Optimization and Planning of Limited Resources for Assisting Non-Profits in Improving Maternal and Child Health

Aditya Mate

aditya_mate@g.harvard.edu

School of Engineering and Applied Sciences

Harvard University

ABSTRACT

The maternal mortality rate in India is appalling, largely fueled by lack of access to preventive care information, especially in low resource households. We partner with non-profit, ARMMAN, that aims to use mobile health technologies to improve the maternal and child health outcomes.

To assisst ARMMAN and such non-profits, we develop a Restless Multi-Armed Bandit (RMAB) based solution to help improve accessibility of critical health information, via increased engagement of beneficiaries with their program. We address fundamental research challenges that crop up along the way and present technical advances in RMABs and Planning Algorithms for Limited-Resource Allocation. Transcending the boundaries of typical laboratory research, we also deploy our models in the field, and present results from a first-of-its-kind pilot test employing and evaluating RMABs in a real-world public health application.

1. INTRODUCTION

The maternal mortality rate in India is appalling with 1 woman dying in childbirth every 20 minutes. More disturbingly, most of these deaths are preventable, if only the expectant mothers have timely access to critical live-saving health information. ARMMAN, an Indian NGO is focused on reducing maternal and child mortality rates in underserved and underprivileged communities. Quoting Dr. Aparna Hedge, founder of ARMMAN, "Pregnancy is not a disease. Childhood is not an ailment. Dying due to a natural life event is unacceptable". Having impacted lives of 26 million women so far, ARMMAN serves new and expectant mothers

Figure 1: Picture credits: ARMMAN

by leveraging the penetration of cellphone technology in India to provide timely and free, automated voice calls or text messages conveying critical health information to these mothers. Studies have shown that when mothers listen to these automated messages, it significantly improves the health outcomes.

However, one recurrent issue such ARMMAN faces is that many women tend to drop out over a period of time or do not listen to the voice messages completely, missing out on the critical information and leading to negative health consequences. To alleviate this issue, health workers at ARMMAN may provide service calls to pregnant women, encouraging them to listen to the health information. However, while serving millions of mothers, the limited health worker staff at ARMMAN can only reach out to a very small fraction of these enrolled mothers each week. This makes it critical to utilize the limited health worker resources optimally, and to try and target those beneficiaries likely to benefit the most from these service calls.

Viewed algorithmically, the key challenge is to optimize the allocation of limited resources. The objective is to maximize the health messages listened to by the beneficiary cohort. With limited resources, the key question is to identify which k out of N (where $k \ll N$) mothers to select each week for service calls.

Our work is the first to cast this problem as a Restless Multi-Armed Bandit (RMAB), popularly studied in the operations research literature. The RMAB solution approach however, doesn't work straight out-of-the-box and needs several fundamental research advances to tackle the unique challenges faced by such non-profits. To this end, we develop new methods to make the RMAB techniques computationally inexpensive and scalable, making them accessible to such non-profits. We also build methodology to handle real-world considerations such as risk-aware planning and dynamically changing beneficiary cohorts. We further innovate new clustering methods to infer necessary RMAB parameters that underpin the RMAB solution approaches, but are unknown in the real-world. Without stopping at merely these technical innovations in the laboratory, we also conducted field visits (Figure **??**), interacting with the actual beneficiaries and the healthworkers at ARMMAN to understand and identify where innovation would benefit the beneficiaries the most. We are also the first to run a large-scale field trial deploying our RMAB solution, in a real-world public health application, in partnership with ARMMAN. We present results from the field experiment which show that our algorithm cuts engagement drops among beneficiaries by 30% in comparison to the current standard of care.

2. OPERATIONS RESEARCH TECHNIQUES AND CONTRIBUTIONS

The central Operations Research question we try to address in our work, is to decide how to allocate the limited health worker resources — specifically, how to identify the subset of beneficiaries to deliver service calls to, each week — so as to maximize the overall benefits of the service calls to the beneficiary cohort. Here we present an

(a) (b) (c)

Figure 2: Pictures from field visit in Mumbai, in July 2022. (a),(b): Site of the field visit (c) During a house visit to a beneficiary for interview accompanied by an ARMMAN health worker (called "Sakhi", translating to "female friend" in Hindi).

overview of the challenges encountered on the way and our key innovative contributions aimed at tackling these challenges.

Restless Multi-Armed Bandits: Restless Multi-Armed Bandits (RMAB) is a popular framework that has seen significant theoretical investigation in the past for handling resource allocation problems, in a myriad of domains such as communication systems [1], UAV routing [2], sensor and machine maintenance [3] and so on. We are the first to cast the health-worker challenge of engagement monitoring and internvention planning as an RMAB problem [4].

Planning the optimal allocation policy in RMABs has been shown to be PSPACE hard in general [5]. Previously available solution techniques are computationally too expensive needing hours to run on a computing cluster, rendering it inaccessible to resource strapped non-profits who may not have access to such computation machinery. Towards allowing RMABs to be utilized by such non-profits, we study a special subclass of RMABs that captures our problem setting, that we call "Collapsing Bandits", and design a fast algorithm that exploits the special structure of the application. We show that our algorithm achieves a 3-order-of-magnitude speedup, while maintaining similar performance quality. This enables applying the RMAB planning techniques in the context of ARMMAN, in determining beneficiaries to deliver interventions to, without needing access to powerful computational resources. Details of the algorithmic idea, theoretical and empirical results are described in Section 3.

Streaming Bandits: The beneficiary cohort served by ARMMAN changes dynamically — there is an incoming stream of new beneficiaries joining the program and an outgoing stream of existing enrolled beneficiaries that leave the system. The stay of beneficiaries in the program is thus naturally finite, spanning only a very limited number of weeks. Unfortunately, existing RMAB solutions assume the beneficiary cohort to begin and end the program sychronously and assume an intermediate stay of an infinite duration. We show that performance of such algorithms degrades when the stay of beneficiaries gets shorter, even when they all start synchronously. The performance only dwindles further if the beneficiaries join asynchronously. Approaches that do account for the finite stay are computationally expensive, and can't scale well to problems of ARMMAN's size.

Recognizing that the available accurate solutions do not scale well and that the scalable solutions ride on unrealistic assumptions leading to poor performance in practice, we focus on planning interventions for a dynamically changing beneficiary cohort. We propose 'Streaming Bandits' [6] to accommodate incoming and outgoing streams of bandit arms (beneficiaries), while also accounting for the finite stay of the beneficiaries. Specifically, we propose an interpolation technique and show that interpolating between the cheaply available solutions for the infinite- and small-horizon problems is nearly as effective as solving the finite horizon problem exactly. In context of ARMMAN, this speeds up the planning algorithm by 2-orders-of-magnitude, without sacrificing on performance, even when planning interventions for a dynamically changing cohort. More details are in Section 4

Real-world deployment challenges: Although RMABs have seen significant theoretical investigation in the literature, none of these have been tested in the field or have seen real-world deployment in context of public health. Existing works either assume knowledge of intermediate RMAB transition parameters underpinning the planning algorithms or rely on being able to learn those easily online. For new and expectant mothers however, both assumptions are untrue, as these parameters are unknown in the real-world and learning those online is stymied by the beneficiaries' short stay in the program. This poses a new fundamental challenge to overcome to be able to deploy RMABs in the real-world for improving maternal healthcare.

In Section 5 we present novel clustering methods [7] that leverage abundant historical data on previously enrolled beneficiaries to infer these parameters for new, unseen beneficiaries that join the program.

Field Trial and Evaluation: Transcending previous RMAB studies and theoretical investigations, we take the RMAB model into the field, in a first-of-its-kind large scale field trial in partnership with ARMMAN. This study, spanning a period of 7 weeks, involves 23,0003 real-world beneficiaries enrolled with ARMMAN. The results from the field trial show that our RMAB algorithm achieves a statistically significant improvement in the engagement behavior of the beneficiaries and manages to cut engagement drops among mothers by 30% in

comparison to the current standard of care. Detailed analysis is presented in Section 6, followed by additional discussions on lessons learnt through this study, and conclusions in Sections 7 and Section 8 respectively.

