Fairness of Exposure in Online Restless Multi-armed Bandits

Extended Abstract

Archit Sood
Indian Institute of Technology Ropar
Rupnagar, India
2020mcb1230@iitrpr.ac.in

Shweta Jain
Indian Institute of Technology Ropar
Rupnagar, India
shwetajain@iitrpr.ac.in

Sujit Gujar
International Institute of Information Technology Hyderabad
Hyderabad, India
sujit.gujar@iiit.ac.in

ABSTRACT

Restless multi-armed bandits (RMABs) generalize the multi-armed bandits where each arm exhibits Markovian behavior and transitions according to their transition dynamics. Solutions to RMAB exist for both offline and online cases. However, they do not consider the distribution of pulls among the arms. Studies have shown that optimal policies lead to unfairness, where some arms are not exposed enough. Existing works in fairness in RMABs focus heavily on the offline case, which diminishes their application in real-world scenarios where the environment is largely unknown. In the online scenario, we propose the first fair RMAB framework, where each arm receives pulls in proportion to its merit. We define the merit of an arm as a function of its stationary reward distribution. We prove that our algorithm achieves sublinear fairness regret in the single pull case $O(\sqrt{T}\ln T)$, with T being the total number of episodes. Empirically, we show that our algorithm performs well in the multi-pull scenario as well.

KEYWORDS

Restless bandits; Online learning; Fairness

ACM Reference Format:
Archit Sood, Shweta Jain, and Sujit Gujar. 2024. Fairness of Exposure in Online 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-Armed Bandits (RMABs) are a class of Multi-armed Bandits where each arm has a Markov Decision Process (MDP) associated with it. Each arm has its own states, actions, transition dynamics, and reward functions. The arms transition from one state to the next state, irrespective of whether they are pulled or not. It is this *restless* nature of the arms that makes RMABs applicable to many domains such as network scheduling [16], anti-poaching [19], healthcare [14], etc. Recently, a lot of works have been using restless bandits to model preventive interventions in public healthcare scenarios [3, 5, 9, 10, 13] where an arm is modeled as the patient.

Multiple works have devised algorithms to find optimal policies in RMABs. This includes both the *offline* setting, where the transition probabilities of each arm's MDP are known [1, 24, 25], and the *online setting*, in which the transition probabilities are unknown [2–4, 6, 7, 17, 22]. However, all these approaches focus only on finding the optimal policy – leading to some arms being completely ignored [18]. As in our running example where arms model patients, this represents a major problem: the optimal policies would focus only on patients who require the most interventions and ignore the patients who rarely need interventions. However, in public healthcare, it becomes important to focus on all kinds of patients so as to provide unbiased healthcare to society. Current work on fairness in RMABs includes [5, 12, 15, 18]; these works assume that the transition probabilities are known beforehand and construct their policies based on this assumption. To our knowledge, only Li and Varakantham [11] explore fairness in online RMABs; their fairness notion ensures that each arm is pulled at least once every fixed time period. We propose that arms should be pulled in proportion to their *merit*, which is defined as the difference at steady state when we always pull the arm compared to when we never pull the arm. We call our notion of fairness as *Merit Fair* which is better than the existing notions in online RMABs [11] because unlike the fairness notion of Li and Varakantham [11] which simply classifies arms as optimal and sub-optimal and then accordingly provides fairness, Merit Fair instead pulls arms in proportion to their merit, and ensures that sub-optimal arms with high merit and sub-optimal arms with low merits receive different levels of exposure.

In this paper [21], for theoretical analysis, we primarily focus on single pull settings for the following reasons. Meritocratic fairness [23] has been designed for pulling a single arm at each round. It is unclear how such merit-based fairness can be extended to multiple pulls. However, our algorithm can be extended to multiple pulls and we study its efficacy empirically. To the best of our knowledge, we are the first one to extend the Fairness of Exposure [23] notion to Restless bandits with theoretical guarantees.

