Incorporating Healthcare Motivated Constraints in Restless Bandit Based Resource Allocation

Aviva Prins
University of Maryland, College Park
aviva@cs.umd.edu

Aditya Mate
Harvard University
aditya_mate@g.harvard.edu

Jackson A. Killian
Harvard University
jkillian@g.harvard.edu

Rediet Abebe
Harvard Society of Fellows
rta36@cornell.edu

Milind Tambe
Harvard Center for Research on Computation and Society
milind_tambe@harvard.edu

Abstract

As reinforcement learning plays an increasingly important role in healthcare, there is a pressing need to identify mechanisms to incorporate practitioner expertise. One notable case is in improving tuberculosis drug adherence, where a health worker must simultaneously monitor and provide services to many patients. We find that—without considering domain expertise—state-of-the-art restless multi-armed bandit algorithms allocate all resources to a small number of patients, neglecting most of the population. To avoid this undesirable behavior, we propose a human-in-the-loop model, where constraints are imposed by domain experts to improve equitability of resource allocations. Our framework enforces these constraints on the distribution of actions without significant loss of utility on simulations from real-world data.

1 Introduction

Designing automated decision-making pipelines that provide domain experts with autonomy in healthcare contexts is becoming increasingly important. We consider a problem in which a health worker monitors a set of patients and delivers interventions to improve their health. Depending on the care context, such problems can be highly resource constrained, for instance due to a high load of patients per health worker. A key application is that of tuberculosis (TB) care in India, where it is crucial that TB patients adhere to their daily antibiotic regimens to avoid spreading the disease and/or developing multi-drug resistance. It is challenging for patients to adhere to TB treatment plans, which last several months; patient adherence decreases as their symptoms subside. It is therefore crucial for health workers to provide frequent care support to patients via phone calls.

In recent years, there has been a growing interest in the use of algorithms for improving medication adherence and related problems (See for example Baio et al. [2011]). These algorithms are effective in theory, but in practice they can result in undesirable behavior (Epstein and Robertson [2015]). For instance, Mate et al. [2020] present an algorithm for computing policies which maximize patient adherence based on the available daily call budget and the patient's likelihood to respond to an intervention. However, since the approach seeks to maximize the adherence of the cohort as a whole, it may exhibit behaviors that are undesirable in a care context, such as never providing interventions to certain patients.

Challenges of Real-World Reinforcement Learning Workshop, 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Additionally, there are multiple reasons why health workers may ignore an automated tool's recommendations. Practitioners tend to ignore opaque recommender tools (Yeomans et al. [2019], DeMichele et al. [2019]). They may be reluctant to use algorithms that suffer from a lack of agency (Lim and O'Connor [1995]). Users may sacrifice accuracy in favor of gaining some control over the output (Dietvorst et al. [2015]). Finally, the model may not have a mechanism to define or prevent undesirable behavior. If the health worker perceives the algorithm to be wrong, they are unlikely to uphold its recommendations (De-Arteaga et al. [2020], Dietvorst et al. [2015]).

We address this shortcoming in Restless Bandits by introducing and examining a human-in-the-loop framework in which the health worker can impose call-frequency constraints. When domain experts find issues such as inequitable allocation that disregard certain populations, they may intervene by using simple constraint-based heuristics or defaulting to standard policies such as round-robin. Such approaches are widely-adopted in human-in-the-loop settings since they are easy to explain and implement. Despite their wide-spread use, however, little is known about the impact of such policies on the utility and fairness of the resource being allocated. Herein, we investigate such human-in-the-loop models and analyze their utility and fairness consequences.

We evaluate our new framework on both synthetic and real-world data collected during a TB treatment program in India. Through experiments, we show that constraint-based human-in-the-loop models do not come at a significant loss of relative reward, but are able to enforce equity and other desirable characteristics on the distribution of resources. Our work serves as an important bridge between theory and practice. This work has implications for practice and decision-support pipelines in public health. By exploring how heuristics that may be deployed by domain experts impact allocations, we also open up a new line of research inquiry on human-machine interaction in restless multi-armed bandits. We close with a discussion about the societal impact of human-in-the-loop models and these new theoretical and computational directions.

2 Problem Description

Our work is motivated by a TB medication adherence program in India, which aims to increase patient compliance in TB treatment plans (Killian et al. [2019]). Though symptoms of TB may dissipate at the beginning of treatment, patients must fulfill the full regimen for immunity. Prematurely terminating the prescription course may contribute to an antibiotic resistant form of TB.

