Efficient Algorithms for Finite Horizon and Streaming Restless Multi-Armed Bandit Problems

Aditya Mate
Harvard University
Boston, MA, U.S.A.
aditya_mate@g.harvard.edu

Arpita Biswas
Harvard University
Boston, MA, U.S.A.
arpitabiswas@seas.harvard.edu

Christoph Siebenbrunner
Harvard University
Boston, MA, U.S.A.
csiebenbrunner@seas.harvard.edu

Susobhan Ghosh
Harvard University
Boston, MA, U.S.A.
susobhan_ghosh@g.harvard.edu

Milind Tambe
Harvard University
Boston, MA, U.S.A.
milind_tambe@harvard.edu

ABSTRACT

We propose Streaming Bandits, a Restless Multi-Armed Bandit (RMAB) framework in which heterogeneous arms may arrive and leave the system after staying on for a finite lifetime. Streaming Bandits naturally capture the health-intervention planning problem, where health workers must manage the health outcomes of a patient cohort while new patients join and existing patients leave the cohort each day. Our contributions are as follows: (1) We derive conditions under which our problem satisfies indexability, a pre-condition that guarantees the existence and asymptotic optimality of the Whittle Index solution for RMABs. We establish the conditions using a polytime reduction of the Streaming Bandit setup to regular RMABs. (2) We further prove a phenomenon that we call index decay — whereby the Whittle index values are low for short residual lifetimes — driving the intuition underpinning our algorithm. (3) We propose a novel and efficient algorithm to compute the index-based solution for Streaming Bandits. Unlike previous methods, our algorithm does not rely on solving the costly finite horizon problem on each arm of the RMAB, thereby lowering the computational complexity compared to existing methods. (4) Finally, we evaluate our approach via simulations run on real-world data sets from a tuberculosis patient monitoring task and an intervention planning task for improving maternal healthcare, in addition to other synthetic domains. Across the board, our algorithm achieves a 2-orders-of-magnitude speed-up over existing methods while maintaining the same solution quality.

KEYWORDS

Restless Multi-Armed Bandits; Whittle Index; Finite Horizon; Intervention Planning

ACM Reference Format:
Aditya Mate, Arpita Biswas, Christoph Siebenbrunner, Susobhan Ghosh, and Milind Tambe. 2022. Efficient Algorithms for Finite Horizon and Streaming Restless Multi-Armed Bandit Problems. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 12 pages.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

1 INTRODUCTION

In community healthcare settings, adherence of patients to prescribed health programs, that may involve taking regular medication or periodic health checkups, is critical to their well-being. One way to improve patients' health outcomes is by tracking their health or monitoring their adherence to such programs. Such health monitoring programs combined with suitably designed intervention schemes help patients alleviate health issues such as diabetes [26], hypertension [5], tuberculosis [6, 30], depression [20, 25], etc. However, health interventions often require dedicated time of healthcare workers, which is a severely scarce resource, grossly inadequate to meet the total demand. This issue is especially more severe in the global south. Moreover, planning interventions with these limited resources is made more challenging due to the fact that the extent of adherence of patients may be both, uncertain as well as transient. Consequently, the healthcare workers have to grapple with this sequential decision making problem of deciding which patients to intervene on, with limited resources, in an uncertain environment. Existing literature on healthcare monitoring and intervention planning (HMIP) [1, 2, 21–23] casts this as a *restless multi-armed bandit* (RMAB) planning problem. In this setup, the patients are typically represented by the arms of the bandit and the planner must decide which arms to pull (which patients to intervene on) under a limited budget. The RMAB problem formalizes the (restless) behavioral dynamics of the patients both in the presence and in the absence of interventions.

In addition to healthcare, RMABs have caught traction as solution techniques in a myriad of other domains involving limited resource planning for applications such as anti-poaching patrol planning [29], multi-channel communication systems [18], sensor monitoring tasks [7], UAV routing [15] etc. For ease of presentation, we consider the HMIP problem for motivation but our approach is relevant and can be extended to other real-world domains.

The existing literature on RMABs for intervention planning, however, has mainly focused on problems involving an infinite time horizon (i.e., the health programs are assumed to run forever) and, moreover, the results are limited to settings where no new patients (or bandit arms) arrive midway during the health program. We consider a 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 patient cohorts. S-RMABs display a special structure in the presence of streaming arms and a finite horizon, which the existing methods fail to utilize. Our approach exploits this structure to arrive at approaches that perform better in the streaming bandit setting.

A fairly general approach, proposed by [29] may be applied even when patients arrive and leave asynchronously after staying for a finite duration. The method allows to approximate the exact solution arbitrarily well, but it is computationally expensive as the number of patients or arms increases. A more recent approach, proposed by [21], exploits the structure of the HMIP and is considerably faster, but the method relies on the assumption of an infinite planning horizon. This algorithm suffers a severe deterioration in performance when employed on shorter horizon settings.

Our **contribution** consists of proposing a new approach, designed for the finite-horizon and asynchronous arrival settings, that achieves a combination of the advantages of existing methods, i.e. high solution quality and low runtime, in those settings. We provide theoretical justifications for the use of Whittle indices in streaming RMABs, as well as for the setup of our algorithms, designed to leverage the structure of the finite horizon and asynchronous cases. We further show that our method also applies to S-RMAB arms exhibiting *reverse threshold optimality*, while previous methods only applied to settings with *forward threshold optimality*. We perform experimental evaluations of our algorithms using real-world data from two domains, as well as synthetic and adversarial domains. Our algorithms provide a 2-orders-of-magnitude speed-up compared to existing accurate methods, without loss in performance.

2 RELATED WORK

The RMAB problem was introduced by [33]. The paper studied the RMAB problem with the goal of maximizing the average reward in a dynamic programming framework. Whittle formulated a relaxation of the problem and provided a heuristic called *Whittle Index policy*. This policy is optimal when the underlying Markov Decision Processes satisfy indexability, which is computationally intensive to verify. Later, [28] established that solving RMAB is PSPACE hard, even when the transition rules are known. Since then, specific classes of RMABs have been studied extensively. [29] studied the infinite horizon RMAB problem and proposed a binary search based algorithm to find Whittle index policy. However, the algorithm becomes computationally expensive as the number of arms grows. [2] models the problem of maximizing health information coverage as an RMAB problem and proposes a hierarchical policy which leverages the structural assumptions of the RMAB model. [1] provide a solution for the class of bandits with "controlled restarts" and state-independent policies, possessing the indexability property. [21] model a health intervention problem, assuming that the uncertainty about the state collapses when an intervention is provided. They provide an algorithm called *Threshold Whittle* to compute the Whittle indices for infinite horizon RMAB. There are many other papers that provide Whittle indexability results for different subclasses of Partially Observable Markovian bandits [7, 9, 19, 31]. However, these papers focus on infinite horizon, whereas we focus on the more challenging setting when there is a fixed finite horizon.

The RMAB problem with finite horizon has been comparatively less studied. [27] provided solutions to the one-armed restless bandit problem, where only one arm is activated at each time before a time horizon T. Their solution do not directly extend to the scenario when multiple arms can be pulled at each time step. [10] considered finite horizon multi-armed restless bandits with identically distributed arms. They show that an index based policy based on the Lagrangian relaxation of the RMAB problem, similar to the infinite horizon setting, provides a near-optimal solution. [16] study the problem of selecting patients for early-stage cancer screening, by formulating it as a very restricted subclass of RMAB. All these works consider that all the arms are available throughout T time steps. Some other works, such as [8, 24] also adopt different approaches to decomposing the bandit arms, which may be applicable to finite horizon RMABs. These techniques to solving weakly coupled Markov Decision Processes are more general, but consequently less efficient than the Whittle Index approach in settings where indexability assumption holds.