3. COLLAPSING BANDITS

We adopt the solution framework of Restless Multi-Arm Bandits (RMABs), focusing on a special subclass, that we call *"Collapsing Bandits"*, capturing the health worker planning problem we consider. In the Collapsing Bandits framework, the planner must act on k out of N binary-state processes each round. The planner fully observes the state of the processes on which she acts, then all processes undergo an action-dependent Markovian state transition; the state of the process may be unobserved in the general case, until it is acted upon again, resulting in uncertainty. The planner's goal is to maximize the number of processes that are in some "good" state over the course of T rounds. This class of problems is natural in the context of non-profits such as ARMMAN, that grapple with such monitoring and intervention planning tasks. For instance, the health workers at ARMMAN must choose a subset of beneficiaries to deliver service calls each week, with the goal of maximizing the number of beneficiaries who engage with the information program.

Solving an RMAB is PSPACE-hard in general [5]. The predominant solution techniques to RMAB problems, is a heuristic known as the *Whittle index policy*, that computes a 'Whittle index' for each arm, that is designed to capture the value of acting upon that arm. The policy then chooses the top k arms with the largest indices for delivery of action. The Whittle index policy generally performs well empirically, and has been shown to be asymptotically optimal under a technical condition called 'indexability'. The existence of such an index policy also hinges upon satisfying the indexability condition.

In summary, using the Whittle index policy requires two key components: (i) a fast method for computing the index and (ii) proving that the problem satisfies indexability. Without (i) the approach can be prohibitively slow, and without (ii) asymptotic performance guarantees are sacrificed and in fact, such an index may not exist in the first place. Neither (i) nor (ii) are known for RMABs in general or for Collapsing Bandits, in particular. In the following subsections, we prove nice theoretical properties on Collapsing Bandits, which help prove indexability. Next we leverage these theoretical properties to build a fast algorithm that unlocks a 3-order-of-magnitude speedup over existing methods and finally present empirical results.

3.1 Indexability in Collapsing Bandits

Indexability guarantees the existence and asymptotic optimality of the Whittle index approach porposed by Whittle in [8]. Intuitively, the Whittle index captures the value of acting on an arm in a particular state by finding the minimum *subsidy* m the agent would accept to *not act*, where the subsidy is some exogenous

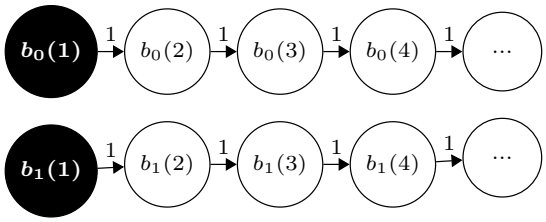


Figure 3: Belief-state MDP under the policy of always being passive. There is one chain for each observation $\omega \in \{0, 1\}$ with the head marked black. Belief states deterministically transition down the chains.

"donation" of reward. Indexability, requires that for all states, the optimal action at the state cannot switch from passive to active as m increases.

To address the potential partial observability of states in the collapsing bandit framework, we adopt the belief state representation in our analysis. This presentation is central to our indexability results as well as the fast computation algorithm.

3.1.1 Belief State MDP representation

In limited observability settings, belief-state MDPs have organized chain-like structures, which we will exploit. In particular, the only information that affects our belief of an arm being in state 1 is the number of days since that arm was last pulled and the state ω observed at that time. Therefore, we can arrange these belief states into two "chains" of length T, each for an observation ω. A sketch of the belief state chains under the passive action is shown in Fig. 3. Let $b_\omega(u)$ denote the belief state, which in context of maternal health,is the probability that the engagement state is 1 at the current time step, if the beneficiary was observed to be in state $\omega \in \{0, 1\}$ when it was acted upon, u times teps ago. Note that $b_\omega(u)$ is also the expected reward associated with that belief state, and let $\mathcal{B}$ be the set of all belief states.

3.1.2 Indexability Theorem

Definition 3.1 (Indexability). *Let Π^*_m be the set of policies that maximize a given reward criterion under subsidy m. An arm is indexable if $\mathcal{B}^*(m) = \{b : \forall \pi \in \Pi^*_m, \pi(b) = 0\}$ monotonically increases from $\emptyset$ to the entire state space as m increases from $-\infty$ to ∞. An RMAB is indexable if every arm is indexable.*

We identify the following special type of MDP policy, central to our analysis, that helps prove indexability and also yields a fast index computation algorithm.

Definition 3.2 (Threshold Policies). *A policy is a* forward (reverse) threshold policy *if there exists a threshold b_{th} such that $\pi(b) = 0$ $(\pi(b) = 1)$ if $b > b_{th}$ and $\pi(b) = 1$ $(\pi(b) = 0)$ otherwise.*

Theorem 3.3. *If for each arm and any subsidy $m \in \mathbb{R}$, there exists an optimal policy that is a forward or*

reverse threshold policy, the Collapsing Bandit is indexable under discounted and average reward criteria.

Proof. (Sketch)

Using linearity of the value function in subsidy m for any fixed policy, we first argue that when forward (reverse) threshold policies are optimal, proving indexability reduces to showing that the threshold monotonically decreases (increases) with m. Unfortunately, establishing such a monotonic relationship between the threshold and m is a well-known challenging task in the literature that often involves problem-specific reasoning [1]. Our proof features a sophisticated induction argument exploiting the finite size of $\mathcal{B}$ and relies on tools from real analysis for limit arguments. □

All formal details of the complete proof can be found in [4]. We remark that Thm. 3.3 generalizes the result in the seminal work by [1] who proved the indexability for only a special class of collapsing bandits. To bolster our proof, we also identify conditions on the transition matrix P of the beneficiaries, under which the optimal policy is of forward or reverse threshold type.

3.1.3 Optimality of Threshold Policies

Let P denote the MDP transition function of a beneficiary, where $P^a_{s,s'}$ denotes the probability of transitioning from state s to s' when action a is taken. We theoretically identify conditions on P, which determine whether the optimal policy for the beneficiary is of forward or reverse threshold type.

THEOREM 3.4. *Consider a belief-state MDP corresponding to an arm in a Collapsing Bandit. For any subsidy m, there is a* forward *threshold policy that is optimal under the condition:*

$$(P^p_{1,1} - P^p_{0,1})(1 + \beta(P^a_{1,1} - P^a_{0,1}))(1 - \beta) \geq P^a_{1,1} - P^a_{0,1} \tag{1}$$

Proof. (Sketch) Forward threshold optimality requires that if the optimal action at a belief b is passive, then it must be so for all $b' > b$. This can be established by requiring that the derivative of the passive action value function is greater than the derivative of the active action value function w.r.t. b. The main challenge is to distill this requirement down to measurable quantities so the final condition can be easily verified. We accomplish this by leveraging properties of belief state update and using induction to derive both upper and lower bounds on the difference in value functions of the beneficiaries in different belief states. □

Intuitively, the condition requires that the intervention effect on beneficiaries in the "non-engaging" state

must be large, making $P^a_{1,1} - P^a_{0,1}$ small.

THEOREM 3.5. *Consider a belief-state MDP corresponding to an arm in a Collapsing Bandit. For any subsidy m, there is a* reverse *threshold policy that is optimal under the condition:*

$$(P^p_{1,1} - P^p_{0,1})\Big(1 + \frac{\beta(P^a_{1,1} - P^a_{0,1})}{1-\beta}\Big) \leq P^a_{1,1} - P^a_{0,1} \tag{2}$$

Intuitively, the condition requires small intervention effect on processes in the "non-engaging" state, the opposite of the forward threshold optimal requirement. Note that both Thm. 3.4 and Thm. 3.5 also serve as conditions for the average reward case as $\beta \to 1$ (a proof based on Dutta's Theorem [9] is given in [4]).

3.2 Fast Algorithm

Although the Whittle index is known to be challenging to compute in general [8], we are able to design an algorithm that computes the Whittle index efficiently, using the knowledge of optimality of threshold policies.