In the program, health workers interact with a subset k out of N patients every day, $k \ll N$, via counseling phone calls. In each call, the health worker learns whether the patient has been adhering to their treatment plan and provides an intervention that encourages future adherence.

We model each patient as a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, P, r)$. For a given time $t \in \mathbb{N}_{\leq T}$ and arm $i \in \mathbb{N}_{\leq N}$, available *actions* are $a_t^i \in \mathcal{A} = \{0, 1\}$. If $a_t^i = 1$, patient i is called and their *state* $s_t^i \in \mathcal{S} = \{0, 1\}$ is learned. If $s_t^i = 1(0)$, the patient took (did not take) their medication on day t. Each patient has a transition matrix P^i is given by $P_{s,s'}^{i,a} = \Pr(s_{t+1}^i = s^{i\prime} \mid s_t^i = s, a_t^i = a)$. Each entry is determined by the probability that the patient adheres to their treatment plan at time t, depending on whether they are being acted on as well as their previous state. For the purposes of this paper, we assume all P^i are known. The reward for each MDP is $r^i(s^i) = s^i$.

A crucial component is that if a patient is not called $(a_t^i = 0)$ the health worker does not know the patient state s_t^i. Therefore our model is a two-state *partially-observable* MDP. Kaelbling et al. [1998] introduce an equivalent multi-state belief MDP, where states are represented by belief $b_t^i \in \mathcal{B} = [0, 1]$:

$$b_{t+1}^i = \begin{cases} s_{t+1}^i & \text{if } a = 1 \\ b_t^i P_{1,1}^{i,a=0} + (1 - b_t^i) P_{0,1}^{i,a=0} & \text{else.} \end{cases} \tag{1}$$

Our goal is to find a policy $\pi \colon \mathcal{S} \to \mathcal{A}$ that maximizes the number of patients in the adhering state over all arms, $\pi^* = \arg\max_\pi R^\pi(S)$. We define the reward to be $R^\pi(S) = \frac{1}{N} \sum_{i=1}^N \sum_{t=0}^T r\left(s_i^t\right)$. Mate et al. [2020] consider both discounted and time-average rewards. We use a variant of the latter since it puts equal weight on adherence throughout treatment.

A common Restless Bandit solution approach is the Whittle index, which is a heuristic that, at each time step, computes the value (i.e., index) of playing each arm in its current state, then greedily selects the k arms with greatest index (Whittle [1988]). Computing the index involves reasoning

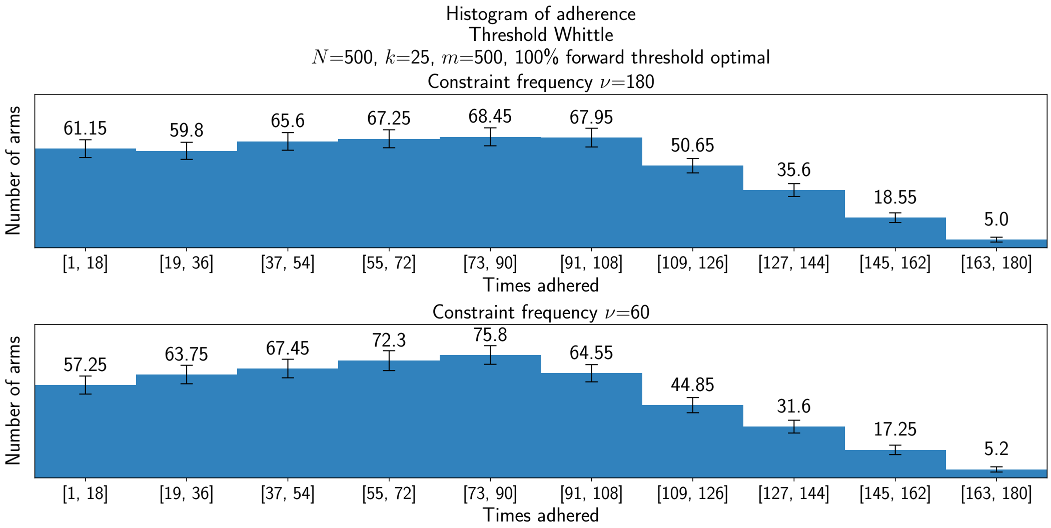

Figure 1: Histogram of adherence. (a) No constraints ($\nu = T$). (b) $m = N$, $\nu = 60$.

about an augmented version of each arm's MDP. Mate et al. [2020] show that when the the optimal policy of the augmented MDPs follow 'forward' or 'reverse' threshold policies, the Whittle index approach is asymptotically optimal. Sufficient conditions on P^i for an arm to be threshold optimal are given in Mate et al. [2020]; they provide an algorithm, Threshold Whittle, to find these policies.