The S-RMAB problem has been studied in a more restricted setting by [34]. They assume that, at each time step, arms may randomly arrive and depart due to random abandonment. However, the main limitation of their solution is the assumption that all arms have the same state-transition dynamics. This assumption does not hold in most of the real-world instances and thus, in this paper, we consider heterogeneous arms—arms are allowed to have their own transition dynamics. We show empirically that our algorithms perform well even with heterogeneous arms.

Another related category of work studied *sleeping arms* for the *stochastic multi-armed bandits* (SMAB) problem, where the arms are allowed to be absent at any time step [4, 12, 14]. However, the SMAB is different from RMAB because, in the former, when an arm is activated, a reward is drawn from a Bernoulli reward distribution (and not dependent on any state-transition process). Thus, the algorithms and analysis of SMABs do not translate to the RMAB setting.

3 STREAMING BANDITS

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 arms. Each arm i of the S-RMAB is a Partially Observable Markov Decision Process (POMDP)—represented by a 4-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$. $\mathcal{S} = \{0, 1\}$ denotes the state space of the POMDP, representing the "bad" state (say, patient not adhering to the health program) and "good" state (patient adhering), respectively. $\mathcal{A}$ is the action space, consisting of two actions $\mathcal{A} = \{a, p\}$ where an action a (or, p), denotes the active (or, passive) action. The state $s \in \mathcal{S}$ of the arm, transitions according to a known transition function, $P^{a,i}_{s,s'}$ if the arm is pulled and according to the known function, $P^{p,i}_{s,s'}$ otherwise. We also assume the transition function to conform to two natural constraints often considered in existing literature [18, 21]: (i) Interventions

should positively impact the likelihood of arms being in the good state, i.e. $P^a_{01} > P^p_{01}$ and $P^a_{11} > P^p_{11}$ and (ii) Arms are more likely to remain in the good state than to switch from the bad state to the good state, i.e. $P^a_{11} > P^a_{01}$ and $P^p_{11} > P^p_{01}$. Though the transitions probabilities are known to the planner, the actual state change is stochastic and is only partially observable—that is, when an arm is pulled, the planner discovers the true state of the arm; however, when the arm is not pulled, uncertainty about the true state persists. Under such uncertainties, it is customary to analyze the POMDP using its equivalent belief state MDP representation instead [21]. The state space of this MDP is defined by a set of all possible "belief" values that the arm can attain, denoted by $\mathcal{B}_i$. Each belief state $b \in \mathcal{B}_i$ represents the likelihood of the arm being in state 1 (good state). This likelihood is completely determined by the number of days passed since that arm was last pulled and the last observed state of the arm [18]. At each time step t, the planner accrues a state-dependent reward r_t from an active arm i, defined as:

$$r_t(i) = \begin{cases} 0 \text{ if } s_t(i) = 0 \text{ (arm } i \text{ is in the bad state at time } t) \\ 1 \text{ if } s_t(i) = 1 \text{ (arm } i \text{ is in the good state at time } t). \end{cases}$$

The total reward[1] of $R_t = \sum_{i\in[N]}(r_t(i))$ is accrued by the planner at time t, which is the sum of individual rewards obtained from the available arms. The planner's goal is to maximize her total reward collected, $\bar{R} := \sum_{t\in[T]} R_t$. This reward criterion is motivated by our applications in the healthcare intervention domain: interventions here correspond to reminding patients to adhere to their medication schedules and the good and bad states refer to patients either adhering or not adhering. The planner's goal is to maximize the expected number of times that all patients in the program adhere to their medication schedules. However, due to the limited budget, the planner is constrained to pull at most k arms per time step. Assuming a set of N arms, the problem then boils down to determining a policy, $\pi : \mathcal{B}_1 \times \ldots \mathcal{B}_N \rightarrow \mathcal{A}^N$ which governs the action to choose on each arm given the belief states of arms, at each time step, maximizing the total reward accumulated across T time steps.

Contrary to previous approaches that typically consider arms to all arrive at the beginning of time and stay forever, in this paper 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 the health program for a patient, 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,

$$\bar{R} := \sum_{t\in[T]} \sum_{i\in[N(t)]} r_t(i). \tag{1}$$

[1]For a natural number N, we use the notation $[N] := \{1, \ldots N\}$.

4 METHODOLOGY

The dominant paradigm for solving RMAB problems is the Whittle index approach. The central idea of the Whittle approach is to decouple the RMAB arms and then compute indices for each arm that capture the "value" of pulling that arm. The Whittle Index policy then proceeds by pulling the k arms with the largest values of Whittle Index. This greedy approach makes the time complexity linear in the number of arms, as indices can be computed independently for each arm. The computation of the index hinges on the notion of a "passive subsidy" m, which is the amount rewarded to the planner for each arm kept passive, in addition to the usual reward collected from the arm. The Whittle Index for an arm is defined as the infimum value of subsidy, m that must be offered to the planner, so that the planner is indifferent between pulling and not pulling the arm. To formalize this notion, consider an arm of the bandit in a belief state b. Its active and passive value functions, under a discount factor of $beta$, and when operating under a passive subsidy m, can be written as:

$$V^p_{m,T}(b) = b + m + \beta V_{m,T-1}(bP^p_{11} + (1-b)P^p_{01}) \tag{2}$$

$$V^a_{m,T}(b) = b + \beta b V_{m,T-1}(P^a_{11}) + \beta(1-b)V_{m,T-1}(P^a_{01}) \tag{3}$$

The value function for the belief state b is $V_{m,T}(b) = \max\{V^p_{m,T}(b), V^a_{m,T}(b)\}$. The Whittle Index for the belief state b, with a residual lifetime T is defined as: $\inf_m\{m : V^p_{m,T}(b) == V^a_{m,T}(b)\}$. The Whittle Index approach is guaranteed to be asymptotically optimal when a technical condition called *indexability* holds for all the arms. Intuitively, indexability requires that if for some passive subsidy m, the optimal action on an arm is passive, then $\forall m' > m$, the optimal action should still remain passive. Equivalently, indexability can be expressed as: $\frac{\partial}{\partial m}V^p_{m,T}(b) \geq \frac{\partial}{\partial m}V^a_{m,T}(b)$.

In this section we first show theoretically that the Streaming Bandit setup is indexable (subsection 4.1). Next, in subsection 4.2, we observe and formalize a useful phenomenon about the Whittle Index in the finite horizon setting. We use this phenomenon to design fast algorithms for S-RMABs in subsection 4.3 and we provide runtime complexity analysis for the same in subsection 4.4. Finally in subsection 4.5 we identify cases beyond those identified by previous work to which our efficient algorithm extends.

4.1 Conditions for indexability of streaming bandits

In this section, we extend the conditions for indexability that [21] originally established for infinite horizon, to the finite horizon setting of Streaming bandits. To show indexability, we first show in Theorem 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 2).

Definition 1 (Threshold Optimality [21]). *An RMAB instance is called threshold optimal if either a forward threshold policy or a*

reverse threshold policy is optimal. A forward (or reverse) threshold policy π is optimal if there exists a threshold b^* such that it is optimal to take a passive (or active) action whenever the current belief of the arm is greater than b^*, that is, $\pi(b) = 0$ (or $\pi(b) = 1$) whenever $b > b^*$ and $\pi(b) = 1$ (or $\pi(b) = 0$) whenever $b \leq b^*$.

