Collapsing Bandits and Their Application to Public Health Interventions

Aditya Mate[*]
Harvard University
Cambridge, MA, 02138
aditya_mate@g.harvard.edu

Jackson A. Killian[*]
Harvard University
Cambridge, MA, 02138
jkillian@g.harvard.edu

Haifeng Xu
University of Virginia
Charlottesville, VA, 22903
hx4ad@virginia.edu

Andrew Perrault
Harvard University
Cambridge, MA, 02138
aperrault@g.harvard.edu

Milind Tambe
Harvard University
Cambridge, MA, 02138
milind_tambe@harvard.edu

Abstract

We propose and study Collapsing Bandits, a new restless multi-armed bandit (RMAB) setting in which each arm follows a binary-state Markovian process with a special structure: when an arm is played, the state is fully observed, thus "collapsing" any uncertainty, but when an arm is passive, no observation is made, thus allowing uncertainty to evolve. The goal is to keep as many arms in the "good" state as possible by planning a limited budget of actions per round. Such Collapsing Bandits are natural models for many healthcare domains in which health workers must simultaneously monitor patients *and* deliver interventions in a way that maximizes the health of their patient cohort. Our main contributions are as follows: (i) Building on the Whittle index technique for RMABs, we derive conditions under which the Collapsing Bandits problem is *indexable*. Our derivation hinges on novel conditions that characterize when the optimal policies may take the form of either "forward" or "reverse" threshold policies. (ii) We exploit the optimality of threshold policies to build fast algorithms for computing the Whittle index, including a closed form. (iii) We evaluate our algorithm on several data distributions including data from a real-world healthcare task in which a worker must monitor and deliver interventions to maximize their patients' adherence to tuberculosis medication. Our algorithm achieves a 3-order-of-magnitude speedup compared to state-of-the-art RMAB techniques, while achieving similar performance. The code is available at: https://github.com/AdityaMate/collapsing_bandits

1 Introduction

Motivation. This paper considers scheduling problems in which a 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 is unobserved 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 *monitoring tasks* which arise in many domains such as sensor/machine maintenance [12, 10, 1, 33], anti-poaching patrols [27], and especially healthcare. For example, nurses or community health workers are employed to monitor and improve the adherence of patient cohorts to medications for diseases like diabetes [24], hypertension [4], tuberculosis [28, 5]

[*] equal contribution.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

and HIV [17, 16]. Their goal is to keep patients adherent (i.e., in the "good" state) but a health worker can only intervene on (visit) a limited number of patients each day. Health workers can play a similar role in monitoring and delivering interventions for patient mental health, e.g., in the context of depression [21, 23] or Alzheimer's Disease [19].

We adopt the solution framework of *Restless Multi-Arm Bandits* (RMABs), a generalization of Multi-Arm Bandits (MABs) in which a planner may act on k out of N arms each round that each follow a Markov Decision Process (MDP). Solving an RMAB is PSPACE-hard in general [25]. Therefore, a common approach is to consider the Lagrangian relaxation of the problem in which the $\frac{k}{N}$ budget constraint is dualized. Solving the relaxed problem gives Lagrange multipliers which act as a greedy index heuristic, known as the Whittle index, for the original problem. Specifically, the *Whittle index policy* computes the Whittle index for each arm, then plays the top k arms with the largest indices. The Whittle index policy has been shown to be asymptotically optimal (i.e., $N \to \infty$ with fixed $\frac{k}{N}$) under a technical condition [34] and generally performs well empirically [3] making it a common solution technique for RMABs.

Critically, using the Whittle index policy requires two key components: (i) a fast method for computing the index and (ii) proving the problem satisfies a technical condition known as *indexability*. Without (i) the approach can be prohibitively slow, and without (ii) asymptotic performance guarantees are sacrificed [34]. Neither (i) nor (ii) are known for general RMABs. Therefore, to capture the scheduling problems addressed in this work, we introduce a new subclass of RMABs, *Collapsing Bandits*, distinguished by the following feature: when an arm is played, the agent fully observes its state, "collapsing" any uncertainty, but when an arm is passive, no observation is made and uncertainty evolves. We show that this RMAB subclass is more general than previous models and leads to new theoretical results, including conditions under which the problem is indexable and under which optimal policies follow one of two simple threshold types. We use these results to develop algorithms for quickly computing the Whittle index. In experiments, we analyze the algorithms' performance on (i) data from a real-world healthcare scheduling task in which our approach ties state-of-the-art performance at a fraction the runtime and (ii) various synthetic distributions, some of which the algorithm achieves performance comparable to the state of the art even outside its optimality conditions.

