 8 (2018), 2343–2358.

[30] Gabriel Zayas-Caban, Stefanus Jasin, and Guihua Wang. 2019. An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits. *Advances in Applied Probability* 51, 3 (2019), 745–772.

[31] Xiangyu Zhang and Peter I Frazier. 2021. Restless bandits with many arms: Beating the central limit theorem. *arXiv preprint arXiv:2107.11911* (2021).

[32] Xiangyu Zhang and Peter I Frazier. 2022. Near-optimality for infinite-horizon restless bandits with many arms. *arXiv preprint arXiv:2203.15853* (2022).

[33] Qing Zhao, Bhaskar Krishnamachari, and Keqin Liu. 2008. On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance. *IEEE Transactions on Wireless Communications* 7, 12 (2008), 5431–5440.