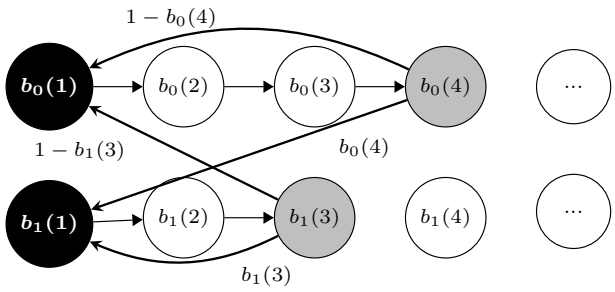
The main algorithmic idea we use is the Markov chain structure that arises from imposing a forward threshold policy on an MDP. A forward threshold policy can be defined by a tuple of the first belief state in each chain that is less than or equal to some belief threshold $b_{th} \in [0,1]$. In the two-observation setting we consider, this is a tuple $(X_0^{b_{th}}, X_1^{b_{th}})$, where $X_\omega^{b_{th}} \in 1, \ldots, T$ is the index of the first belief state in each chain where it is optimal to act (i.e., the belief is less than or equal to b_{th}). We now drop the superscript b_{th} for ease of exposition. See Fig. 4a for a visualization of the transitions induced by such an example policy. For a forward threshold policy (X_0, X_1), the occupancy frequencies induced for each state $b_\omega(u)$ are:

$$f^{(X_0,X_1)}(b_\omega(u)) = \begin{cases} \alpha & \text{if } \omega = 0, u \leq X_0 \\ \beta & \text{if } \omega = 1, u \leq X_1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$\alpha = \left(\frac{(X_1 b_0(X_0))}{1 - b_1(X_1)} + X_0\right)^{-1}, \beta = \left(\frac{X_1 b_0(X_0)}{1 - b_1(X_1)} + X_0\right)^{-1} \frac{b_0(X_0)}{1 - b_1(X_1)} \tag{4}$$

These equations are derived from standard Markov chain theory. These occupancy frequencies do not depend on the subsidy. Let $J_m^{(X_0,X_1)}$ be the average reward of policy (X_0, X_1) under subsidy m. We decompose the average reward into the contribution of the state reward and the subsidy

$$J_m^{(X_0,X_1)} = \sum_{b \in \mathcal{B}} b f^{(X_0,X_1)}(b) + m(1 - f^{(X_0,X_1)}(b_1(X_1)) - f^{(X_0,X_1)}(b_0(X_0))) \tag{5}$$



(a)

Figure 4: (a) Visualization of forward threshold policy ($X_0 = 4$,$X_1 = 3$). Black nodes are the head of each chain and grey nodes are the thresholds.

Recall that for any belief state $b_\omega(u)$, the Whittle index is the smallest m for which the active and passive actions are both optimal. Given forward threshold optimality, this translates to two corresponding threshold policies being equally optimal. We show that such policies must have adjacent belief states as thresholds. Note that for a belief state $b_0(X_0)$ the only adjacent threshold policies with active and passive as optimal actions at $b_0(X_0)$ are (X_0, X_1) and (X_0+1, X_1) respectively. Thus the subsidy which makes these two policies equal in value must necessarily be the Whittle Index for $b_0(X_0)$, which we obtain by solving: $J_m^{(X_0,X_1)} = J_m^{(X_0+1,X_1)}$ for m. We use this idea to construct the following fast Whittle index algorithm.

Sequential index computation algorithm Alg. 1 precomputes the Whittle index of every belief state for each process, having time complexity $\mathcal{O}(|\mathcal{S}|^2TN)$. Then, the per-round complexity to retrieve the top k indices is $\mathcal{O}(N \min\{k, log(N)\})$. This gives a great improvement over the more general method given by Qian et al. [10] (state-of-the-art) which has per-round complexity of $\approx \mathcal{O}(N \log(\frac{1}{\epsilon})(|\mathcal{S}|T)^{2+\frac{1}{18}})$, where $\log(\frac{1}{\epsilon})$ is due to a bifurcation method for approximating the Whittle index to within error ϵ on each arm and $(|\mathcal{S}|T)^{2+\frac{1}{18}}$ is due to the best-known complexity of solving a linear program with $|\mathcal{S}|T$ variables.

Alg. 1 is optimized for settings in which the Whittle index can be precomputed. However, for online learning settings, we also give an alternative method that computes the Whittle index on-demand, in a closed form.

Algorithm 1: Sequential index computation algorithm

Initialize counters to heads of the chains: $X_1 = 1$, $X_0 = 1$
while $X_1 < T$ *or* $X_0 < T$ **do**
  Compute $m_1 := m$ such that $J_m^{(X_0,X_1)} = J_m^{(X_0,X_1+1)}$
  Compute $m_0 := m$ such that $J_m^{(X_0,X_1)} = J_m^{(X_0+1,X_1)}$
  Set $i = \arg\min\{m_0, m_1\}$ and $W(X_i) = \min\{m_0, m_1\}$
  Increment X_i
end

3.3 Empirical Results

We evaluate our algorithm on several domains using both real and synthetic data distributions. We test the following algorithms: **Threshold Whittle** is the algorithm developed in this paper. [10], a slow, but precise general method for computing the Whittle index, is the state-of-the-art that we improve upon. **Random** selects k process to act on at random each round. **Myopic** acts on the k processes that maximize the expected reward at the immediate next time step. Formally, at time t, this policy picks the k processes with the largest values of $\Delta b_t = (b_{t+1}|a=1) - (b_{t+1}|a=0)$. **Oracle** fully observes all states and uses [10] to calculate Whittle indices. We measure performance in terms of *intervention benefit*, where 0% corresponds to the reward of a policy that is always passive and 100% corresponds to Oracle. All results are averaged over 50 independent trials.

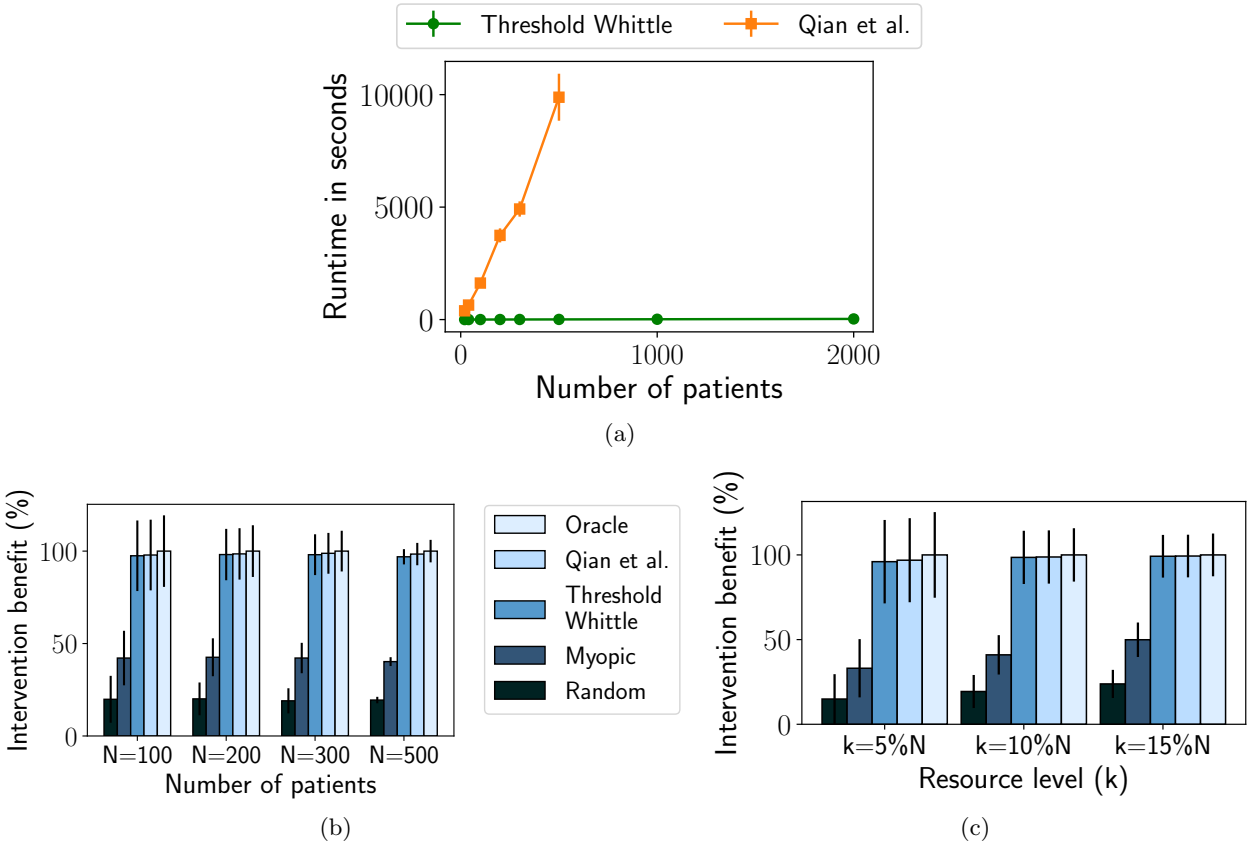


Figure 5: (a),(b): Threshold Whittle is several orders of magnitude faster than Qian et al. and scales to thousands of patients without sacrificing performance on realistic data. (c) Intervention benefit of Threshold Whittle is far larger than naive baselines and nearly as large as Oracle.