To summarize, our contributions are as follows: (i) We introduce a new subclass of RMABs, Collapsing Bandits, (ii) Derive theoretical conditions for Whittle indexability and for the optimal policy to be threshold-type, and (iii) Develop an efficient solution that achieves a 3-order-of-magnitude speedup compared to more general state-of-the-art RMAB techniques, without sacrificing performance.

2 Restless Multi-Armed Bandits

An RMAB consists of a set of N arms, each associated with a *two-action* MDP [26]. An MDP $\{\mathcal{S}, \mathcal{A}, r, P\}$ consists of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a state-dependent reward function $r : \mathcal{S} \to \mathbb{R}$, and a transition function P, where $P^a_{s,s'}$ denotes the probability of transitioning from state s to s' when action a is taken. An MDP *policy* $\pi : \mathcal{S} \to \mathcal{A}$ represents a choice of action to take at each state. We will consider both discounted and average reward criteria. The long-term *discounted reward* starting from state $s_0 = s$ is defined as $R^\pi_\beta(s) = E\left[\sum_{t=0}^{\infty} \beta^t r(s_{t+1} \sim T(s_t, \pi(s_t), s_{t+1}) | \pi, s_0 = s\right]$ where $\beta \in [0, 1)$ is the discount factor and actions are selected using π. To define average reward, let $f^\pi(s) : \mathcal{S} \to [0, 1]$ denote the *occupancy frequency* induced by policy π, i.e., the fraction of time spent in each state of the MDP. The *average reward* $\overline{R}^\pi$ of policy π be defined as the expected reward computed over the occupancy frequency: $\overline{R}^\pi = \sum_{s \in \mathcal{S}} f^\pi(s) r(s)$.

Each arm in an RMAB is an MDP with the action set $\mathcal{A} = \{0, 1\}$. Action 1 (0) is called the *active* (*passive*) action and denotes the arm being pulled (not pulled). The agent can pull at most k arms at each time step. The agent's goal is to maximize either her discounted or average reward across the arms over time. Some RMAB problems need to account for partial observability of states. It is sufficient to let the MDP state be the *belief state*: the probability of being in each latent state [15]. While intractable in general due to infinite number of reachable belief states, most partially observable RMABs studied (including our Collapsing Bandits) have polynomially many belief states due to a finite time horizon or other structures.

Related work RMABs have been an attractive framework for studying various stochastic scheduling problems since Whittle indices were introduced [36]. Because general RMABs are PSPACE-hard [25], RMAB studies usually consider restricted classes under which some performance guarantees can be derived. Collapsing Bandits form one such novel class that generalizes some existing results which we note in later sections. Liu and Zhao [20] develop an efficient Whittle index policy for a 2-state partially observable RMAB subclass in which the state transitions are unaffected by the actions taken and reward is accrued from the active arms only. Akbarzadeh and Mahajan [2] define a class of bandits with "controlled restarts," giving indexability results and a method for computing the Whittle index. However, "controlled restarts" define the active action as state independent, a stronger assumption than Collapsing Bandits which allow state-dependent action effects. Glazebrook et al. [10] give Whittle indexability results for three classes of restless bandits: (1) A machine maintenance regime with deterministic active action effect (we consider stochastic active action effect) (2) A switching regime in which the passive action freezes state transitions (in our setting, states always change regardless of action) (3) A reward depletion/replenishment bandit which deterministically resets to a start state on passive action (we consider stochastic passive action effect). Hsu [11] and Sombabu et al. [30] augment the machine maintenance problem from Glazebrook et al. [10] to include either i.i.d. or Markovian evolving probabilities of an active action having no effect, a limited form of state-dependent action. Meshram et al. [22] introduce Hidden Markov Bandits which, similar to our approach, consider binary state transitions under partial observability, but do not allow for state dependent rewards on passive arms. In sum, our Collapsing Bandits introduce a new, more general RMAB formulation than special subclasses previously considered. Qian et al. [27] present a generic approach for any indexable RMAB based on solving the (partially observable) MDPs on arms directly. Because we derive a closed form for the Whittle index, our algorithm is orders of magnitude faster.