2 PRELIMINARIES

An RMAB problem is defined by a set of N independent arms. Each arm $i \in [N]$ is characterized by a Markov Decision Process (MDP) given by $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P_i)$ with $\mathcal{S}, \mathcal{A}$, and $\mathcal{R} : \mathcal{S} \to \mathbb{R}$ denoting the state space, action space, and reward function respectively. $P_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probablity matrix for arm i. In the traditional RMAB setting, each arm differs only by their transition matrix P_i. The states are assumed to be fully observable. The action the decision-maker takes is governed by a policy π. The total number of episodes is T, where a policy π^t is fixed for $t \leq T$, and is run for H timesteps, where H is the time horizon of an episode. For each timestep $h \in H$ in episode t, the decision-maker has to select $K \leq N$ arms according to π^t, where K is the budget.

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).

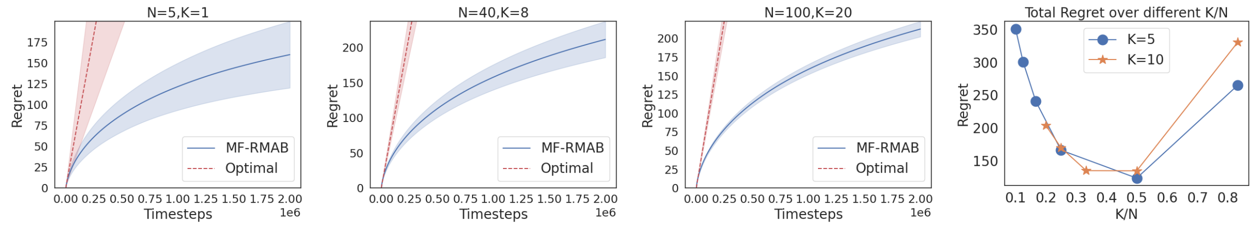

Figure 1: The first three plots show Regret vs. Time for different K and N settings on CPAP dataset. The last plot shows the Regret with different K/N values for $T \times H = 2 \times 10^6$ timesteps.

In our setting, we assume $\mathcal{S} := \{0, 1\}$, where 0 denotes *bad* state and 1 denotes *good* state. There are two possible actions, i.e. $\mathcal{A} := \{1, 0\}$, indicating whether an arm is pulled or not respectively. The arm receive a reward of 1 for being in the good state and 0 otherwise. Let us denote the true transition matrix for an arm i as P_i^*. We assume that P_i^*'s are *non-degenerate*, i.e., there exists an $\epsilon > 0$ such that $\epsilon \leq P_i^*(s, a, s') \leq 1 - \epsilon \quad \forall i \in [N], a \in \mathcal{A}, s, s \in \mathcal{S}'$. Let μ_i^* denote the true reward of arm i, which we define formally later. Along a similar line to Wang et al. [23], let us define the Optimal Fair Policy as $Pr^*(K)$, where $Pr_i^*(K)$ is the probability that arm i is among the K chosen out of the N total arms. Observe that $Pr_i^*(N) = 1$ and that $Pr_i^*(1) = \pi_i^*$, where π^* is the probability distribution of being chosen over the arms. Let $g(\cdot)$ be a non-decreasing merit function that maps the reward of the arm to a positive value. Then, for the optimal fair policy, we have $\frac{Pr_i^*(K)}{g(\mu_i^*)} = \frac{Pr_j^*(K)}{g(\mu_j^*)} \quad \forall i, j \in [N]$. Let $\pi = \{\pi^t\}_{t=1}^T$ be the policy learnt by our algorithm with π^t being the employed policy at episode t. We define *Fairness Regret* FR^T as the difference between the optimal fair policy π^* and our policy π up to episode T. Mathematically, $FR^T = \sum_{t=1}^T \sum_{i \in [N]} \left|\pi_i^* - \pi_i^t\right|$. Here, π_i^t denotes the probability of pulling an arm i in episode t.