Real Data Experiments We test on real data obtained for a tuberculosis medication adherence monitoring task, analogous to the beneficiary engagement monitoring challenge we face. This tuberculosis data set contains

daily adherence information recorded for each real tuberculosis patient in the system, as obtained from [11]. The "good" and "bad" states of the arm (here, patient) correspond to "Adhering" and "Not Adhering" to medication, respectively. State transition probabilities are estimated from the data. Reward is measured as the undiscounted sum of patients (arms) in the adherent state over all rounds, where each trial lasts $T = 180$ days (matching the length of first-line TB treatment) with N patients and a budget of k calls per day. Recall that this setup is analogous to the problem setting of ARMMAN, where the reward is measured as the sum of beneficiaries across the cohort in the engaging state, and the goal is to maximize this sum, while selecting some k beneficiaries for intervention out of a cohort of strength N, over a period of T timesteps.

In Fig. 5a, we plot the runtime in seconds vs the number of patients N. Fig. 5b compares the intervention benefit for $N = 100, 200, 300, 500$ patients and $k = 10\%$ of N. In the $N = 200$ case, the runtimes of a single trial of Qian et al. and Threshold Whittle index policy are 3708 seconds and 3 seconds, respectively, while attaining near-identical intervention benefit. Our algorithm is thus 3 orders of magnitude faster than the previous state of the art without sacrificing performance.

We next test Threshold Whittle as the resource level k is varied. Fig. 5c shows the performance in the $k = 5\%N$, $k = 10\%N$ and $k = 15\%N$ regimes ($N = 200$). Threshold Whittle outperforms Myopic and Random by a large margin in these low resource settings.

4. HANDLING DYNAMICALLY CHANGING BENEFICIARY COHORT

The mathematical underpinnings of the RMAB framework developed for intervention planning, assume a setting involving an infinite time horizon (i.e., they assume the ARMMAN information programs to run forever) and, moreover, the results are limited to settings where no new beneficiaries (or bandit arms) arrive midway during the program. While this initial setup yields useful algorithms and reveals valuable technical insights, the static cohorts assumption can prove to be restrictive in deploying RMABs for use by ARMMAN. To counter this issue, we propose a new, general class of RMABs, which we call *streaming restless multi-armed bandits*, or S-RMAB. In an S-RMAB instance, the arms of the bandit are allowed to arrive asynchronously, that is, the planner observes an incoming and outgoing stream of bandit arms. The classic RMAB (both with infinite and finite horizon) is a special case of the S-RMAB where all arms appear (leave) at the same time. Additionally, each arm of an S-RMAB is allowed to have its own transition probabilities, capturing the potentially heterogeneous nature of beneficiary cohorts.

We develop new theory and algorithms to extend the benefits of our fast index computation algorithm to the streaming case, which we present in this section as follows. We first crystallize the additional challenge to

address in the Streaming bandits setup. We then show that the indexability condition still holds true, even for dynamically changing cohorts in S-RMAB. Next, we identify a key phenomenon responsible for making planning in S-RMAB challenging and finally propose a fast algorithm that counters this, yielding a policy that is both, inexpensive to run and displays near-optimal performance. Finally we show with empirical evaluation on real data demonstrating the utility of our approach in planning service call interventions better for ARMMAN.

4.1 Problem Formulation

The *streaming restless multi-armed bandit* (S-RMAB) problem is a general class of RMAB problem where a stream of arms arrive over time (both for finite and infinite-horizon problems). Similar to RMAB, at each time step, the decision maker is allowed to take active actions on at most k of the available N arms.

Contrary to previous approaches that typically consider arms to all arrive at the beginning of time and stay forever, in this setup, we consider streaming multi-armed bandits—a setting in which arms are allowed to arrive asynchronously and have finite lifetimes. We denote the number of arms arriving and leaving the system at a time step $t \in [T]$ by $X(t)$ and $Y(t)$, respectively. Each arm i arriving at time t, is associated with a fixed lifetime L_i (for example, L_i can be used to represent the duration of stay of beneficiaries in the ARMMAN program, which is known to the planner). The arm consequently leaves the system at time $t + L_i$. Thus, instead of assuming a finite set of N arms throughout the entire time horizon, we assume that the number of arms at any time t is denoted by the natural number $N(t)$, and can be computed as $N(t) = \sum_{s=1}^{t}(X(s) - Y(s))$. Thus, the goal of the planner is to decide, at each time step t, which k arms to pull (out of the $N(t) \gg k$ arms, relabeled as [N(t)] each timestep for ease of representation), in order to maximize her total reward,

$$\bar{R} := \sum_{t \in [T]} \sum_{i \in [N(t)]} r_t(i). \tag{6}$$

4.2 Indexability in Streaming Bandits

We extend the conditions for indexability that established previously for static cohorts and infinite horizon, to the finite horizon setting of Streaming bandits. To show indexability, we first show in Theorem 4.1, that S-RMABs can be reduced to a standard RMAB with augmented belief states. We build on this result and prove another useful Lemma, both of which combined can be used to show that indexability holds for this augmented RMAB instance, and ultimately for S-RMABs (Theorem 4.3).

Our strategy in proving indexability is as follows. First, we show that the belief state MDP of a Streaming Bandit arm with deterministic arrival and departure time can be formulated as an augmented belief state MDP

of the same instance with infinite horizon. Using this, we prove that, whenever the infinite horizon problem satisfies threshold optimality for a passive subsidy m, then the augmented belief state MDP for finite horizon also satisfies threshold optimality. Finally, leaning on the result from previous section proving that indexability holds whenever threshold optimality is satisfied, we imply that the Streaming Bandits problem is also indexable whenever threshold optimality on the underlying infinite horizon problem is satisfied.

THEOREM 4.1. *The belief state transition model for a 2-state Streaming Bandit arm with deterministic arrival time T_1 and departure time T_2 can be reduced to a belief state model for the standard restless bandit arm with $T_2 + (T_2 - T_1)^2$ states.*

Proof Sketch. We incorporate the arrival and departure of Streaming arms by constructing a new belief model where each state is represented by a tuple $\langle$ `behavior, time-step` $\rangle$ where `behavior` may either take a value in $(0, 1)$ or U (unavailable). Details of the proof are deferred to the supplementary material.

LEMMA 4.2. *If a forward (or reverse) threshold policy π is optimal for a subsidy m for the belief states MDP of the infinite horizon problem, then π is also optimal for the augmented belief state MDP.*

Proof Sketch The proof relies on the key observation that it is never optimal to take the *active* action on an arms in the $\langle U, t\rangle$ state of the augmented belief model. In this state, because the actions have no effect, both actions are already equally optimal for a passive subsidy of $m = 0$, which is strictly less than the minimum passive subsidy required when the arm is in any other state.

THEOREM 4.3. *A Streaming Bandits instance is indexable when there exists an optimal policy, for each arm and every value of $m \in \mathbb{R}$, that is forward (or reverse) threshold optimal policy.*

Proof. Using Theorem 1 and Lemma 1, it is straightforward to see that an optimal threshold policy for infinite horizon problem can be translated to a threshold policy for Streaming bandits instance. Moreover, using the fact that the existence of threshold policies for each subsidy m and each arm $i \in N$ is sufficient for indexability to hold (Theorem 1 of [4]), we show that the Streaming bandit problem is also indexable. □

4.3 Index Decay Phenomenon

Despite having shown indexability, the Whittle Indices computed using Threshold Whittle in the previous section cannot be used out-of-the-box, as those are computed assuming an infinite stay of beneficiaries in the system. In reality however, some beneficiaries in the program may have only recently joined, while others may be nearing completion and about to leave. This differential in the duration of stay remaining, translates to differing impacts of calling beneficiaries and thus affect the prioritization order. We show that performance of Threshold

Whittle algorithm degrades when the lifetime of arms gets shorter, even when all arms start synchronously. The performance only dwindles further if arms were to arrive asynchronously.