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. Using the result that indexability holds whenever threshold optimality is satisfied [21], we imply that the Streaming Bandits problem is indexable whenever threshold optimality on the underlying infinite horizon problem is satisfied.

Theorem 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. Consider a streaming arm, that arrives (or, becomes available to the system) at time step T_1 and exits (or, becomes unavailable) at time step T_2. To capture the arm's arrival and departure in the belief model, we construct a new belief model with each state represented by a tuple $\langle$ `behavior, time-step` $\rangle$, where `behavior` takes a belief value in the interval $(0, 1)$ or is set to U (unavailable). U can be set to any constant value (such as $U = 0$). The transition probabilities are constructed as follows:

- The first $T_1 - 1$ states represent the unavailability of the arm and have deterministic transitions, i.e., for an action a, $P^a_{\langle U,t-1\rangle,\langle U,t\rangle} = 1$ for all $t \in \{2, \ldots, T_1 - 1\}$.
- At time T_1, the arm can either be in good state or bad state, so we create two states $\langle 1, T_1\rangle$ and $\langle 0, T_1\rangle$. For each $x \in \{0, 1\}$, $P^a_{\langle U,T_1-1\rangle,\langle x,T_1\rangle} = p_x$ where p_x represents the probability that the arm starts at a good (1) or bad (0) state. Note that, in our experiments, we assume that the initial state of an arm is fixed to 0 or 1, and can be captured by using either $p_x = 0$ or $p_x = 1$, respectively.
- For each time step $t \in \{T_1 + 1, T_2 - 1\}$, we create $2(t - T_1 + 1)$ states: $\langle b_w(0), t\rangle, \ldots, \langle b_w(t - T_1), t\rangle$ for each action $w \in \{0, 1\}$. For any $t', t'' \in \{0, 1, \ldots, t - T_1\}$, the probability of transitioning from the state $\langle b_w(t'), t - 1\rangle$ to the state $\langle b_w(t''), t + 1\rangle$ is same as the probability of changing from belief value $b_w(t')$ to $b_w(t'')$ in one time step on taking action w.
- For time step $t \geq T_2$, we create one sink state $\langle U, T_2\rangle$. This state represents unavailability of the arm subsequent to time step $T_2 - 1$. For any $t' \in \{0, 1, \ldots, T_2 - T_1\}$, the probability of transitioning from $\langle b_w(t'), T_2\rangle$ to $\langle U, T_2\rangle$ is 1.

Thus, the number of states in the new belief network is:

$$\begin{aligned} & T_1 - 1 + 2(1 + \ldots + (T_2 - T_1)) + 1 \qquad (4)\\ = \;& T_1 + (T_2 - T_1)(T_2 - T_1 + 1)\\ = \;& T_2 + (T_2 - T_1)^2 \end{aligned}$$

Thus, $T_2 + (T_2 - T_1)^2$ states are required for converting a belief network representing 2-state streaming bandits problem to a classic RMAB problem. □

Lemma 1. *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. First, we define the value function for the modified belief states.

$$V^p_m(\langle b, t\rangle) = \begin{cases} b + m + \beta V_m(\langle bP^p_{11} + (1-b)P^p_{01}, t+1\rangle) & \text{if } b \neq U \\ b + m + V_m(\langle b', t+1\rangle) & \text{otherwise} \end{cases}$$

$$V^a_m(\langle b, t\rangle) = \begin{cases} b + \beta(V_m(\langle bP^a_{11}, t+1\rangle) + (1-b)V_m(\langle P^a_{01}, t+1\rangle)) & \\ \qquad \text{if } b \neq U & \\ b + V_m(\langle b', t+1\rangle) \qquad \text{otherwise} & \end{cases}$$

where b' is the next belief state.

The minimum value of m_U that makes the passive action as valuable as active action at the states $\langle U, t\rangle$ for $T_1 \leq t < T_2$, can be obtained by equating

$$\begin{aligned} V^p_{m_U}(\langle U, t\rangle) &= V^a_{m_U}(\langle U, t\rangle) && (5)\\ \Rightarrow U + m_U + V_{m_U}(\langle b', t+1\rangle) &= U + V_{m_U}(\langle b', t+1\rangle) && (6)\\ \Rightarrow m_U &= 0. && (7) \end{aligned}$$

Assuming that there exists a forward (or reverse) threshold policy, $m_U = 0$ implies that, even without any subsidy, passive action is as valuable as active action.

Further, we show in the Appendix that the minimum subsidy at any other belief state is greater than 0. As the belief states $b \neq U$ require a positive subsidy for the passive action to be optimal, while for the belief state U, passive is already optimal for a subsidy of zero, a policy that maximizes value while paying minimum subsidy, would never choose to set arms currently in the u state to active. □

Theorem 2. *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 [21]), we show that the Streaming bandit problem is also indexable. □

4.2 Index decay for finite horizons

In this section we describe a phenomenon called *index decay* which is observed considering short horizon. Here, the Whittle index values are low when the residual lifetime of an arm is 0 or 1. We formalize this observation in Theorem 3. We use this phenomenon as an anchor to develop our algorithm (detailed in 4.3). We proceed by stating one fact and proving one useful Lemma, building up towards the Theorem.

Fact 1. *For two linear functions, $f(x)$ and $g(x)$ of x, such that $f'(x) \geq g'(x)$, whenever $f(x_1) < g(x_1)$ and $f(x_2) = g(x_2)$, the following holds: $x_2 > x_1$.*

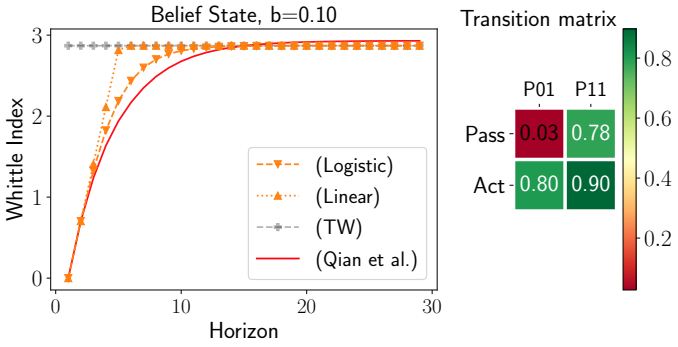


Figure 1: Whittle Indices for a belief state as computed by different algorithms. Both our algorithms capture index decay providing good estimates.

Lemma 2. *Consider an arm operating under a passive subsidy m. Assuming an initial belief state b_0, let $\rho^a(b_0,t)$ and $\rho^p(b_0,t)$ denote the probability of the arm being in the good state at time t $\forall t < T$ when policies $\pi^a(t)$ and $\pi^p(t)$ are adopted respectively, such that $\pi^a(0) = a$, $\pi^p(0) = p$, and $\pi^a(t) = \pi^p(t)$ $\forall t \in \{1,\ldots,T\}$. Then, $\rho^a(b_0,t) > \rho^p(b_0,t)$ $\forall t \in \{1,\ldots,T\}$.*

Theorem 3 (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 horizon T). Assuming indexability holds, we show that: $\forall T > 1$: $m_T > m_1 > m_0 = 0$.*

Proof. We provide our argument for a more general reward criterion than the total reward introduced in Section 3. Consider a discounted reward criterion with discount factor $\beta \in [0,1]$ (where $\beta = 1$ corresponds to total reward). m_0 is simply the m that satisfies: $V^p_{m,0}(b) = V^a_{m,0}(b)$ i.e., $b + m = b$, thus $m_0 = 0$. Similarly, m_1 can be solved by equating $V^p_{m_1,1}(b)$ and $V^a_{m_1,1}(b)$ and obtained as: $m_1 = \beta\Delta b = \beta\Big(\big(b\ P^a_{11} + (1-b)\ P^a_{01}\big) - \big(b\ P^p_{11} + (1-b)\ P^p_{01}\big)\Big)$

Using the natural constraints $P^a_{s1} > P^p_{s1}$ for $s \in \{0,1\}$, we obtain $m_1 > 0$.