2.1 Our Framework

In this paper, we present a human-in-the-loop framework for RMAB problems, with the goal of incorporating practitioners in the decision-making pipeline. Without domain expertise, the algorithm is unaware of key criteria and background information, so even in the best case it cannot make a fully-informed decision. We add a mechanism for practitioners to define additional criteria $g(\cdot)$:

$$\pi^* = \arg\max_{\pi} R^{\pi}(S) \text{ such that } g(\vec{a}) \geq c \tag{2}$$

This is a new general and flexible framework, which simplifies the problem of specifying and regulating policy characteristics. It may even be adapted to include statistical notions of fairness. However, this approach does not avoid undesirable behavior that the user never considered.

Throughout this paper, we use frequency constraints as an example; in this constraint, the arm must pulled at least once in any interval of ν time steps. Though the constraint considered here is over all time steps, a practitioner may selectively apply this constraint over the course of treatment instead.

3 Experiments on Simulated Data

We test the performance of our framework on simulations with $N = 500, T = 180$, and $k = 5\%$ of N. We vary the number of constrained arms $m \in \{0, 25, 50, 100, \ldots, N\}$ and the constraint frequency $\nu \in \left\{\frac{N}{k} = 20, 40, \ldots, T\right\}$, with 20 trials per combination of parameters. In each trial, the transition matrices were sampled from a uniform distribution, such that each was forward threshold optimal.

We apply the frequency constraint as follows: first, we pull any arms which would violate these specified constraints if not pulled. Next, the algorithm applies the optimal policy on the remaining budget. Any queued pulls that are over the budget k are instead pulled in the subsequent time step.

We measure the *relative reward*, defined as the reward normalized between the reward obtained via a policy that never acts and the reward obtained via the unconstrained Threshold Whittle solution. If $m = 0$ or $\nu = T$, no constraints are applied and the highest reward is obtained, but as m increases or ν decreases (making constraints tighter), the relative reward decreases. We also compare the performance of our framework against pulling random arms (*Random*) and the short-sighted greedy algorithm *Myopic*. At every time step t, the Myopic policy picks the top k processes with the largest values of $\Delta b_t = (b_{t+1} \mid a_t = 1) - (b_{t+1} \mid a_t = 0)$.

The source of decreased reward is investigated in Figure 1, which compares the distribution of adherences without our new framework against $m = N, \nu = 60$. The mean and variance of the distribution both decrease slightly, meaning that the number of arms that adhere consistently decreases, as do the number of arms that rarely adhere. Thus, we are able to improve the adherence pattern of poorly behaving arms, at the expense of arms that generally perform well.

Figure 2 compares the relative rewards for Threshold Whittle with other policies as (a) the number of constrained arms or (b) the enforced frequency of actions varies. In the top figure, the relative reward decreases only slightly as the number of constrained arms increases. At $m = N = 500, \nu = \frac{N}{k} = 20$, the constraints reduce the solution space of Threshold Whittle to only one feasible policy that is effectively same as Round Robin. Thus, the relative reward decreases similarly to logarithmically with ν in Figure 2(b). At $\nu = 20$, the relative reward is 55.52%. However, if the constraint frequency ν is increased to 60 (i.e., acting every 60 time steps), the relative reward increases to 89.44%. The reward changes continuously the set of feasible solutions is reduced. Thus, we could set a threshold preemptively, restricting the number of constraints to apply in the system interface.

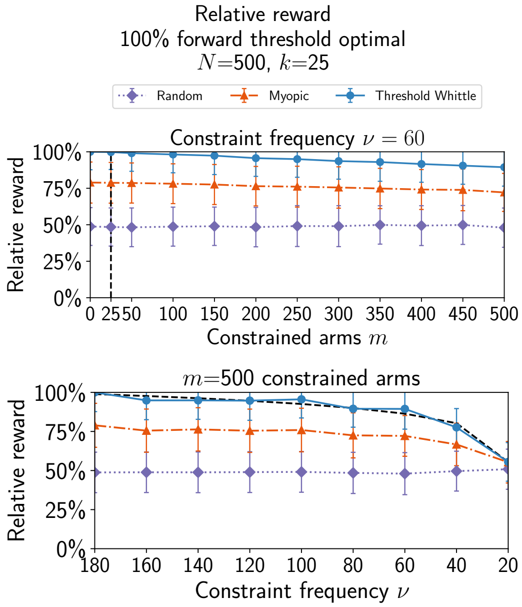

Figure 2: Relative reward. (a) $\nu = 60$. (b) $m = N$.