3 Collapsing Bandits

We introduce *Collapsing Bandits* (CoB) as a specially structured RMAB with partial observability. In CoB, each arm $n \in \{1, \dots, N\}$ has binary latent states $\mathcal{S} = \{0, 1\}$, representing *bad* and *good* state, respectively. The agent acts during each of finite days $t \in 1, \dots, T$. Let $a_t \in \{0, 1\}^N$ denote the vector of actions taken by the agent on day t. Arm n is said to be *active* at t if $a_t(n) = 1$ and *passive* otherwise. The agent acts on k arms per day, i.e., $\|a_t\| = k$, where $k \ll N$ because resources are limited. When acting on arm n, the true latent state of n is fully observed by the agent and thus its uncertainty "collapses" to a realization of the binary latent state. We denote this observation as $\omega \in \mathcal{S}$. States of passive arms are completely unobservable by the agent.

Active arms transition according to the *transition matrix* $P^{a,n}_{s,s'}$ and passive arms transition according to $P^{p,n}_{s,s'}$. We drop the superscript n when there is no ambiguity. Our scheduling problem, like many problems in analogous domains, exhibits the following natural structure: (i) processes are more likely to stay "good" than change from "bad" to "good"; (ii) when acted on, they tend to improve. These natural structures are respectively captured by imposing the following constraints on P^p and P^a for each arm: (i) $P^p_{0,1} < P^p_{1,1}$ and $P^a_{0,1} < P^a_{1,1}$; (ii) $P^p_{0,1} < P^a_{0,1}$ and $P^p_{1,1} < P^a_{1,1}$. To avoid unnecessary complication through edge cases, all transition probabilities are assumed to be nonzero. The agent receives reward $r_t = \sum_{n=1}^{N} s_t(n)$ at t, where $s_t(n)$ is the latent state of arm n at t. The agent's goal is to maximize the long term rewards, either discounted or average, defined in Sec. 2.

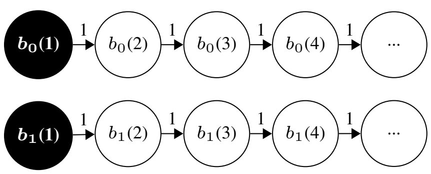


Figure 1: Belief-state MDP under the policy of always being passive. There is one chain for each observation $\omega \in \{0, 1\}$ with the head marked black. Belief states deterministically transition down the chains.

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. 1. Let $b_\omega(u)$ denote the belief state, *i.e., the probability that the state is* 1, if the agent received observation $\omega \in \{0, 1\}$ when it acted on the process u days 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.

When the belief-state MDP is allowed to evolve under some policy, the following mechanism arises: first, after an action, the state ω is observed (uncertainty "collapses"), then one round passes causing the agent's belief to become $P^a_{\omega,1}$, representing the head of the chain determined by ω. Subsequent passive actions cause the process to transition deterministically down the same chain (though, the transition in the latent state is still stochastic). Then when the process's arm is active, it transitions to the head of one of the chains with probability equal to the belief that the corresponding observation would be emitted (see Fig. 2a for an illustration).