3 PROPOSED SOLUTION

We first define a reward that is based on steady state and is indicative of how much intervention an arm requires. Consider the policy discussed by Herlihy et al. [5] where each arm is pulled with some fixed probability p_i, i.e., $\pi_{PF} : \{i \mid i \in [N]\} \rightarrow [1 - p_i,\ p_i]^N$. Let us denote $f(P_i, p_i)$ to be the steady state probability of arm i being in state 1, when followed a policy π_{PF}. At steady state, we should have $f(P_i, p_i)[(1-p_i)P_i(1, 0, 1)+p_i P_i(1, 1, 1)] + (1-f(P_i, p_i))[(1-p_i)P_i(0, 0, 1) + p_i P_i(0, 1, 1)] = f(P_i, p_i)$. The reward of an arm can be naturally defined as: $\mu_i = f(P_i, 1) - f(P_i, 0)$ which represents the benefit of pulling an arm in the long run as compared to the loss the algorithm would have incurred if it had not pulled the arm.

As we are in an online setting, we also need to estimate the true transition matrices P_i^*'s. We use the Upper Confidence Bound (UCB) approach which maintains an optimistic bound on the true transition matrix corresponding to each state-action-state [22]. Let $N_i^t(s, a, s')$ be the number of times (s, a, s') transition has been observed for arm i by episode t. Further, we define $N_i^t(s, a) = \sum_{s'} N_i^t(s, a, s')$. Then at episode t, we estimate the true transition matrix $P_i^*(s, a, s')$ with empirical mean $\hat{P}_i^t(s, a, s') := \frac{N_i^t(s,a,s')}{N_i^t(s,a)}$ and confidence radius $d_i^t(s, a) := \sqrt{\frac{2|\mathcal{S}|\ln(2|\mathcal{S}||\mathcal{A}|N\frac{t^4}{\delta})}{\max\{1, N_i^t(s,a)\}}}$ where $\delta > 0$ is a user defined constant. We can now define the ball B^t of possible values of P^* as $B^t = \{P \mid \|P_i(s, a, \cdot) - \hat{P}_i^t(s, a, \cdot)\|_1 \leq d_i^t(s, a)\ \forall i, s, a\}$. In particular, B_i^t is the ball of possible values of P_i^* at episode t for some particular arm i. It can be proven that P^* belongs to this ball with high probability [22].

MF-RMAB calculates the estimated reward of each arm i as $\mu_i^t = f(P_i^{+,t}, 1) - f(P_i^{+,t}, 0)$ for an episode t. Then the probability distribution over arms being chosen π^t is given by $\pi_i^t = \frac{g(\mu_i^t)}{\sum_j g(\mu_j^t)}$. MF-RMAB then samples K arms without replacement from π^t. We show the following theoretical result:

THEOREM 3.1. *MF-RMAB incurs $O(\sqrt{T \ln T})$ fairness regret for sufficiently large T.*

4 EXPERIMENTAL RESULTS

The dataset used for the experiments is the Markov model of CPAP treatment given by Kang et al. [8] and adapt their three-state model into two states in a similar fashion to [5, 12]. We compare MF-RMAB with an "Optimal" baseline, where in each episode t, Optimal policy pulls the arms with the K highest values for μ_i^t. We use $\delta = 0.01$ and set the merit function $g(\mu) = e^{c\mu}$. We set $c = 3$ and in line with our non-degeneracy assumption on P^*, we clip the transition probabilities in the range $[\epsilon, 1 - \epsilon]$ with $\epsilon = 0.01$. The results are averaged over 30 independent runs with different seed values. We run the experiments for T=10k episodes and H=200 timesteps per episode for a total of $T \times H = 2 \times 10^6$ timesteps.For $K > 1$ cases, we use the same definition of fairness regret as for $K = 1$ cases. The source code is available on Github [20].

