Time-Constrained Restless Multi-Armed Bandits with Applications to City Service Scheduling

Extended Abstract

Yi Mao
The Ohio State University
Columbus, United States
mao.496@osu.edu

Andrew Perrault
The Ohio State University
Columbus, United States
perrault.17@osu.edu

ABSTRACT

Municipalities maintain critical infrastructure through inspections, both proactive and in response to complaints. For example, the Chicago Department of Public Health (CDPH) periodically inspects 7000 food establishments to maintain the safety of food bought, sold, or prepared for public consumption. Restless multi-armed bandits (RMABs) appear to be a useful tool for optimizing the scheduling of inspections, as the schedule aims to keep as many establishments in the "passing" state subject to an action limit per period. However, a key challenge arises: satisfying timing and frequency constraints. Municipal agencies often provide an inspection window to each establishment (e.g., a two-week period where an inspection will occur) and guarantee the minimum frequency of inspection (e.g., once per year). We develop an extension to Whittle index-based systems for RMABs that can guarantee both action window constraints and minimum frequencies. Briefly, we take a Whittle index-based view, enforcing window constraints by integrating the window structure into individual MDPs, and frequency constraints through a higher-level scheduling algorithm that aims to maximize the Whittle index. We demonstrate our methods' performance and scalability in experiments using synthetic and real data (with 7000 establishments inspected per year). Not only does our approach enforce constraints more effectively than naive methods, but it also achieves higher rewards, up to 20%.

KEYWORDS

Restless Multi-Arm Bandits, Scheduling with Constraints

ACM Reference Format:
Yi Mao and Andrew Perrault. 2024. Time-Constrained Restless Multi-Armed Bandits with Applications to City Service Scheduling: 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-armed bandits (RMABs) [12] describe a sequential decision problem where an agent aims to manage a large population of Markov decision processes (MDPs) that are independent except for a shared action budget at each time step. We argue that RMABs are a valuable framework to study for two reasons. First, they are a practical way of introducing a sequential component to many real-world, large-scale optimization problems that are naturally sequential without adding much complexity. Second, RMABs can often be solved approximately optimally in a computationally efficient manner through the use of the Whittle index heuristic [12] if RMABs are under a technical condition known as indexability. As a result, RMABs have attracted wide interest over the past several decades in a large variety of resource allocation tasks, including wireless networking [5], machine maintenance [1, 4], and planning health interventions [3, 7, 9].

In practical applications of RMABs, it is common to place constraints on the timing and frequency of arm pulls. As a city service example for this work, the Chicago Department of Public Health (CDPH) provides establishments with an inspection window: they state a particular time period during which the routine inspection will occur. This window makes the inspection less disruptive to the establishment. A similar constraint is used by a field study of applying bandits in the child health setting [9], where each beneficiary receives at most one call every fixed number of weeks. In addition, CDPH guarantees at least one inspection per year, per establishment, providing a baseline level of service. In this paper, we develop methods for integrating action constraints into RMABs.

2 PROBLEM STATEMENT AND SOLUTIONS

2.1 Motivation Inspection RMAB

Motivated by the food establishment setting, we define a model RMAB with action constraints. Due to unobserved binary states(good or bad), each food establishment is described as a partially observed Markov decision process (POMDP) [2], which we can rewrite as a fully observed belief state MDP. Such an RMAB can be viewed as a collapsing bandit [8] or a resetting bandit [6], both with indexability guarantees. For each underlying MDP, we have binary state passive transitions $P_i^{(0)}$ and active transitions $P_i^{(1)}$. Converting this POMDP to a belief-state MDP yields a set of belief states that are reachable from the passing state $b_1 = [0, 1]$ (as a column vector), i.e., $(P_i^{(0)})^t b_1$, where t is any non-negative integer. In practice, the number of belief states of one MDP depends on the rate of MDP mixing.

2.2 Action Windows and MDP Encoding

Action windows are introduced here as an exemplar for the family constraints where the constraint can be directly encoded into the RMAB structure, then we can apply whatever existing state-of-the-art algorithms directly. In the food inspection domains in our work,

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).

		Budget					
		380	400	416	430	450	500
2*IP	Coverage	**4327.40±26.02**	**4507.40±26.02**	**4650.20±25.55**	4775.60±24.76	4939.10±21.59	5000.00 ± 0.00
	Reward	**40534.73±38.30**	**40751.29±38.28**	**40919.68±37.64**	41065.62±36.41	41253.96±37.52	41324.07 ± 58.26
2*TCB	Coverage	4091.00±32.96	4254.40±34.84	4385.40±33.91	4502.60±35.33	4654.70±32.60	4889.10 ± 25.36
	Reward	40323.12±30.00	40515.14±28.15	40667.54±29.92	40800.59±30.54	40970.34±40.13	41242.30 ± 68.88
2*IP	Coverage	-	-	-	**5000.00 ± 0.00**	**5000.00 ± 0.00**	**5000.00 ± 0.00**
(Equality)	Reward	-	-	-	**41186.96±40.37**	**41309.57±33.44**	**41331.99 ± 34.02**

Table 1: Coverage (number of arms pulled in a year) and reward, tested on 5000 arms with a variable budget over 12 timesteps. IP lookahead substantially increases coverage and slightly increases reward relative to TCB. IP (Equality) adds an equality constraint that states that each arm must be pulled exactly once per year. Once the budget is high enough to make this problem feasible, coverage is increased to full, and reward is not decreased by enforcing this constraint.