Now, to show $m_T > m_1$ $\forall T > 1$, we first show that $V^a_{m_1,T}(b) > V^p_{m_1,T}(b)$. Combining this with the fact that $V_m(.)$ is a linear function of m and by definition, m_T is a point that satisfies $V^p_{m_T,T}(b) = V^a_{m_T,T}(b)$, we use Fact 1 and set $f = V^p_{m,T}(b)$, $g = V^a_{m,T}(b)$, $x_1 = m_1$ and $x_2 = m_T$ to obtain $m_1 < m_T$, and the claim follows. To complete the proof we now show that $V^a_{m_1,T}(b) > V^p_{m_1,T}(b)$.

Starting from an initial belief state b_0, let $\rho^p(b_0,t)$ be the expected belief for the arm at time t, if the passive action was chosen at $t = 0$ and the optimal policy, $\pi^p(t)$ was adopted for $0 < t < T$. Similarly let $\rho^a(b_0,t)$ be the expected belief at time t, if the active action was chosen at $t = 0$ and the *same* policy, $\pi^p(t)$ (which may not be optimal now) was adopted for $0 < t < T$. Then, $\beta\big(\rho^a(b_0,1) - \rho^p(b_0,1)\big) = m_1 > 0$ as shown above. Note that if we took actions according to $\pi^p(t)$ for $t \in \{1,\ldots,T-1\}$ with active action taken at the 0^{th} time step, the total expected reward so obtained is upper bounded by the active action value function, $V^a_{m_1,T}(b_0)$. Thus,

$$V^p_{m_1,T}(b_0) = b_0 + m_1 + \beta\rho^p(b_0,1) + \sum_{t=2}^{T}\beta^t\rho^p(b_0,t) + \Big(\sum_{t=1}^{T}\beta^t m_1.\mathbb{1}_{\{\pi^p(t)=passive\}}\Big) \tag{8}$$

$$= b_0 + \beta\rho^a(b_0,1) + \sum_{t=2}^{T}\beta^t\rho^p(b_0,t) + \Big(\sum_{t=1}^{T}\beta^t m_1.\mathbb{1}_{\{\pi^p(t)=passive\}}\Big)$$

$$< b_0 + \beta\rho^a(b_0,1) + \sum_{t=2}^{T}\beta^t\rho^a(b_0,t) + \Big(\sum_{t=1}^{T}\beta^t m_1.\mathbb{1}_{\{\pi^p(t)=passive\}}\Big) \tag{9}$$

(by Lemma 2)

$$\leq V^a_{m_1,T}(b_0)$$

□

4.3 Proposed algorithms

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.

Recall that we establish in Theorem 3 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:

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

For the linear interpolation, we assume $\hat{W}(h)$, our estimated Whittle Index, to be a piece-wise-linear function of 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 $\overline{W}$. 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

$$\hat{W}(h, \Delta b, \overline{W}) = \min\{h\ \Delta b, \overline{W}\}. \tag{10}$$

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

Algorithm 1: Interpolation Algorithm Template

1: 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 $\mathcal{B}_i$.
2: **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.
3: Initialize $\hat{W}_{N\times 1}$ to store estimated Whittle Indices.
4: **for** each arm i in N **do**
5: Let $b := \bar{b}_i, h := \bar{h}_i$ and let P be i's transition matrix.
6: Compute the myopic index Δb as: $\Delta b = \left(b\ P^a_{11} + (1-b)\ P^a_{01}\right) - \left(b\ P^p_{11} + (1-b)\ P^p_{01}\right)$.
7: Set $\hat{W}_i(h, \Delta b, \overline{W})$ according to one of the interpolation functions (10) or (11).
8: **end for**
9: Pull the k arms with the largest values of $\hat{W}$.

moderately higher rewards in many cases for a small additional compute time. For the logistic interpolation, we let

$$\hat{W}(h, \Delta b, \overline{W}) = \frac{C_1}{1 + e^{-C_2 h}} + C_3. \tag{11}$$

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 [29]. Figure 1 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.

Infinite horizon index: For transition matrices that satisfy the conditions for forward threshold policies to be optimal, Mate et al. [21] present an algorithm that computes $\overline{W}$ cheaply. The cornerstone of their technique is to leverage forward threshold optimality to map the passive and active actions to two different forward threshold policies, and find the value of subsidy m that makes the expected reward of the policies equal. We extend this reasoning to reverse threshold optimal arms.

4.4 Complexity analysis

For the complexity analysis of the algorithms, we denote by $\bar{X}$ the expected number of arms arriving each time step and $\bar{L}$ their average expected lifetimes. The expected number of arms at any point in time is then $O(\bar{X}\bar{L})$ [17]. Our algorithms (both versions) require a per-period cost of $O(\bar{X} * |\mathcal{B}_i| = \bar{X} * 2\bar{L})$ for the Threshold Whittle pre-computations, plus $O(\bar{X})$ computations for the myopic cost, plus $O(\bar{X}\bar{L} * \bar{L})$ calculations (for $\bar{X}\bar{L}$ arms, each requiring up to $\bar{L}$ additions or multiplications) and $O(\bar{X}\bar{L})$ for determining the top k indices. The overall per-period complexity of our algorithm is thus $O(\bar{X}\bar{L}^2)$.

For comparison, Qian et al. has a per-period complexity of $\approx O\left(\bar{X}\bar{L}^{(3+\frac{1}{18})}\log(\frac{1}{\epsilon})\right)$, where $\log(\frac{1}{\epsilon})$ is due to a bifurcation method for approximating the Whittle index to within error ϵ on each arm and $\bar{L}^{2+\frac{1}{18}}$ is due to the best-known complexity of solving a linear program with $\bar{L}$ variables [11] .

4.5 Reverse Threshold Arms

Computing the infinite horizon Whittle index cheaply ($\overline{W}$) is key to the runtime efficiency of our approach. Existing methods provide techniques to compute $\overline{W}$ used in the previous subsection, when the transition matrices satisfy the forward threshold optimality conditions. In this subsection, we describe how the technique can be extended to the case when reverse threshold optimality conditions are satisfied.

All the belief states that an arm can ever visit during its lifetime L can be enumerated and organized into two chains — each chain corresponding to one of the two possible observations ($\omega \in \{0, 1\}$) last observed for that arm. These chains are shown in Figure 2. [21] present an algorithm to compute the index for forward threshold arms with belief states belonging to the NIB process (i.e. whenever $b > b_{stationary} = \frac{P^p_{01}}{P^p_{01}+P^p_{10}}$). The algorithm relies on mapping the active and passive actions to two different forward threshold policies (with corresponding threshold states on the two chains indexed as X_0, X_1) and equating the policies' rewards to solve for the passive subsidy m, that makes the two actions equal.

We extend this reasoning to reverse threshold arms with belief chains belonging to the $\omega = 0$ chain of the SB (split-belief) process, as shown in Figure 2. The belief states belonging to the increasing chain ($\omega = 0$ chain) satisfy $b < b_{stationary} = \frac{P^p_{01}}{P^p_{01}+P^p_{10}}$. We identify two different reverse threshold policies that correspond to the active and passive actions, which can be used to set up similar indifference equations. For a given belief state on the increasing chain with index in the chain X, the corresponding reverse threshold policies can be indexed by $(X_0, X_1) = (1, X)$ and $(X_0, X_1) = (1, X + 1)$ and used to solve for the whittle index using the indifference equation outlined in Algorithm 1 of [21].