In this section we identify the key phenomenon responsible for this issue, that we call *index decay*. This effect manifests as the beneficiaries approach the end of their stay in the program. Simply put, the Whittle index values are low when the residual lifetime of an arm is 0 or 1. We formalize this observation in Theorem 4.4. We use this phenomenon as an anchor to develop our algorithm (detailed in the next subsection).

THEOREM 4.4 (INDEX DECAY). *Let $V^p_{m,T}(b)$ and $V^a_{m,T}(b)$ be the T-step passive and active value functions for a belief state b with passive subsidy m. Let m_T be the value of subsidy m, that satisfies the equation $V^p_{m,T}(b) = V^a_{m,T}(b)$ (i.e. m_T is the Whittle Index for a residual life time of T). Assuming indexability holds, we show that:* $\forall T > 1$: $m_T > m_1 > m_0 = 0$.

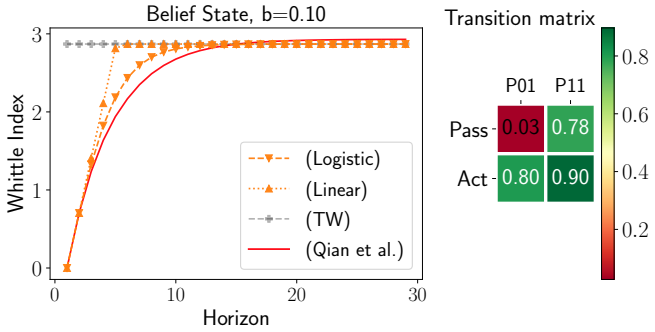


Figure 6: Whittle Indices for a belief state as computed by different algorithms. Both our algorithms capture index decay providing good estimates.

4.4 Interpolation Algorithm

The key insight driving the design of our solution is that, by accounting for the index decay phenomenon, we can bypass the need to solve the costly finite horizon problem. We make use of the fact that we can cheaply compute index values for arms with residual lifetime 0 and 1, where the index decay phenomenon occurs, and for infinite horizon bandits. Our proposed solution for computing indices for arbitrary residual lifetime is to use a suitable functional form to interpolate between those three observations. We propose an interpolation template, that can be used to obtain two such algorithms, one using a piece-wise linear function and the other using a logistic function.

We establish in Theorem 4.4 that the Whittle Index for arms with a zero residual lifetime, is always zero. Similarly, indices for arms with residual lifetime of 1 are simply the myopic indices, computed as:

$$\Delta b = \left(b\ P^a_{11} + (1-b)\ P^a_{01}\right) - \left(b\ P^p_{11} + (1-b)\ P^p_{01}\right).$$

Algorithm 2: Interpolation Algorithm Template

Pre-compute $\overline{W}(b, P^i)\ \forall b \in \mathcal{B}_i,\ \forall\ i \in [N]$, with transition matrix P^i and set of belief states $\mathcal{B}_i$.
Input: $\bar{b}_{N\times 1} \in [0,1]^N,\ \bar{h}_{N\times 1} \in [L]^N$, containing the belief values and remaining lifetimes for the N arms.
Initialize $\hat{W}_{N\times 1}$ to store estimated Whittle Indices.
for each arm i in N **do**
    Let $b \coloneqq \bar{b}_i, h \coloneqq \bar{h}_i$ and let P be i's transition matrix.
    Compute the myopic index Δb as:
    $\Delta b = \left(b\ P^a_{11} + (1-b)\ P^a_{01}\right) - \left(b\ P^p_{11} + (1-b)\ P^p_{01}\right)$.
    Set $\hat{W}_i(h, \Delta b, \overline{W})$ according to one of the interpolation functions (7) or (8).
end for
Pull the k arms with the largest values of $\hat{W}$.

For the linear interpolation, we assume $\hat{W}(h)$, our estimated Whittle Index, to be a piece-wise-linear function of the horizon, h (with two pieces), capped at a maximum value of the Whittle Index for the infinite horizon problem, corresponding to $h = \infty$. We denote Whittle Index for infinite horizon as $\overline{W}$. Note that $\overline{W}$ is simply the 'Threshold Whittle' index computed using the Collapsing Bandits machinery described in the previous section. The first piece of the piece-wise-linear $\hat{W}(h)$ must pass through the origin, given that the Whittle Index is 0 when the residual lifetime is 0. The slope is determined by $\hat{W}(h=1)$ which must equal the myopic index, given by Δb. The second piece is simply the horizontal line $y = \overline{W}$ that caps the function to its infinite horizon value. The linear interpolation index value is thus given by

$$\hat{W}(h, \Delta b, \overline{W}) = \min\{h\ \Delta b, \overline{W}\}. \tag{7}$$

The linear interpolation algorithm performs well and has very low run time, as we will demonstrate in the later sections. However, the linear interpolation can be improved by using a logistic interpolation instead. The logistic interpolation algorithm yields moderately higher rewards in many cases for a small additional compute time. For the logistic interpolation, we let

$$\hat{W}(h, \Delta b, \overline{W}) = \frac{C_1}{1 + e^{-C_2 h}} + C_3. \tag{8}$$

We now apply the three constraints on the Whittle Index established earlier and solve for the three unknowns $\{C_1, C_2, C_3\}$ to arrive at the logistic interpolation model. For the residual lifetimes of 0 and 1, we have that $\hat{W}(0) = 0$ and $\hat{W}(1) = \Delta b$. As the horizon becomes infinity, $\hat{W}(.)$ must converge to $\overline{W}$, giving the final constraint $\hat{W}(\infty) = \overline{W}$. Solving this system yields the solution:

$$C_1 = 2\overline{W},\ C_2 = -\log\left(\left(\frac{\Delta b}{C_1} + \frac{1}{2}\right)^{-1} - 1\right),\ C_3 = -\overline{W}.$$

We note that both interpolations start from $\hat{W} = 0$ for $h = 0$ and saturate to $\hat{W} = \overline{W}$ as $h \to \infty$.

We compare the index values computed by our interpolation algorithms with the exact solution by [10]. Figure 6 shows an illustrative example, plotting the index values as a function of the residual lifetime and shows that the interpolated values agree well with the exact values.

4.5 Empirical Evaluation

We evaluate the performance and runtime of our proposed algorithms against several baselines, using both, real as well as synthetic data distributions. LOGISTIC and LINEAR are our proposed algorithms. Similar to previous evaluations, our main baselines are: (1) QIAN ET AL. [10], providing a precise, but slow algorithm which accounts for the residual lifetime by solving the expensive finite-horizon POMDP on each of the N arms and finds the k best arms to pull and (2) Threshold-Whittle (marked in figures as TW), a much faster algorithm developed in the previous section, but designed to work for infinitely long residual time horizons. We again include the MYOPIC policy, that plans interventions optimizing for the expected reward of the immediate next time step. RANDOM is a naive baseline that pulls k arms at random.