4 Discussion and Conclusion

This paper opens a new line of work related to restless multi-arm bandits (RMABs). We have demonstrated that including humans-in-the-loop for RMAB planning is a promising avenue towards safer and more responsible applications that can ensure a user-specified level of fairness while preserving benefits obtained from state-of-the-art algorithmic planning.

This is of increased importance as automated data-driven decision-support pipelines are introduced in a wide-range of public health applications. Increasingly, tools such as digital adherence technologies are being integrated to help support treatment of a number of diseases such as HIV (Haberer et al. [2017]), depression (Corden et al. [2016]), TB (Liu et al. [2015]), diabetes, and hypertension (Conway and Kelechi [2017]). However, without humans-in-the-loop, such tools can fall prey to the faults and biases that exist within that data and algorithms and lead to inequitable delivery of care. Our framework allows for domain experts to make corrections to state-of-the-art algorithms that currently do not account for the possibility of such disparities.

Further, within our framework, users can easily define and regulate criteria on the recommender tool's behavior that otherwise might render the tool unusable. In our TB domain, we have found that the optimal policy makes disparate recommendations such that some patients are visited every day, while others are never recommended for a visit. If recommendations are not sufficiently aligned or are entirely mis-aligned with the worker's domain expertise, and there is no mode by which the worker can give feedback, they may be unwilling to expend the effort of interacting with the tool altogether.

5 Acknowledgements

Aviva Prins was funded by the Harvard Center for Research on Computation and Society (CRCS)'s Summer Fellowship.

References

Nima Akbarzadeh and Aditya Mahajan. Restless bandits with controlled restarts: Indexability and computation of whittle index. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 7294–7300. IEEE, 2019.

PS Ansell, Kevin D Glazebrook, José Nino-Mora, and M O'Keeffe. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research*, 57(1):21–39, 2003.

Gianluca Baio, Mario Barbagallo, Giovanni D'Avola, Andrea Di Luccio, Gian Luca Di Tanna, Paolo Falaschi, Giovanni Iolascon, Nazzarena Malavolta, Federico Robbiati, and Fabio Massimo Ulivieri. Improving adherence in osteoporosis: a new management algorithm for the patient with osteoporosis. *Expert Opinion on Pharmacotherapy*, 12(2):257–268, 2011.

Cheryl Moseley Conway and Teresa J Kelechi. Digital health for medication adherence in adult diabetes or hypertension: an integrative review. *JMIR diabetes*, 2(2):e20, 2017.

Marya E Corden, Ellen M Koucky, Christopher Brenner, Hannah L Palac, Adisa Soren, Mark Begale, Bernice Ruo, Susan M Kaiser, Jenna Duffecy, and David C Mohr. Medlink: A mobile intervention to improve medication adherence and processes of care for treatment of depression in general medicine. *Digital Health*, 2:2055207616663069, 2016.

Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Apr 2020. doi: 10.1145/3313831.3376638. URL http://dx.doi.org/10.1145/3313831.3376638.

Matthew DeMichele, Peter Baumgartner, Kelle Barrick, Megan Comfort, Samuel Scaggs, and Shilpi Misra. What do criminal justice professionals think about risk assessment at pretrial? *Fed. Probation*, 83:32, 2019.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

Robert Epstein and Ronald E Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

Kevin D Glazebrook, Diego Ruiz-Hernandez, and Christopher Kirkbride. Some indexable families of restless bandit problems. *Advances in Applied Probability*, 38(3):643–672, 2006.

Jessica E Haberer, Nicholas Musinguzi, Alexander C Tsai, BM Bwana, C Muzoora, PW Hunt, JN Martin, DR Bangsberg, et al. Real-time electronic adherence monitoring plus follow-up improves adherence compared with standard electronic adherence monitoring. *AIDS (London, England)*, 31(1):169–171, 2017.