The belief associated with a belief state can be calculated in closed form with the given transition probabilities. Formally,

$$b_\omega(u) = \tau_{u-1}(P^a_{\omega,1})\ \forall u \in [T] \text{ where } \tau_u(b) = \frac{P^p_{0,1} - (P^p_{1,1} - P^p_{0,1})^u (P^p_{0,1} - b(1 + P^p_{0,1} - P^p_{1,1}))}{(1 + P^p_{0,1} - P^p_{1,1})} \quad (1)$$

4 Collapsing Bandits: Threshold Policies and Whittle Indexability

Because of the well-known intractability of solving general RMABs, the widely adopted solution concept in the literature of RMABs is the Whittle index approach; for a comprehensive description, see Whittle [36]. 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 "donation" of reward. Formally, the modified reward function becomes $r_m : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, where $r_m(s, 0) = r(s) + m$ and $r_m(s, 1) = r(s)$. Let $R^\pi_{\beta,m}(s) = E\left[\sum_{t=0}^{\infty} \beta^t r_m(s_t, \pi(s_t)) | \pi, s_0 = s\right]$ and $\overline{R}^\pi_m = \sum_{s \in \mathcal{S}} f^\pi(s) r_m(s, \pi(s))$ be the discounted and average reward criteria for this new subsidy setting, respectively. The former is maximized by the discounted value function (we give a value function for the average reward criterion in **Fast Whittle Index Computation**):

$$V_m(b) = \max \begin{cases} m + b + \beta V_m(\tau_1(b)) & \text{passive} \\ b + \beta(b V_m(P^a_{1,1}) + (1 - b) V_m(P^a_{0,1})) & \text{active} \end{cases} \quad (2)$$

where τ is defined in Eq. 1 and b is shorthand for $b_\omega(u)$. In a CoB, the Whittle index of a belief state b is the smallest m s.t. it is equally optimal to be active or passive in the current state. Formally:

$$W(b) = \inf_m \{m : V_m(b; a = 0) \geq V_m(b; a = 1)\} \quad (3)$$

Critically, performance guarantees hold only if the problem satisfies *indexability* [34, 36], a condition which says that for all states, the optimal action cannot switch to active as m increases. Let Π^*_m be the set of policies that maximize a given reward criterion under subsidy m.

Definition 1 (Indexability). *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.*

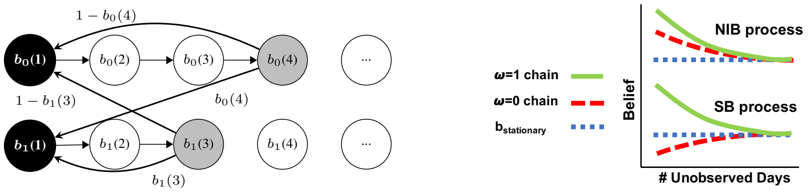
The following special type of MDP policy is central to our analysis.

Definition 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 1. *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 [20]. Our proof features a sophisticated induction argument exploiting the finite size of $\mathcal{B}$ and relies on tools from real analysis for limit arguments.

□



(a) (b)

Figure 2: (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. (b) Non-increasing belief (NIB) process has non-increasing belief in both chains. A split belief process (SB) has non-increasing belief after being observed in state 1, but non-decreasing belief after being observed in state 0.

All formal proofs can be found in the appendix. We remark that Thm. 1 generalizes the result in the seminal work by Liu and Zhao [20] who proved the indexability for a special class of CoB. In particular, the RMAB in Liu and Zhao [20] can be viewed as a CoB setting with $P^a = P^p$, i.e., transitions are independent of actions.

Though the Whittle index is known to be challenging to compute in general [36], we are able to design an algorithm that computes the Whittle index efficiently assuming the optimality of threshold policies, which we now describe.

Fast Whittle Index Computation 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 $X_\omega^{b_{th}} \in 1, \dots, T$ 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. 2a 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:

$$f^{(X_0,X_1)}(b_\omega(u)) = \begin{cases} \alpha & \text{if } \omega = 0, u \leq X_0 \\ \beta & \text{if } \omega = 1, u \leq X_1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$\alpha = \left(\frac{(X_1 b_0(X_0))}{1 - b_1(X_1)} + X_0 \right)^{-1}, \beta = \left(\frac{X_1 b_0(X_0)}{1 - b_1(X_1)} + X_0 \right)^{-1} \frac{b_0(X_0)}{1 - b_1(X_1)} \tag{5}$$

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

$$J_m^{(X_0,X_1)} = \sum_{b \in \mathcal{B}} b f^{(X_0,X_1)}(b) + m(1 - f^{(X_0,X_1)}(b_1(X_1)) - f^{(X_0,X_1)}(b_0(X_0))) \tag{6}$$

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. Such policies must have adjacent belief states as thresholds, as can be concluded from Lemma 1 in Appendix 7. 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 thus 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 two fast Whittle index algorithms.