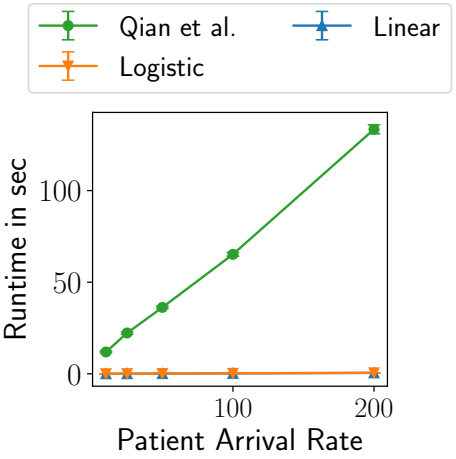
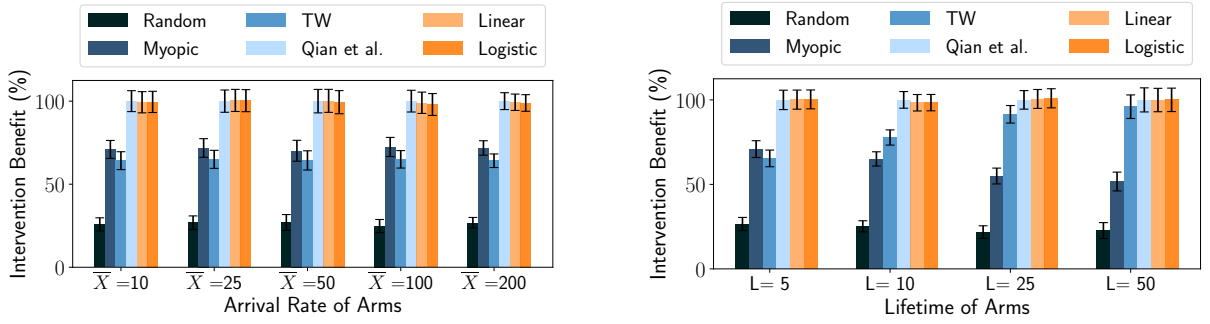


Figure 7: Linear and Logistic interpolation algorithms are nearly 200× faster than Qian et al.

Performance is again measured as the excess average intervention benefit over a 'do-nothing' policy, measuring the sum of rewards over all arms and all timesteps minus the reward of a policy that never pulls any arms. Intervention benefit is normalized to set QIAN ET AL. equal to 100% and can be obtained for an algorithm `ALG` as: $\frac{100\times(\overline{R}^{\texttt{ALG}} - \overline{R}^{\text{No intervention}})}{\overline{R}^{\text{Qian et al.}} - \overline{R}^{\text{No intervention}}}$ where $\overline{R}$ is the average reward. All simulation results are measured and averaged over 50 independent trials and error bars denote the standard errors.



(a) (b)

Figure 8: The interpolation algorithms achieve the speedup without sacrificing on performance, while other fast algorithms like Threshold Whittle deteriorate significantly for small residual horizons.

Evaluation on ARMMAN data In Figures 7 and 8, we again consider the finite horizon setting with a deterministic incoming stream of patients. In Figure 7, we plot the runtimes of our algorithms and that of QIAN ET AL., as a function of the weekly arrival rate, $\bar{X}$ of the incoming stream. Figure 8a measures the intervention benefits of these algorithms for these values of $\bar{X}$. The lifetime of each arm, L is fixed to 5 weeks and the number of resources, k is set to $10\% \times (\bar{X}L)$. Each simulation was run for a total length T such that $\bar{X}T = 5000$, which is the total number of arms involved in the simulation. Runtime is measured as the time required to simulate L decision timesteps. The runtime of [10] quickly far exceeds that of our algorithms. For the $\bar{X} = 200$ case, a single trial of QIAN ET AL. takes 106.69 seconds to run on an average, while the proposed Linear and Logistic interpolation algorithms take 0.47 and 0.49 seconds respectively, while attaining virtually identical intervention benefit. Threshold Whittle, while being similarly fast, assumes an infinite residual horizon, and consequently suffer a severe degradation in performance for such short residual horizons. Our algorithms thus manage to achieve a dramatic speed up over existing algorithms, without sacrificing on performance.

In Figure 8b, we consider an S-RMAB setting, in which arms continuously arrive according to a deterministic schedule, and leave after staying on for a lifetime of L, which we vary on the x-axis. We also study the isolated effects of small lifetimes and asynchronous arrivals separately as well as performance in settings with stochastic arrivals. Across the board, we find that the performance of TW degrades as the lifetime becomes shorter and that this effect only exacerbates with asynchronous arrivals. The performance of our algorithms remains at par with QIAN ET AL., in all of the above.

5. REAL-WORLD DEPLOYMENT CHALLENGES

Despite the significant algorithmic contributions of existing RMAB literature, these works have remained confined to their theoretical significance because they assume the transition parameters underpinning the RMAB to be available as an input. However, these transition parameters of real-world beneficiaries such as pregnant mothers are not just unknown, but also hard to infer in practice. To address this issue, we use clustering techniques that exploit historical data D_{train} of beneficiaries, to estimate an offline RMAB problem instance relying solely on the beneficiaries' static features and state transition data.

Clustering also helps solve the issue of limited samples (time-steps) available per beneficiary. While there is limited historical service call data (active transition samples) for any single beneficiary, clustering on the beneficiaries allows us to combine their data to infer transition probabilities for the entire group. Clustering offers the added advantage of reducing computational cost for resource limited NGOs; because all beneficiaries within a cluster share identical transition probability values, we can compute their Whittle indices all at once.

5.1 PPF Clustering

The key motivation here is to group together beneficiaries with similar transition behaviors, irrespective of their features. To this end, we use k-means clustering on passive transition probabilities (to avoid issues with missing active data) of beneficiaries in D_{train} and identify cluster centers. We then learn a map ϕ from the feature vector f to the cluster assignment of the beneficiaries that can be used to infer the cluster assignments of new beneficiaries at test-time solely from f. We use a random forest model as ϕ.

5.2 Evaluation of Clustering Method

In [7], we compare against additional three clustering methods, suitable for our application and find the PPF clustering method performs the best. We compare against the following alternative clustering methods:

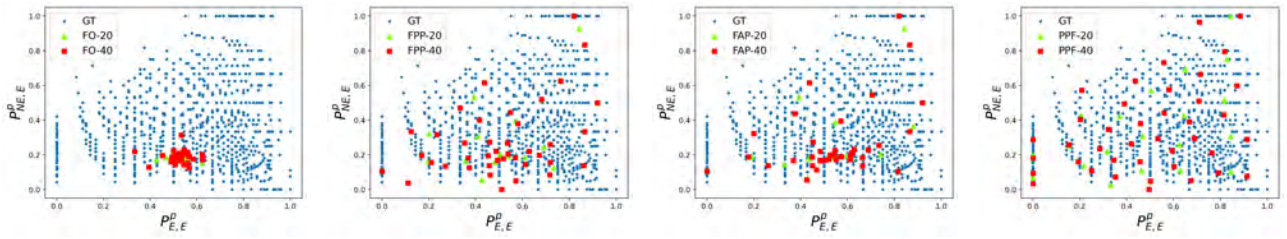
1. Features-only Clustering (FO): This method relies on the correlation between the beneficiary feature vector f and their corresponding engagement behavior. We employ k-means clustering on the feature vector f of all beneficiaries in the historic dataset D_{train}, and then derive the representative transition probabilities for each cluster by pooling all the (s, α, s') tuples of beneficiaries assigned to that cluster. At test time, the features f of a new, previously unseen beneficiary in D_{test} map the beneficiary to their corresponding cluster and estimated transition probabilities.

2. Feature + All Probabilities (FAP) In this 2-level hierarchical clustering technique, the first level uses a rule-based method, using features to divide beneficiaries into a large number of pre-defined buckets, B. Transition probabilities are then computed by pooling the (s, α, s') samples from all the beneficiaries in each bucket. Finally, we perform a k-means clustering on the transition probabilities of these B buckets to reduce them to k clusters ($k \ll B$). However, this method suffers from several smaller buckets missing or having very few active transition samples.

3. Feature + Passive Probabilities (FPP): This method builds on the FAP method, but only considers the passive action probabilities to preclude the issue of missing active transition samples.

5.2.1 Comparison against Baselines

We use a historical dataset, $\mathcal{D}_{train}$ from ARMMAN consisting of 4238 beneficiaries in total, who enrolled into the program between May-July 2020. We compare the clustering methods empirically, based on the criteria described below.



(a) FO clustering (b) FPP clustering (c) FAP clustering (d) PPF clustering

Figure 9: Comparison of passive transition probabilities obtained from different clustering methods with cluster sizes $k = \{20, 40\}$ with the ground truth transition probabilities. Blue dots represent the true passive transition probabilities for every beneficiary while red or green dots represent estimated cluster centres.

1. Representation: Cluster centers that are representative of the underlying data distribution better resemble the ground truth transition probabilities. This is of prime importance to the planner, who must rely on these values to plan actions. Fig 9 plots the ground truth transition probabilities and the resulting cluster centers determined using the proposed methods. Visual inspection reveals that the *PPF* method represents the ground truth well, as is corroborated by the quantitative metrics of Table 1 that compares the RMSE error across different clustering methods.