Yu-Pin Hsu. Age of information: Whittle index for scheduling stochastic arrivals. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 2634–2638. IEEE, 2018.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Jackson A Killian, Bryan Wilder, Amit Sharma, Vinod Choudhary, Bistra Dilkina, and Milind Tambe. Learning to prescribe interventions for tuberculosis patients using digital adherence data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2430–2438, 2019.

Joa Sang Lim and Marcus O'Connor. Judgemental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making*, 8(3):149–168, 1995.

Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, Nov 2010. ISSN 1557-9654. doi: 10.1109/tit.2010.2068950. URL `http://dx.doi.org/10.1109/TIT.2010.2068950`.

Xiaoqiu Liu, James J Lewis, Hui Zhang, Wei Lu, Shun Zhang, Guilan Zheng, Liqiong Bai, Jun Li, Xue Li, Hongguang Chen, et al. Effectiveness of electronic reminders to improve medication adherence in tuberculosis patients: a cluster-randomised trial. *PLoS medicine*, 12(9):e1001876, 2015.

Aditya Mate, Jackson A Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health interventions. *arXiv preprint arXiv:2007.04432*, 2020.

Rahul Meshram, D. Manjunath, and Aditya Gopalan. On the whittle index for restless multiarmed hidden markov bandits. *IEEE Transactions on Automatic Control*, 63(9):3046–3053, Sep 2018. ISSN 2334-3303. doi: 10.1109/tac.2018.2799521. URL `http://dx.doi.org/10.1109/TAC.2018.2799521`.

Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, May 1999. ISSN 1526-5471. doi: 10.1287/moor.24.2.293. URL `http://dx.doi.org/10.1287/moor.24.2.293`.

Bejjipuram Sombabu, Aditya Mate, D Manjunath, and Sharayu Moharir. Whittle index for aoi-aware scheduling. In *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 630–633. IEEE, 2020.

Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, pages 637–648, 1990.

P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25(A):287–298, 1988. ISSN 1475-6072. doi: 10.1017/s0021900200040420. URL `http://dx.doi.org/10.1017/s0021900200040420`.

Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414, 2019.

A Further Related Works

In general, solving an RMAB problem for an optimal solution is PSPACE hard when solved directly (Papadimitriou and Tsitsiklis [1999]). The Whittle index can find an optimal solution by decoupling arms using indices (Whittle [1988]). The Whittle index is asymptotically optimal and performs well empirically (Weber and Weiss [1990], Ansell et al. [2003]).

Whittle indexability has been investigated for many related bandit problems. Akbarzadeh and Mahajan [2019] give Whittle indexability for a bandit problem with "controlled restarts", where state-independent actions restart processes. However, we allow actions to be determined using state information. Glazebrook et al. [2006] give Whittle indexability for three classes of restless bandits where a subset of the action space has a deterministic effect on a processes' state. Hsu [2018] and Sombabu et al. [2020] expand their argument to limited forms of action-dependent state bandit problems, where an active action has some probability of no effect on the process. However, we have a fully stochastic state transition. Liu and Zhao [2010] give Whittle indexability for a special class of RMAB where the transition probabilities P are action-independent, i.e., $P^p = P^a$. Finally, Meshram et al. [2018] use Hidden Markov Bandits which is similar to the problem here, the only difference being that there are no state dependent rewards on passive arms.

B Why human-in-the-loop?

To show the utility of our framework, we demonstrate that without it, Threshold Whittle produces policies with undesirable behavior for the health applications we consider. We simulate experiments with $N = 500$, $T = 180$, and $k = 5\%$ of N. In each trial, the transition matrices for each arm were

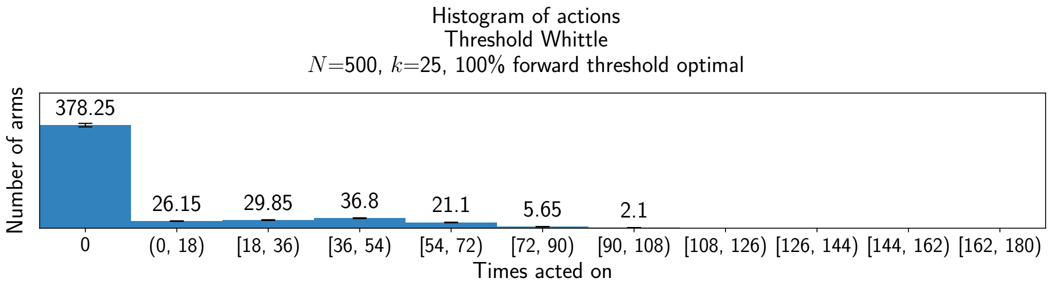

Figure 3: Without our framework, the majority of arms are never pulled.

generated from a uniform distribution within the respective trial condition. We find that in each case, the majority of arms are never called, while others are called with remarkably high frequency.