		Number of Arms				
Policy	Budget	10	50	100	1000	5000
4*IP	1%	-	-	**2.5%**	**2.4%**	**2.6%**
	5%	-	**9.6%**	**12.1%**	**12.2%**	**12.3%**
	10%	**17.8%**	**19.0%**	**20.4%**	**20.1%**	**20.2%**
	20%	**19.5%**	19.1%	**19.6%**	**19.4%**	**19.4%**
4*TCB	1%	-	-	2.4%	**2.4%**	2.4%
	5%	-	9.4%	11.9%	12.0%	12.1%
	10%	16.0%	18.4%	19.4%	**20.1%**	20.1%
	20%	19.3%	**19.5%**	19.4%	19.3%	19.3%
4*RFP	1%	-	-	1.0%	0.9%	1.1%
	5%	-	8.4%	11.0%	11.1%	11.0%
	10%	15.4%	17.7%	18.6%	19.5%	19.6%
	20%	19.0%	19.1%	19.0%	18.9%	19.1%
4*WIP	1%	-	-	0.3%	0.3%	0.3%
	5%	-	1.1%	1.4%	1.5%	1.5%
	10%	2.0%	2.5%	2.5%	2.6%	2.6%
	20%	3.7%	3.7%	3.7%	3.6%	3.6%

Table 2: Percent improved reward vs. RP in synthetic data. RFP and WIP do not explicitly model constraints and thus achieve minimal improvement over RP. TCB and IP yield dramatically increased rewards and larger increases with more arms and budget, with IP outperforming TCB.

we add 2 pieces of information to the states in addition to the belief $b \in [0, 1]$, then modify the transitions to remove impacts of actions out of the window.

- t: the current timestamp. It will always increase by one after one passive or active transition.
- m: a counter for the number of inspections remaining in the window. An active transition decrements m by 1.

This encoding increases the number of states in the MDP by a factor of $O(LM)$, where L is the number of all timestamps needed to track and M is the total number of allowed actions during the window.

2.3 Whittle Index and Lookahead

For RMBAs, the most state-of-the-art solution is the Whittle Index under a technical condition called *indexability*.

We are not aware of an existing class of indexable RMABs that includes the action window MDPs with counters that we define in this section. We empirically check for indexability by tracking the set of passive states as the subsidy changes and find no violations.

Though editing the MDPs enforces the maximum action limits, it is not possible to guarantee the minimums. Therefore, we replace the Whittle Index heuristic with a sequential planning that aims to maximize the sum of Whittle Indices of pulled arms over a lookahead window via an Integer Programming technique. We seek to maximize $\sum_{i=1}^{N}\sum_{t=1}^{T} a_{i,t} w_{i,t}$, subject to some constraints on $a_{i,t}$(e.g., $\sum_{t=1}^{T} a_{i,t} = 1$), then minimum or exact number of actions constraints are fulfilled. This lookahead problem is reducible into a variant of weighted b-matching problem [11], which can be solved in polynomial time $O(|V|^2 \max_v b(v))$ [10].

3 EXPERIMENTS AND RESULTS

We compare our policies with other 3 policies: **RP** is the random policy for the baseline; **RFP** is the risk-first policy which prioritizes arms with the worst beliefs; **WIP** is the Whittle Index Policy without our MDP encoding; **TCB** is the time-constrained RMAB polices without minimum constraints; **IP** is the integer programming policy to fulfill minimum action constraints. We generate synthetic transitions and windows for all MDPs and run in 60 timestamps. The total reward improvement compared with RP in percentage is shown in Table 2, and coverage constraints are checked in Table 1. The TCB and IP can improve the total rewards by as much as 20%. What's more, IP lookahead substantially increases coverage without sacrificing rewards. Once the budget is high enough to make the problem feasible, full coverage is satisfied.

4 CONCLUSION

To the best of our knowledge, we present the first RMAB study to optimize scheduling problems under timing and frequency constraints. Synthetic data results suggest that our methods for explicitly modeling constraints are critical for RMABs to have an impact in this setting. We hope our work paves the way for applying RMABs to other critical infrastructure maintenance and public service problems under constraints.

REFERENCES

[1] Abderrahmane Abbou and Viliam Makis. 2019. Group maintenance: A restless bandits approach. *INFORMS Journal on Computing* 31, 4 (2019), 719–731.

[2] Mauricio Araya, Olivier Buffet, Vincent Thomas, and Françcois Charpillet. 2010. A POMDP Extension with Belief-dependent Rewards. In *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (Eds.), Vol. 23. Curran Associates, Inc.

[3] 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.

[4] Pedro Cesar Lopes Gerum, Ayca Altay, and Melike Baykal-Gürsoy. 2019. Data-driven predictive maintenance scheduling policies for railways. *Transportation Research Part C: Emerging Technologies* 107 (2019), 137–154.

[5] Kesav Kaza, Varun Mehta, Rahul Meshram, and S. N. Merchant. 2018. Restless bandits with cumulative feedback: Applications in wireless networks. In *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. 1–6.

[6] Ali Al Khansa, Raphael Visoz, Yezekael Hayel, and Samson Lasaulce. 2021. Resource allocation for multi-source multi-relay wireless networks: A multi-armed bandit approach. In *Ubiquitous Networking: 7th International Symposium, UNet 2021, Virtual Event, May 19–22, 2021, Revised Selected Papers 7*. Springer, 62–75.

[7] 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.

[8] Aditya Mate, Jackson A. Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. 2020. Collapsing Bandits and Their Application to Public Health Interventions. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

[9] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12017–12025.

[10] Williamk Pulleyblank. 1973. *Facets of I-matching polyhedra*. Ph.D. Dissertation.

[11] Alexander Schrijver et al. 2003. *Combinatorial optimization: polyhedra and efficiency*. Vol. 24. Springer.

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