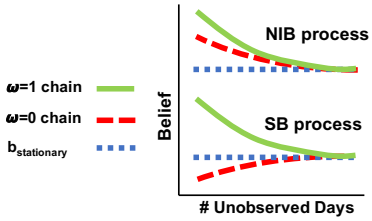
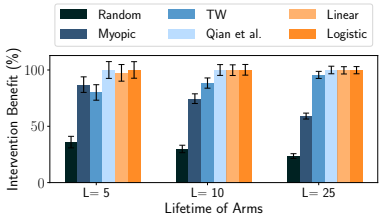


Figure 2: Belief values arranged in chains as presented in [21]. For every possible last observed state of the arm, ω, there is a corresponding chain of belief states.

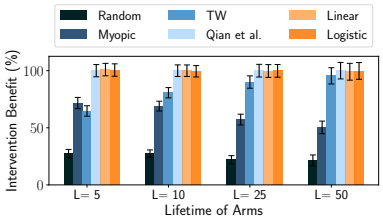
5 EXPERIMENTAL EVALUATION

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. Our main baselines are: (1) a precise, but slow algorithm by QIAN ET AL., 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 [21] (marked as TW), a much faster algorithm, that is only designed to work for infinitely long residual time horizons. MYOPIC policy is a popularly used baseline [18, 21, 29] 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.

Performance is 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 [29] equal to 100% and can be obtained for an algorithm ALG as: $\frac{100\times(\overline{R}^{\text{ALG}}-\overline{R}^{\text{No intervention}})}{\overline{R}^{\text{Qian et al.}}-\overline{R}^{\text{No intervention}}}$ 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) Arms arriving synchronously



(b) Arms arriving asynchronously as a stochastic process

Figure 3: (a) Performance of Threshold Whittle algorithm degrades when the lifetime of arms gets shorter, even when all arms start synchronously (b) The performance dwindles further if arms arrive asynchronously.

5.1 Real domain: Monitoring tuberculosis medication adherence

We first test on an anonymized real-world data set used by [13], consisting of daily adherence data of tuberculosis patients in Mumbai, India following a prescribed treatment regimen for six months. For our study, we only obtain the summary statistics capturing the transition probabilities of these patients moving between the adherent and non-adherent states as extracted from the dataset. We then follow the same data imputation steps adopted by [21] for arriving at the transition matrices, $P^a_{ss'}$ and $P^{\hat{P}}_{ss'}$ for each patient. We sample transition matrices from this real-world patient distribution and run simulations over a simulation length much longer than the lifetimes of the patients in the simulation.

In Figure 3a, we first demonstrate the impact of a short horizon alone on the performance of various algorithms in a simple, non-streaming setting. In Figure 3b, we contrast this with a similar comparison for the short horizon setting combined with a stochastic incoming stream of patients.

In Figure 4, we again consider the finite horizon setting with a deterministic incoming stream of patients. In Figure 4a, we plot the runtimes of our algorithms and that of Qian et al., as a function of the daily arrival rate, $\bar{X}$ of the incoming stream. Figure 4b measures the intervention benefits of these algorithms for these values of $\bar{X}$. The lifetime of each arm, L is fixed to 5 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 days. The runtime of Qian et al. 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. Other competing fast algorithms like Threshold Whittle, which assume an infinite residual horizon, 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 4c, 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. The details about the other parameters are deferred to the appendix. We also study the isolated effects of small lifetimes and asynchronous arrivals separately as well as performance in settings with stochastic arrivals, in the appendix. 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 on par with Qian et al., in all of the above.

5.2 Real domain: ARMMAN for improving maternal healthcare

Considering an alternate real-world domain, we again only use summary statistics (transition probabilities) from an application domain consisting of intervention planning for improving maternal healthcare [3]. Individuals (arms) are labeled to be in one of three states at any time step, of which one is the good state. [22] cast the problem as an RMAB with 2-state MDP on each arm. We also focus on maximizing the number of individuals in the good state, merging the other two states from the data into a single bad state. The data set consists of three types of transition matrices for different groups, only one of which satisfies the constraints mentioned in Section 3 and is used in our subsequent analysis, which is otherwise analogous to Section 5.1. Figure 5a establishes similar large runtime gains achieved by our algorithm as against other baselines, while

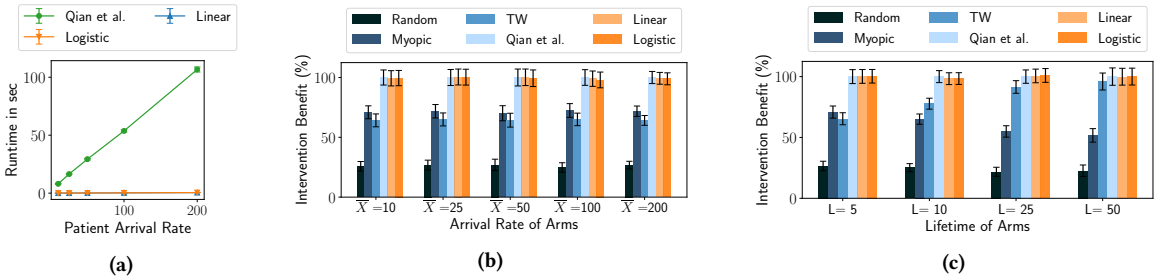


Figure 4: (a) Linear and Logistic interpolation algorithms are nearly 200× faster than Qian et al. (b) & (c) The interpolation algorithms achieve the speedup without sacrificing on performance, while other fast algorithms like Threshold Whittle deteriorate significantly for small residual horizons.

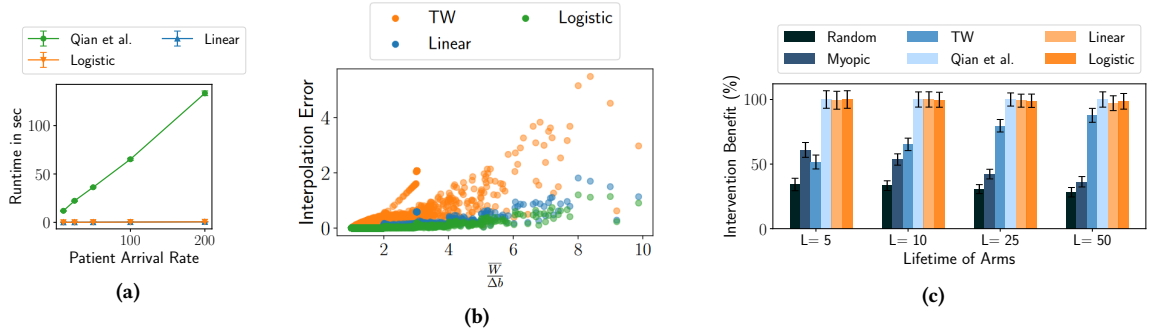


Figure 5: (a) The interpolation algorithms achieve a speedup of about 250× over baselines.(b) The error between the actual and estimated indices is largest for TW and lower for our interpolation algorithms (c) The good performance is maintained even for reverse threshold optimal arms.

maintaining similar performance figures in this domain. In the supplementary material we also present more details and analyses of the performance of our algorithms and baselines for this domain.