Figure 3 shows the distribution of actions found by Threshold Whittle on a 100% forward threshold optimal cohort. Here, 75.65% of arms are never pulled. The remaining 24.35% of arms are pulled with a distribution centered at 20-30% of $T = 180$, which corresponds to being pulled every 4.5 time steps on average.

This bipartite behavior is undesirable for our health application: if a health worker were to strictly adhere to this policy, over three-quarter of the patients would never receive a call, while others would receive one every day. The former would be unacceptable if those patients had low adherence in general and the latter may be overbearing by certain patients, lowering the effect of future interventions.

Recall that forward threshold optimal means that it is only optimal to act if our belief that an arm is adhering falls below a certain threshold. Threshold Whittle's disparate treatment of arms is consistent across threshold optimal cohorts. When the cohort is 100% *reverse* threshold optimal, the distribution of actions is completely bi-modal – 474 out of 500 individuals (95%) are never called and 24 individuals (5%) are called at every time step of the algorithm. Similar results are obtained for mixtures of the two types of cohorts.

It is not unlikely, however, that we will be able to obtain near-optimal utility when constraints are applied. As shown in Figure 4, the value of reward $R^{\pi}(S) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T} s_t^i$ differs by as little as 0.4297 between Threshold Whittle and the next best policy available (Myopic). This translates to a loss of 215 potential instances of adherence. For context, note that 90,000 adherence instances corresponds to 100% compliance across all individuals and times, for a reward $R(S^*) = 180$. As the forward threshold optimal increases, the difference between Threshold Whittle and Myopic policies also increases. At 100% forward threshold optimal, the difference is 2.3977, or a loss of 1,199 'good' state instances.

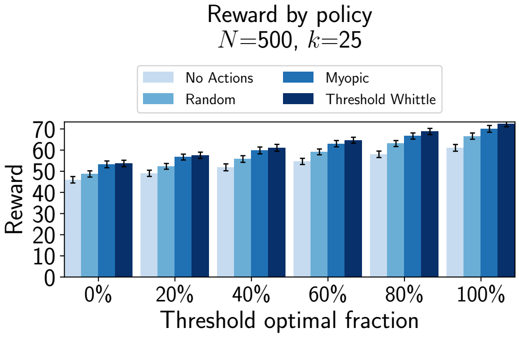

Figure 4: Reward $R^{\pi}(S)$ by policy and threshold optimal fraction.

We observe that current state-of-the-art RMAB approaches based on the Whittle index are inequitable. There is no mechanism to specify and constrain such behavior; this is the primary benefit of our new framework.

C Experiments on Patient Data

Courtesy of Killian et al. [2019], we have adherence information for $N = 7,676$ patients that participated in a tuberculosis pilot in India for at least 90 days. The pilot group has high adherence,

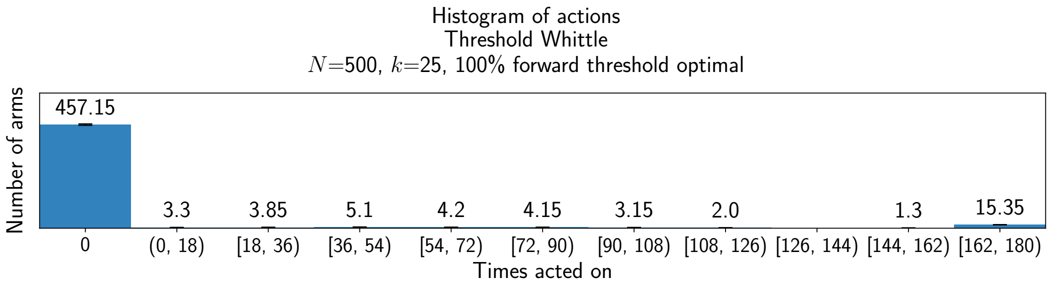

Figure 5: Histogram of adherence on tuberculosis pilot data, no constraints.