The first three plots of Figure 1 show the various trends of fairness regret across different values of N and K. We can see that MF-RMAB incurs a sublinear regret, while Optimal is unable to learn a fair policy and exhibits linear regret. The rightmost plot of Figure 1 shows the variation of total regret FR^T, T = 10k over increasing $\frac{K}{N}$ ratio. We can observe that the minima is around $\frac{K}{N} \approx 0.5$. Therefore, we conclude that increasing K does not necessarily help in learning the transition probabilities faster, and can end up increasing the regret instead.

ACKNOWLEDGMENTS

The author Shweta Jain would like to acknowledge the DST grant MTR/2022/000818 for providing the support to carry out this work.

REFERENCES

[1] Nima Akbarzadeh and Aditya Mahajan. 2019. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *2019 IEEE 58th conference on decision and control (CDC)*. IEEE, 7294–7300.

[2] Konstantin E Avrachenkov and Vivek S Borkar. 2022. Whittle index based Q-learning for restless bandits with average reward. *Automatica* 139 (2022), 110186.

[3] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. 2021. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. *arXiv preprint arXiv:2105.07965* (2021).

[4] 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.

[5] Christine Herlihy, Aviva Prins, Aravind Srinivasan, and John P Dickerson. 2023. Planning to fairly allocate: Probabilistic fairness in the restless bandit setting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 732–740.

[6] Young Hun Jung, Marc Abeille, and Ambuj Tewari. 2019. Thompson sampling in non-episodic restless bandits. *arXiv preprint arXiv:1910.05654* (2019).

[7] Young Hun Jung and Ambuj Tewari. 2019. Regret bounds for thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems* 32 (2019).

[8] Yuncheol Kang, Amy M Sawyer, Paul M Griffin, and Vittaldas V Prabhu. 2016. Modelling adherence behaviour for the treatment of obstructive sleep apnoea. *European journal of operational research* 249, 3 (2016), 1005–1013.

[9] Jackson A Killian, Manish Jain, Yugang Jia, Jonathan Amar, Erich Huang, and Milind Tambe. 2023. Equitable Restless Multi-Armed Bandits: A General Framework Inspired By Digital Health. *arXiv preprint arXiv:2308.09726* (2023).

[10] Jackson A Killian, Arshika Lalan, Aditya Mate, Manish Jain, Aparna Taneja, and Milind Tambe. 2023. Adherence Bandits. (2023).

[11] Dexun Li and Pradeep Varakantham. 2022. Efficient resource allocation with fairness constraints in restless multi-armed bandits. In *Uncertainty in Artificial Intelligence*. PMLR, 1158–1167.

[12] Dexun Li and Pradeep Varakantham. 2023. Avoiding Starvation of Arms in Restless Multi-Armed Bandits. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1303–1311.

[13] 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.

[14] 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.

[15] Aditya Mate, Andrew Perrault, and Milind Tambe. 2021. Risk-Aware Interventions in Public Health: Planning with Restless Multi-Armed Bandits.. In *AAMAS*. 880–888.

[16] 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 (2019), 1–19.

[17] Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. 2012. Regret bounds for restless markov bandits. In *International conference on algorithmic learning theory*. Springer, 214–228.

[18] Aviva Prins, Aditya Mate, Jackson A Killian, Rediet Abebe, and Milind Tambe. 2020. Incorporating Healthcare Motivated Constraints in Restless Bandit Based Resource Allocation. *preprint* (2020).

[19] 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*. 123–131.

[20] Archit Sood. 2024. *MF-RMAB*. https://github.com/rchiso/MF-RMAB

[21] Archit Sood, Shweta Jain, and Sujit Gujar. 2024. Fairness of Exposure in Online Restless Multi-armed Bandits. arXiv:2402.06348 [cs.LG]

[22] Kai Wang, Lily Xu, Aparna Taneja, and Milind Tambe. 2023. Optimistic whittle index policy: Online learning for restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10131–10139.

[23] Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. 2021. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*. PMLR, 10686–10696.

[24] Richard R Weber and Gideon Weiss. 1990. On an index policy for restless bandits. *Journal of applied probability* 27, 3 (1990), 637–648.

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