5.3 Synthetic domains

Finally, in this section, we test our algorithms on synthetic domains. We identify corner cases where our solutions do poorly and construct adversarial domains based on those. The ratio between the infinite horizon Whittle Index $\bar{W}$ and the myopic index Δb is an important driver of the approximation quality of our algorithms. The linear interpolation takes $\frac{\bar{W}}{\Delta b}$ steps to reach the finite horizon value, hence the higher this ratio is, higher the potential for approximation errors. In figure 5b we sum the approximation error over this interval $\epsilon := \sum_{h=1}^{h=\frac{\bar{W}}{\Delta b}} (\|\hat{W}(h) - W_{Qian}(h)\|)$ and plot it for different ratios $\frac{\bar{W}}{\Delta b}$. As expected, the approximation error increases with $\frac{\bar{W}}{\Delta b}$. We construct an adversarial domain by simulating cohorts with varying proportions of such patients. The results in the supplementary material show the intervention benefit of our algorithms decreases but remains within one standard error of Qian et al.

In Figure 5c, we simulate a population consisting of reverse threshold optimal patients exclusively and show similar good performance even though the previous theoretical guarantees of Threshold Whittle apply to forward threshold optimal patients only. In the supplementary material, we test multiple synthetic domains by varying the proportion of forward threshold optimal patients. In addition, we perform several other robustness checks varying important problem parameters and find that the run time and strong performance of our algorithms remains consistent across the board.

6 CONCLUSION

We study *streaming bandits*, or S-RMAB, a class of bandits where heterogeneous arms arrive and leave asynchronously under possibly random streams. While efficient RMAB algorithms for computing Whittle Indices for infinite horizon settings exist, for the finite horizon settings however, these algorithms are either comparatively costly or not suitable for estimating the Whittle Indices accurately. To tackle this, we provide a new scalable approach that allows for efficient computation of the Whittle Index values for finite horizon restless bandits while also adapting to more general S-RMAB settings. Our approach leverages a phenomenon called *index decay* to compute the indices for each arm. Through an extensive set of experiments on real-world and synthetic data, we demonstrate that our approach provides good estimates of Whittle Indices, and yield over 200× runtime improvements without loss in performance.

ACKNOWLEDGMENTS

This work was supported in part by the Army Research Office by MURI grant number W911NF1810208. A.B. and C.S. were supported by the Harvard Center for Research on Computation and Society.

REFERENCES

[1] N. Akbarzadeh and A. Mahajan. 2019. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *2019 IEEE Conference on Decision and Control*. IEEE.

[2] Biswarup Bhattacharya. 2018. Restless bandits visiting villages: A preliminary study on distributing public health services. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–8.

[3] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. 2021. Learn to Intervene: An Adaptive Learning Policy for Restless Bandits in Application to Preventive Healthcare. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*.

[4] Arpita Biswas, Shweta Jain, Debmalya Mandal, and Y Narahari. 2015. A truthful budget feasible multi-armed bandit mechanism for crowdsourcing time critical tasks. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. 1101–1109.

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

[6] Alicia H Chang, Andrea Polesky, and Gulshan Bhatia. 2013. House calls by community health workers and public health nurses to improve adherence to isoniazid monotherapy for latent tuberculosis infection: a retrospective study. *BMC public health* 13, 1 (2013), 894.

[7] K.D. Glazebrook, D. Ruiz-Hernandex, and C. Kirkbride. 2006. Some indexable families of restless bandit problems. *Adv. Appl. Probab* (2006), 643–672.

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

[9] Y. Hsu. 2018. Age of information: Whittle index for scheduling stochastic arrivals. In *2018 IEEE International Symposium on Information Theory*. IEEE.

[10] Weici Hu and Peter Frazier. 2017. An asymptotically optimal index policy for finite-horizon restless bandits. *arXiv preprint arXiv:1707.00205* (2017).

[11] S. Jiang, Z. Song, O. Weinstein, and H. Zhang. 2020. Faster dynamic matrix inverse for faster lps. *arXiv preprint arXiv:2004.07470* (2020).

[12] Varun Kanade, H Brendan McMahan, and Brent Bryan. 2009. Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *Artificial Intelligence and Statistics*. PMLR, 272–279.

[13] J. A. Killian, B. Wilder, A. Sharma, V. Choudhary, B. Dilkina, and M. Tambe. 2019. Learning to Prescribe Interventions for Tuberculosis Patients using Digital Adherence Data. In *KDD*.

[14] Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. 2010. Regret bounds for sleeping experts and bandits. *Machine learning* 80, 2 (2010), 245–272.

[15] Jerome Le Ny, Munther Dahleh, and Eric Feron. 2008. Multi-UAV dynamic routing with partial observations using restless bandit allocation indices. In *2008 American Control Conference*. IEEE, 4220–4225.

[16] Elliot Lee, Mariel S Lavieri, and Michael Volk. 2019. Optimal screening for hepatocellular carcinoma: A restless bandit model. *Manufacturing & Service Operations Management* 21, 1 (2019), 198–212.

[17] John DC Little. 1961. A proof for the queuing formula: L= λ W. *Operations research* 9, 3 (1961), 383–387.

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

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

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

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

[22] Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. 2021. Field Study in Deploying Restless Multi-Armed Bandits: Assisting Non-Profits in Improving Maternal and Child Health. *arXiv preprint arXiv:2109.08075* (2021).

[23] Aditya Mate, Andrew Perrault, and Milind Tambe. 2021. Risk-Aware Interventions in Public Health: Planning with Restless Multi-Armed Bandits. In *International Conference on Autonomous Agents and Multiagent Systems*.

[24] Nicolas Meuleau, Milos Hauskrecht, Kee-Eung Kim, Leonid Peshkin, Leslie Pack Kaelbling, Thomas L Dean, and Craig Boutilier. 1998. Solving very large weakly coupled Markov decision processes. In *AAAI/IAAI*. 165–172.

[25] Christopher Mundorf, Arti Shankar, Tracy Moran, Sherry Heller, Anna Hassan, Emily Harville, and Maureen Lichtveld. 2018. Reducing the risk of postpartum depression in a low-income community through a community health worker intervention. *Maternal and child health journal* 22, 4 (2018), 520–528.

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

[27] José Nino-Mora. 2011. Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing* 23, 2 (2011), 254–267.

[28] Christos H Papadimitriou and John N Tsitsiklis. 1994. The complexity of optimal queueing network control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*. IEEE, 318–322.

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

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

[31] B. Sombabu, A. Mate, D. Manjunath, and S. Moharir. 2020. Whittle index for AoI-aware scheduling. In *IEEE International Conference on Communication Systems & Networks (COMSNETS)*. IEEE.

[32] Edward J Sondik. 1978. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations research* 26, 2 (1978), 282–304.

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

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

SUPPLEMENTARY MATERIAL FOR: "EFFICIENT ALGORITHMS FOR FINITE HORIZON AND STREAMING RESTLESSMULTI-ARMED BANDIT PROBLEMS", AAMAS 2022

7 PROOFS

Theorem 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 a restless bandit arm with $T_2 + (T_2 - T_1)^2$ states.*

Let us consider that a streaming arm, that arrives (or, becomes available to the system) at time step T_1 and exits (or, becomes unavailable) at time step T_2. For including their arrival and departure in the belief model, we construct a new belief model with each state represented by a tuple $\langle$ `behavior, time-step` $\rangle$, where `behavior` takes a belief value in the interval $(0, 1)$ or is set to U (unavailable). U can be set to any constant value. The transition probabilities are constructed as follows:

- The first $T_1 - 1$ states represent the unavailability of the arm and have deterministic transitions, i.e., for an action a, $P^a_{\langle U,t-1\rangle,\langle U,t\rangle} = 1$ for all $t \in \{2, \ldots, T_1 - 1\}$.
- At time T_1, the arm can either be in good state or bad state, so we create two states $\langle 1, T_1\rangle$ and $\langle 0, T_1\rangle$. For each $x \in \{0, 1\}$, $P^a_{\langle U,T_1-1\rangle,\langle x,T_1\rangle} = p_x$ where p_x represents the probability that the arm starts at a good (1) or bad (0) state. Note that, in our experiments, we assume that the initial state of an arm is fixed to 0 or 1, that can be captured by using either $p_x = 0$ or $p_x = 1$, respectively.
- For each time step $t \in \{T_1 + 1, T_2 - 1\}$, we create $2t$ states: $\langle b_w(0), t\rangle, \ldots, \langle b_w(t - T_1), t\rangle$ for each action $w \in \{0, 1\}$. For any $t', t'' \in \{0, 1, \ldots, t - T_1\}$, the probability of transitioning from the state $\langle b_w(t'), t - 1\rangle$ to the state $\langle b_w(t''), t + 1\rangle$ is same as the probability of changing from belief value $b_w(t')$ to $b_w(t'')$ in one time step on taking action w.
- For time step $t \geq T_2$, we create one sink state $\langle U, T_2\rangle$. This state represents that unavailability of the arm subsequent to time step $T_2 - 1$. For any $t' \in \{0, 1, \ldots, T_2 - T_1\}$, the probability of transitioning from $\langle b_w(t'), T_2\rangle$ to $\langle U, T_2\rangle$ is 1.

Thus, the new belief network contains the following number of states:

$$
\begin{aligned}
& T_1 - 1 + 2(1 + \ldots + (T_2 - T_1)) + 1 && (12)\\
= \; & T_1 + (T_2 - T_1)(T_2 - T_1 + 1) && (13)\\
= \; & T_2 + (T_2 - T_1)^2 && (14)
\end{aligned}
$$

Thus, $T_2 + (T_2 - T_1)^2$ states are required for converting a belief network representing 2-state streaming bandits problem to a classic RMAB problem. □

Lemma 1. *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. First, we define the value function for the modified belief states.

$$
V^p_m(\langle b, t\rangle) = \begin{cases} b + m + \beta V_m(\langle bP^p_{11} + (1-b)P^p_{01}, t+1\rangle) & \text{if } b \neq U \\ b + m + V_m(\langle b', t+1\rangle) & \text{otherwise} \end{cases}
$$

$$
V^a_m(\langle b, t\rangle) = \begin{cases} b + \beta(V_m(\langle bP^a_{11}, t+1\rangle) + (1-b)V_m(\langle P^a_{01}, t+1\rangle)) & \text{if } b \neq U \\ b + V_m(\langle b', t+1\rangle) & \text{otherwise} \end{cases}
$$

where b' is the next belief state.

The minimum value of m_U that makes the passive action as valuable as active action at the states $\langle U, t\rangle$ for $T_1 \leq t < T_2$, can be obtained by equating

$$
\begin{aligned}
V^p_{m_U}(\langle U, t\rangle) &= V^a_{m_U}(\langle U, t\rangle) && (15)\\
\Rightarrow U + m_U + V_{m_U}(\langle b', t+1\rangle) &= U + V_{m_U}(\langle b', t+1\rangle) && (16)\\
\Rightarrow m_U &= 0. && (17)
\end{aligned}
$$

Assuming that there exists a forward (or reverse) threshold policy, $m_U = 0$ implies that, even without any subsidy, passive action is as valuable as active action. To show that the passive action is optimal at the u states, we now show that the minimum subsidy at any other belief state is greater than 0. We show this by contradiction. Let us assume that the minimum subsidy $m_b = 0$ for a belief state $b \neq U$. Then,

$$
\begin{aligned}
& V^p_{m_b}(\langle b, t\rangle) \geq V^a_{m_b}(\langle b, t\rangle)\\
\Rightarrow \quad & b + m_b + \beta V_{m_b}(bP^p_{11} + (1-b)P^p_{01}) \geq\\
& b + \beta(V_{m_B}(bP^a_{11}) + (1-b)V_{m_b}(P^a_{01}))\\
\Rightarrow \quad & V_{m_b}(bP^p_{11} + (1-b)P^p_{01}) \geq\\
& V_{m_b}(bP^a_{11}) + (1-b)V_{m_b}(P^a_{01})\\
\Rightarrow \quad & V_{m_b}(bP^p_{11} + (1-b)P^p_{01}) >\\
& V_{m_b}(bP^p_{11}) + (1-b)V_{m_b}(P^p_{01})
\end{aligned}
$$

$\because P^a_{x1} > P^p_{x1}$ and $V_m(b)$ is non-decreasing (Corollary 1 in [21]).

The last inequality contradicts the fact that $V_m(b)$ is a convex function of b [32]. Hence, the minimum subsidy required at any belief state $b \neq U$ to make the passive action more valuable is strictly greater than 0. □

Theorem 2. *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 [21]), we show that the Streaming bandit problem is also indexable. □

Theorem 3 (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 horizon T).*

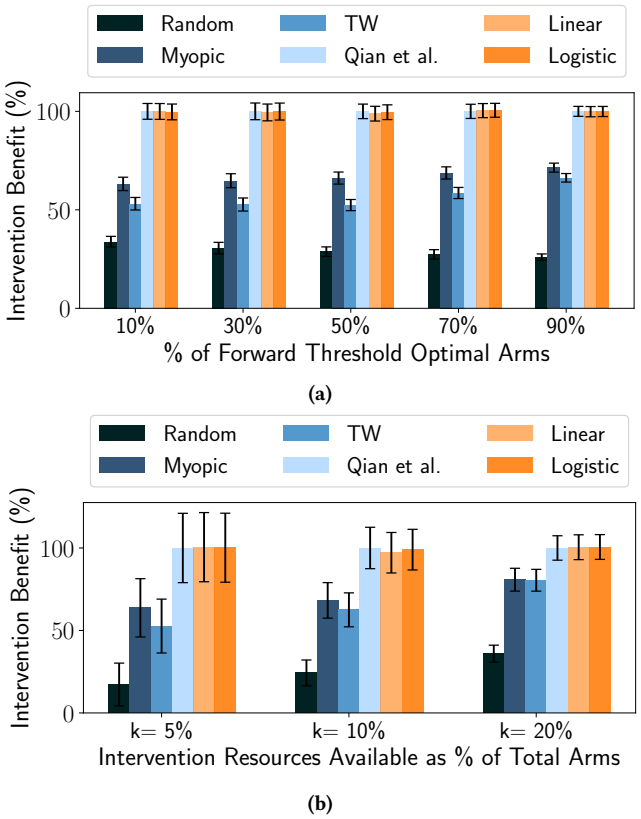


Figure 6: (a)Performance of our algorithm remains robust even when population is composed of a varying fraction of forward threshold optimal arms (b) Performance of our algorithm remains robust under varying levels of available resources

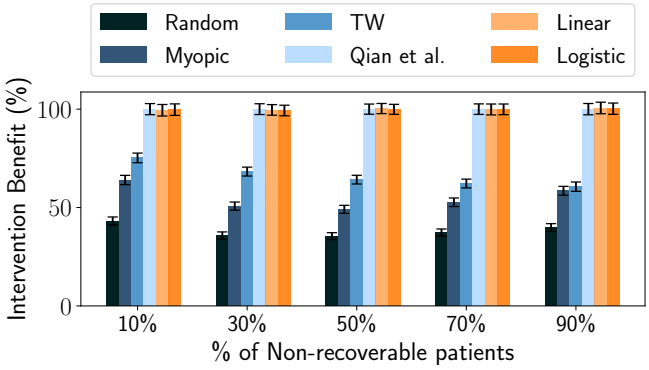


Figure 7: Non-recoverable patients are those that remain in the bad state with high probability, even after receiving an intervention. Performance of Threshold Whittle begins to dwindle when the fraction of non-recoverable patients in the cohort increases, but our interpolation algorithms remain robust.