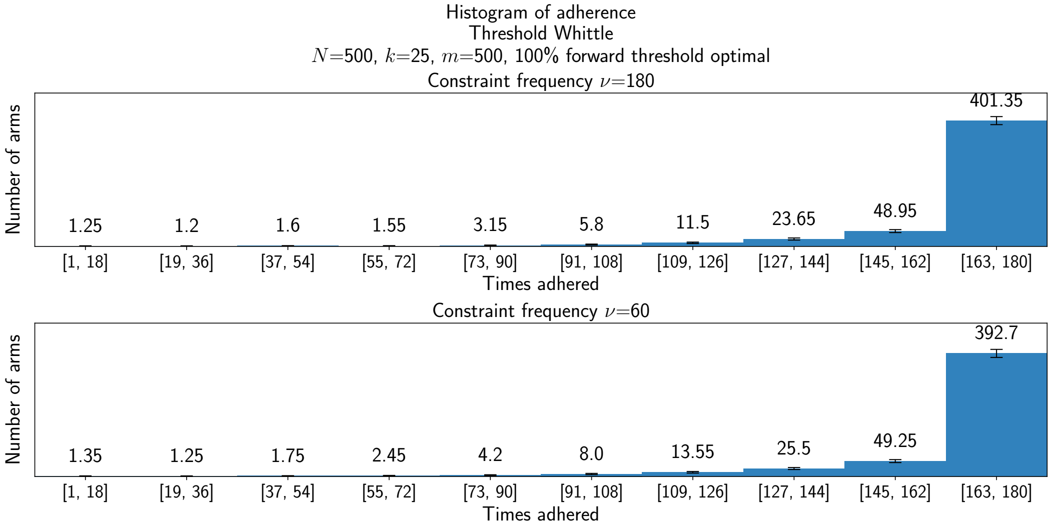

Figure 6: Histogram of adherence on TB pilot data. (a) No constraints. (b) $m = N$, $\nu = 60$.

with 74% of patients adhering at least 90% of the time. In fact, 45% adhere at least 99% of the time, which is exceptional. Less than 1% of patients in the dataset adhere less than 2% of the time.

We repeat the simulations in the previous section on transition matrices approximated from the adherence data. Throughout this section, we consider only forward threshold optimal transition matrices.

Figure 5, captures the distribution of action allocations per arm and shows a bimodal distribution when Threshold Whittle is applied without any constraints. Here, 457.15 out of 500 patients in our simulation are never called on average; 14.95 out of 500 patients in our simulation are called at every time step on average. This policy for arm pulls, while it may produce optimal adherence, also ignores a large majority of arms.

The population in our study has very high adherence rates, consistently (Figure 6(a)), so the transition matrices we generate are not uniformly distributed as in the simulated experiments above. When we apply the frequency constraint $\nu = 60$ to all $m = N$ arms, the rate of consistent adherence drops only slightly (Figure 6(b)).

Since we have high adherence patterns, the reward $R(S)$ is larger than in the simulated trials. However, the *relative* reward varies similarly (Figure 7). As ν approaches $\frac{N}{k} = 20$, the relative reward decreases considerably. Relative reward is normalized reward between the reward accrued upon never acting on arms and that accrued by Threshold Whittle. As the number of constrained arms increases, the reward decreases, albeit not catastrophically.

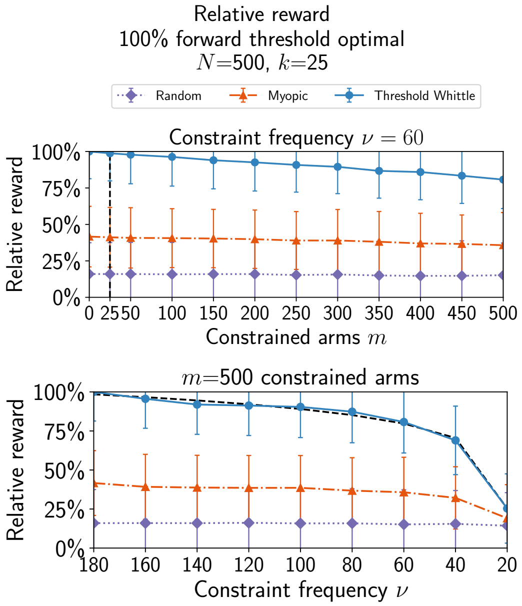

Figure 7: Relative reward on TB pilot data for (a) $\nu = 60$ and (b) $m = N$.