2. Balanced cluster sizes: A low imbalance across cluster sizes is desirable to preclude the possibility of arriving at few, gigantic clusters which will assign identical whittle indices to a large groups of beneficiaries. Working with smaller clusters also aggravates the missing data problem in estimation of active transition probabilities. Considering the variance in cluster sizes and RMSE error for the different clustering methods with $\kappa = \{20, 40\}$ as shown in Table 1, *PPF* outperforms the other clustering methods and was chosen for the pilot study.

Clustering Method	Average RMSE		Standard Deviation	
	$\kappa = 20$	$\kappa = 40$	$\kappa = 20$	$\kappa = 40$
FO	0.229	0.228	143.30	74.22
FPP	0.223	0.222	596.19	295.01
FAP	0.224	0.223	318.46	218.37
PPF	0.041	0.027	145.59	77.50

Table 1: Average RMSE and cluster size variance over all beneficiaries for different methods. Total Beneficiaries = 4238, $\mu_{20} = 211.9$, $\mu_{40} = 105.95$ (μ = average beneficiaries per cluster)

Next we turn to choosing κ, the number of clusters: as κ grows, the clusters become sparse in number of active samples aggravating the missing data problem while a smaller κ suffers from a higher RMSE. We found $\kappa = 40$ to be the optimal choice for the pilot study.

Finally, we adopt the Whittle solution approach for RMABs to plan actions and pre-compute all of the

possible $2 * \kappa$ index values that beneficiaries can take (corresponding to combinations of κ possible clusters and 2 states). The indices can then be looked up at all future time steps in constant time, making this an optimal solution for large scale deployment with limited compute resources.

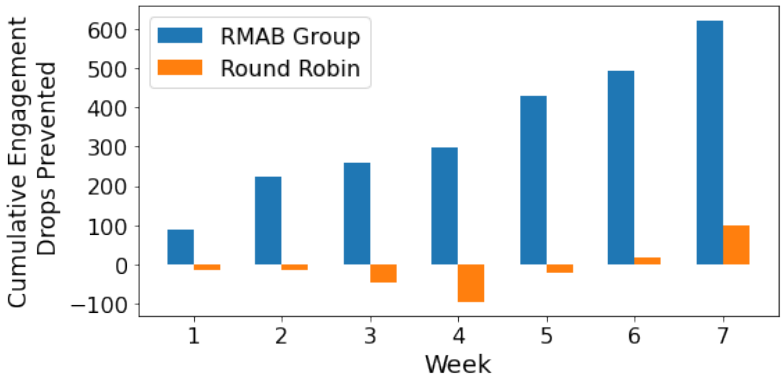
As we prepared this RMAB system for real-world use, there was as an important observation for social impact settings: real-world use also required us to carefully handle several domain specific challenges, which were time consuming. For example, despite careful clustering, a few clusters may still be missing active probability values, which required employing a data imputation heuristic. Moreover, there were other constraints specific to ARMMAN, such as a beneficiary should receive only one service call every η weeks, which was addressed by introducing "sleeping states" for beneficiaries who receive a service call.

6. FIELD TRIAL AND REAL-WORLD EVALUATION

6.1 Pilot Study Results

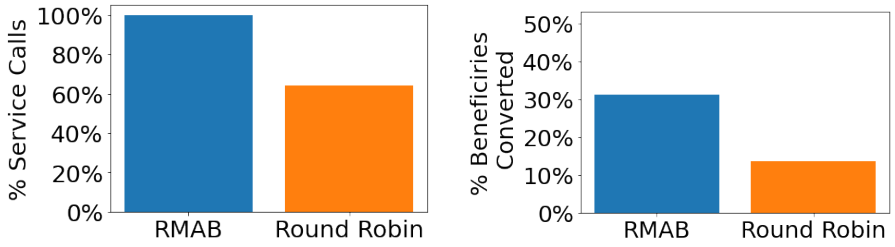
In this first-of-its-kind such effort, we ran a real-world trial in partnership with ARMMAN, implementing our RMAB algorithm for selecting beneficiaries for service call delivery. We ran a field trial involving 23,003 new and expectant mothers over a period of 7 weeks. These mothers were uniformly and equally divided into three groups — (1) RMAB, where serive calls were delivered according to our algorithm (2) Round Robin, where the exact same number of service calls were delivered to beneficiaries selected according to a set order and (3) Current Standard of Care (CSOC) exercising ARMMAN's current mode of operation without service calls. Of the total $\sim$ 7668 beneficiaries per group, an average of 225 women received service calls per week in the RMAB and Round Robin groups. The ARMMAN staff performing service calls were blind to the experimental groups that the beneficiaries belonged to. Beneficiaries across all three groups receive the same automated voice messages regarding pregnancy and post-birth care throughout the program, and no health related information was withheld from any beneficiary. Because engagement generally dwindles over time, we measured the effectiveness of the service calls in terms of the number of engagement drops prevented, in comparison to CSOC.

In Figure 10a, real-world numbers from the experiment show that the RMAB algorithm prevents a total 622 instances of missed engagement with the automated health messages at the end of 7 weeks, as compared to CSOC, which sees a total of 1944 missed engagements. In other words, the RMAB algorithm prevents 32% of the engagement drops seen by the CSOC group. We also show statistical significance in the improvement in the engagement behavior offered by the RMAB algorithm. These encouraging results ratify that our algorithms are not limited to theoretical significance on the whiteboard, but can be translated to actual social impact at scale.



(a)

Figure 10: (a) RMAB prevents a much higher number of engagement drops than the baseline. (b) RMAB allocates service calls to more non-engaging beneficiaries than the baseline. (c) Success rate of RMAB algorithm is higher in converting non-engaging beneficiaries of week 1 to the engaging state by week 7 upon delivery of service calls.



(a) (b)

Figure 11: (a) RMAB prevents a much higher number of engagement drops than the baseline. (b) RMAB allocates service calls to more non-engaging beneficiaries than the baseline. (c) Success rate of RMAB algorithm is higher in converting non-engaging beneficiaries of week 1 to the engaging state by week 7 upon delivery of service calls.

6.2 Statistical Significance

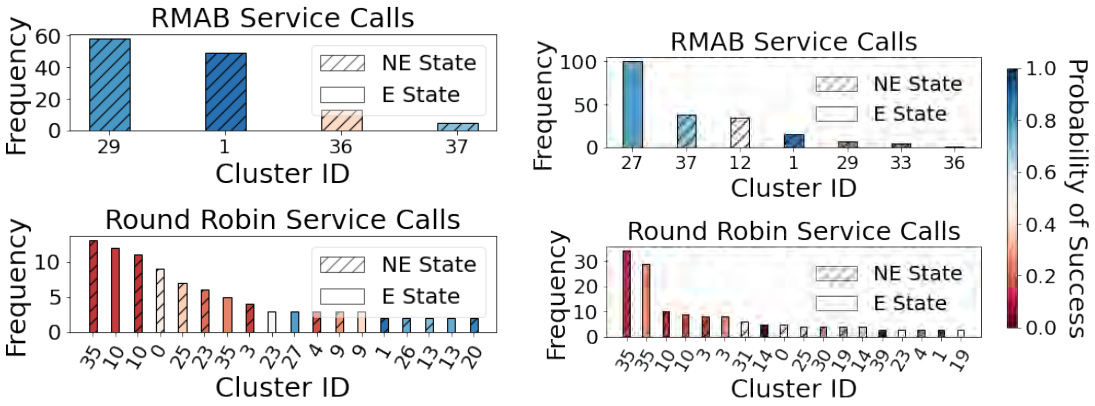
To investigate the benefit from use of RMAB policy over policies in the RR and CSOC groups, we use regression analysis [12]. Specifically, we fit a linear regression model to predict number of cumulative engagement drops at week 7 while controlling for treatment assignment and covariates specified by beneficiary registration features. The model is given by:

$$Y_i = k + \beta T_i + \sum_{j=1}^{J} \gamma_j x_{ij} + \epsilon_i$$

where for the i_{th} beneficiary, Y_i is the outcome variable defined as number of cumulative engagement drops at week 7, k is the constant term, β is the treatment effect, T_i is the treatment indicator variable, x_i is a vector of length J representing the i_{th} beneficiary's registration features, γ_j represents the impact of the j_{th} feature on the outcome variable and ϵ_i is the error term. For evaluating the effect of RMAB service calls as compared to CSOC group, we fit the regression model only for the subset of beneficiaries assigned to either of these two

	RMAB vs CSOC	**RR vs CSOC**	**RMAB vs RR**
% reduction in cumulative engagement drops	32.0%	5.2%	28.3%
p-value	0.044^{*}	0.740	$0.098^{\dagger}$
Coefficient β	-0.0819	-0.0137	-0.0068

Table 2: Statistical significance for service call policy impact at week 7 is tested using a linear regression model. We use: $^{*}p < 0.05$; $^{\dagger}p < 0.1$



(a) Week 1 Service Calls

(b) Week 2 Service Calls

Figure 12: Distributions of clusters picked for service calls by RMAB and RR are significantly different. RMAB is very strategic in picking only a few clusters with a promising probability of success, RR displays no such selection.

groups. T_i is set to 1 for beneficiaries belonging to the RMAB group and 0 for those in CSOC group. We repeat the same experiment to compare RR vs CSOC group and RMAB vs RR group.