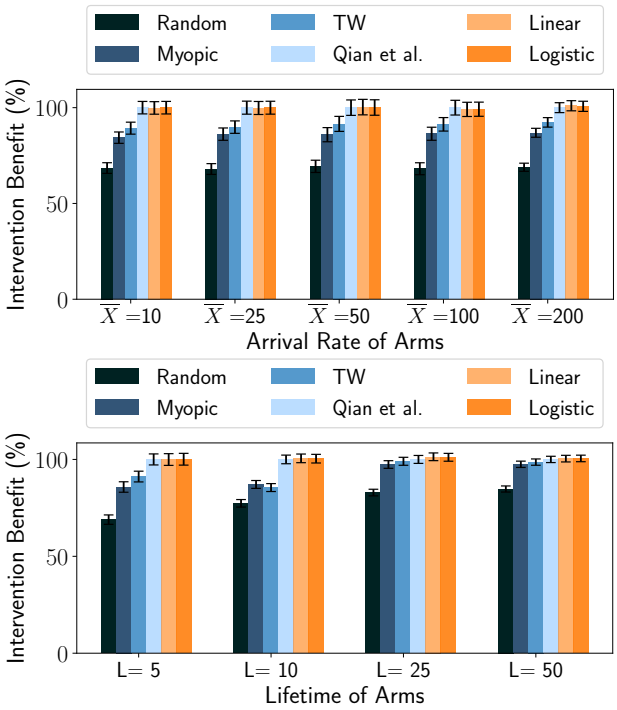


Figure 8: Tests on the ARMMAN domain reveal that the large speedup is achieved while virtually maintaining the same good quality of performance

Assuming indexability holds, we show that the Whittle index decays for short horizons: $\forall T > 1 \colon\ m_T > m_1 > m_0 = 0$.

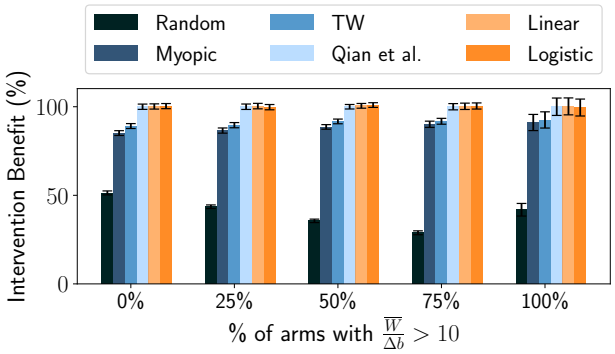
Proof. We provide our argument for a more general reward criterion than the total reward introduced in Section 3. Consider a discounted reward criterion with discount factor $\beta \in [0, 1]$ (where $\beta = 1$ corresponds to total reward). m_0 is simply the m that satisfies: $V^p_{m,0}(b) = V^a_{m,0}(b)$ i.e., $b + m = b$, thus $m_0 = 0$. Similarly, m_1 can be solved by equating $V^p_{m_1,1}(b)$ and $V^a_{m_1,1}(b)$ as follows:

$$\implies b + m_1 + \beta\big(bP^p_{11} + (1-b)P^p_{01}\big) = b + \beta\big(b(P^a_{11}) + (1-b)(P^a_{01})\big)$$
$$\implies m_1 = \beta\big(b(P^a_{11} - P^p_{11}) + (1-b)(P^a_{01} - P^p_{01})\big) \tag{18}$$

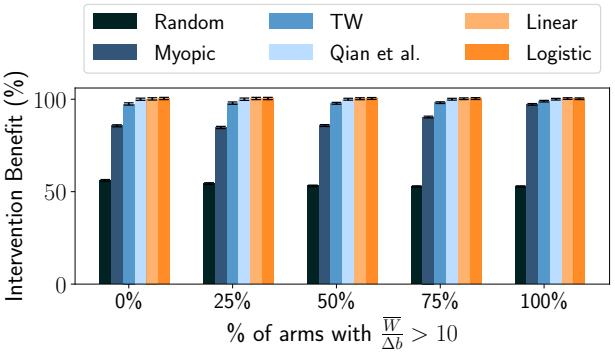
Using the natural constraints $P^a_{s1} > P^p_{s1}$ for $s \in \{0, 1\}$, we obtain $m_1 > 0$.

Now, to show $m_T > m_1\ \forall T > 1$, we first show that $V^a_{m_1,T}(b) > V^p_{m_1,T}(b)$. Combining this with the fact that $V_m(.)$ is a linear function of m and by definition, m_T is a point that satisfies $V^p_{m_T,T}(b) = V^p_{m_T,T}(b)$, we use Fact 1 and set $f = V^p_{m,T}(b)$, $g = V^a_{m,T}(b)$, $x_1 = m_1$ and $x_2 = m_T$ to obtain $m_1 < m_T$, and the claim follows. For completeness, we now show that $V^a_{m_1,T}(b) > V^p_{m_1,T}(b)$.

Starting from an initial belief state b_0, let $\rho^p(b_0, t)$ be the expected belief for the arm at time t, if the passive action was chosen at $t = 0$ and the optimal policy, $\pi^p(t)$ was adopted for $0 < t < T$. Similarly let $\rho^a(b_0, t)$ be the expected belief at time t, if the active action was chosen at $t = 0$ and the *same* policy, $\pi^p(t)$ (which may not be optimal now) was adopted for $0 < t < T$. Then,



(a) Arm lifetime = 10



(b) Arm lifetime = 40.

Figure 9: We generate corner cases consisting of varying proportion of patients with high value of $\frac{W}{\Delta b}$. We test the algorithms under two situations corresponding to lifetime of arms smaller/larger than the ratio $\frac{W}{\Delta b}$ and find that our algorithm still show good performance throughout.

$\rho^a(b_0, 1) - \rho^p(b_0, 1) = m_1 > 0$ as shown above. Note that if we took actions according to $\pi^p(t)$ for $t \in \{1, \ldots, T-1\}$ with active action taken at the 0^{th} time step, the total expected reward so obtained is upper bounded by the active action value function, $V^a_{m_1,T}(b_0)$. Thus,

$$V^p_{m_1,T}(b_0) = b_0 + m_1 + \beta\rho^p(b_0, 1) + \sum_{t=2}^{T} \beta^t \rho^p(b_0, t) \tag{19}$$

$$+ \Big(\sum_{t=1}^{T} \beta^t m_1 . \mathbb{1}_{\{\pi^p(t)=passive\}}\Big)$$

$$= b_0 + \beta\rho^a(b_0, 1) + \sum_{t=2}^{T} \beta^t \rho^p(b_0, t) + \Big(\sum_{t=1}^{T} \beta^t m_1 . \mathbb{1}_{\{\pi^p(t)=passive\}}\Big)$$

$$< b_0 + \beta\rho^a(b_0, 1) + \sum_{t=2}^{T} \beta^t \rho^a(b_0, t) + \Big(\sum_{t=1}^{T} \beta^t m_1 . \mathbb{1}_{\{\pi^p(t)=passive\}}\Big) \tag{20}$$

(by Lemma 2)

$$\leq V^a_{m_1,T}(b_0)$$

□

8 ROBUSTNESS CHECKS

We conduct several robustness checks by varying key parameters important for the simulation and confirm that the good results remain constant across various settings. We also simulate a few additional synthetic domains as described below.