The results are summarized in Table 2. We find that RMAB has a statistically significant treatment effect in reducing cumulative engagement drop (negative $\beta, p < 0.05$) as compared to CSOC group. However, the treatment effect is not statistically significant when comparing RR with CSOC group ($p = 0.740$). Additionally, comparing RMAB group with RR, we find β, the RMAB treatment effect, to be significant ($p < 0.1$). This shows that RMAB policy has a statistically significant effect on reducing cumulative engagement drop as compared to both the RR policy and CSOC. RR fails to achieve statistical significance against CSOC. Together these results illustrate the importance of RMAB's optimization of service calls, and that without such optimization, service calls may not yield any benefits.

6.3 Understanding RMAB Strategy

We analyse RMAB's strategic selection of beneficiaries in comparison to RR using Figure 12, where we group beneficiaries according to their whittle indices, equivalently their ⟨`cluster, state`⟩. Figure 12 plots the frequency distribution of beneficiaries (shown via corresponding clusters) who were selected by RMAB and RR in the first two weeks. For example, the top plot in Figure 12a shows that RMAB selected 60 beneficiaries from cluster 29 (NE state). First, we observe that RMAB was clearly more selective, choosing beneficiaries from just four (Figure 12a) or seven (Figure 12b) clusters, rather than RR that chose from 20. Further, we assign each cluster a hue based on their probability of transitioning to engaging state from their current state given a service call. Figure 12 reveals that RMAB consistently prioritizes clusters with high probability of success (blue hues) while RR deploys no such selection; its distribution emulates the overall distribution of beneficiaries across clusters (mixed blue and red hues).

Furthermore, Figure 11a further highlights the situation in week 1, where RMAB spent 100% of its service calls on beneficiaries in the non-engaging state while RR spent the same on only 64%. Figure 11b shows that RMAB converts 31.2% of the beneficiaries shown in Figure 11a from non-engaging to engaging state by week 7, while RR does so for only 13.7%. This further illustrates the need for optimizing service calls for them to be effective, as done by RMAB.

7. LESSONS LEARNED AND FUTURE VISION

Our work in deploying RMABs for maternal healthcare offered several key lessons on the way, to be adopted in my future research endeavors for social impact [7]. First, partnerships with NGOs including field-visits and beneficiary-focused discussion about the challenges faced reveals unique and fundamental research questions to be addressed before transitioning the technology from the whiteboard to actual impact on the ground. Second, data and compute limitations faced by non-profits are genuine research challenges preventing them from benefiting from the existing technology. Finally, in deploying AI methods for social impact, there may be several technical challenges that don't need innovative solutions, but are critical barriers to large-scale deployment.

It was very gratifying to hear feedback from Dr. Aparna Hedge, founder of ARMMAN, who says "We have seen that when women listen to the information, the health outcomes are phenomenal. We are able to reach out to more and more women each week, bring them back into the fold and save lives because of AI". A mother of a 5-year old enrolled with ARMMAN, vouches "They explained the benefits of listening to the messages. Now I listen to the calls regularly, it feels like someone from your own family is looking after you. I follow all advice and take good care of my baby".

Figure 13: During an in-person discussion with ARMMAN staff and healthworkers

Encouraged by these results, our vision is to work with ARMMAN and roll out the AI-powered algorithm to 1 million women. I am also involved in other ongoing efforts exploring ideas such as decision-focused learning to improve the quality of recommendations or building agent-based simulators to use as a test-bed with instantaneous turnaround time for evaluation before deploying an algorithm in the real-world.

In addition to my technical work, I am also very enthusiastic and passionate about encouraging research, focused on social good. I co-organize Pasteur's Quadrant Seminar Series, a student-run multi-institutional, multi-country initiative that spotlights AI for social good work and builds a community of practitioners in this space. I've also co-organized the Harvard CRCS Rising Stars workshop and speaker series, highlighting prominent work in this space. To work in close collaboration with partner organizations, I've also taken up field trips in the past, visiting ARMMAN at their office in Mumbai and interacting with the health workers, including a recent visit in July this year. Previously, I've also organized and participated in AI and Tuberculosis workshops in Mumbai and met with the city TB officers and health workers in my other research.

8. CONCLUSION

Our project is focused on using optimization and planning techniques to tackle public health challenges in low-resources settings, specifically, those faced by non-profits working towards improving maternal and child health. The widespread use of cell-phones, particularly in the global south, has enabled non-profits to launch massive programs delivering key health messages to a broad population of beneficiaries in a cost-effective manner.

We innovate fundamental technical advances in RMABs that enable building a system to assist these non-profits in optimizing their limited service resources. We present results from a real-world evaluation of our

system, conducted in partnership with ARMMAN, an India-based NGO. To the best of our knowledge, ours is the first study to build and demonstrate the effectiveness of such RMAB-based resource optimization in real-world public health contexts. These encouraging results have led to the transition of our RMAB software to ARMMAN for real-world deployment. We hope this work paves the way for use of RMABs in many other health service applications.

–

References

[1] Liu, K. and Zhao, Q., "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Transactions on Information Theory* **56**(11), 5547–5567 (2010).

[2] Le Ny, J., Dahleh, M., and Feron, E., "Multi-uav dynamic routing with partial observations using restless bandit allocation indices," in [*2008 American Control Conference*], 4220–4225, IEEE (2008).

[3] Glazebrook, K. D., Ruiz-Hernandez, D., and Kirkbride, C., "Some indexable families of restless bandit problems," *Advances in Applied Probability* **38**(3), 643–672 (2006).

[4] Mate, A., Killian, J. A., Xu, H., Perrault, A., and Tambe, M., "Collapsing bandits and their application to public health interventions," in [*Advances in Neural and Information Processing Systems (NeurIPS) 2020*], (2020).

[5] Papadimitriou, C. H. and Tsitsiklis, J. N., "The complexity of optimal queuing network control," *Math. Oper. Res.* **24**(2), 293–305 (1999).

[6] Mate, A., Biswas, A., Siebenbrunner, C., and Tambe, M., "Efficient algorithms for finite horizon and streaming restless multi-armed bandit problems," *arXiv preprint arXiv:2103.04730* (2021).

[7] Mate, A., Madaan, L., Taneja, A., Madhiwalla, N., Verma, S., Singh, G., Hegde, A., Varakantham, P., and Tambe, M., "Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health," in [*Proc. 36th AAAI Conference on Artificial Intelligence (AAAI-22)*], (2022).

[8] Whittle, P., "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.* **25**(A), 287–298 (1988).

[9] Dutta, P., "What do discounted optima converge to?: A theory of discount rate asymptotics in economic models," *Journal of Economic Theory* **55**(1), 64–94 (1991).

[10] Qian, Y., Zhang, C., Krishnamachari, B., and Tambe, M., "Restless poachers: Handling exploration-exploitation tradeoffs in security domains," in [*AAMAS*], (2016).

[11] Killian, J., Wilder, B., Sharma, A., Choudhary, V., Dilkina, B., and Tambe, M., "Learning to prescribe interventions for tuberculosis patients using digital adherence data," in [*KDD*], (2019).

[12] Angrist, J. D. and Pischke, J.-S., [*Mostly harmless econometrics*], Princeton university press (2008).