Sequential index computation algorithm Alg. 1 precomputes the Whittle index of every belief state for each process, having time complexity $\mathcal{O}(|\mathcal{S}|^2 TN)$. 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. [27] (our main competitor) 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 [13].

Alg. 1 is optimized for settings in which the Whittle index can be precomputed. However, for online learning settings, we give an alternative method in Appendix 12 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
```

Our algorithm also requires that belief is decreasing in X_0 and X_1. Formally, we require:

Definition 3 (Non-increasing belief (NIB) processes)**.** *A process has* non-increasing belief *if, for any $u \in [T]$ and for any $\omega \in \mathcal{S}$, $b_\omega(u) \geq b_\omega(u+1)$.*

All possible CoB belief trends are shown in Fig. 2b. We make this distinction because the computation of the Whittle index in Alg. 1 is guaranteed to be exact for NIB processes that are also forward threshold optimal, though we show empirically that our approach works surprisingly well for most distributions. In the next section, we analyze the possible forms of optimal policies to find conditions under which threshold policies are optimal.

Types of Optimal Policies Analyzing Eq. 2 reveals that at most three types of optimal policies exist. This follows directly from the definition of $V_m(b)$, which is a max over the passive action value function and the active action value function. The former is convex in b, a well-known POMDP result [31], and the latter is linear in b. Thus, as shown in Fig. 3, there are three ways in which the value functions of each action may intersect; this defines three optimal policy forms of *forward*, *reverse* and *dual* threshold types, respectively. Forward and reverse threshold policies are defined in Def. 2; dual threshold policies are active between two separate threshold points and passive elsewhere. Not only do threshold policies greatly reduce the optimization search space, they often admit closed form expressions for the index as demonstrated earlier in this section. We now derive sufficient conditions on the state transition probabilities under which each type of policy is verifiably optimal.

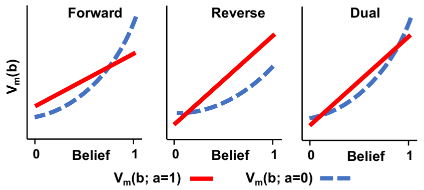


Figure 3: Components of $V_m(b)$ in Eq. 2. Since the passive action is convex in b, active action is linear in b, and value function is a max over these, at most three optimal policy types are possible.

Theorem 2. *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:*

$$(P^p_{1,1} - P^p_{0,1})(1 + \beta(P^a_{1,1} - P^a_{0,1}))(1 - \beta) \geq P^a_{1,1} - P^a_{0,1} \tag{7}$$

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 $\tau(b)$ and using induction to derive both upper and lower bounds on $V_m(b_1) - V_m(b_2)\ \forall\, b_1, b_2$ as well as a lower bound on $\frac{d(V_m(b))}{db}$. □

Intuitively, the condition requires that the intervention effect on processes in the "bad" state must be large, making $P^a_{1,1} - P^a_{0,1}$ small. Note that Liu and Zhao [20] consider the case where $P^a_{1,1} = P^p_{1,1}$ and $P^a_{0,1} = P^p_{0,1}$, which makes Eq. 7 always true. Thus we generalize their result for threshold optimality.

Theorem 3. *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:*

$$(P^p_{1,1} - P^p_{0,1})\Big(1 + \frac{\beta(P^a_{1,1} - P^a_{0,1})}{1-\beta}\Big) \leq P^a_{1,1} - P^a_{0,1} \tag{8}$$

Intuitively, the condition requires small intervention effect on processes in the "bad" state, the opposite of the forward threshold optimal requirement. Note that both Thm. 2 and Thm. 3 also serve as conditions for the average reward case as $\beta \to 1$ (a proof based on Dutta's Theorem [8] is given in Appendix 10).

Conjecture 1. *Dual threshold policies are never optimal for Collapsing Bandits.*

This conjecture is supported by extensive numerical simulations over the random space of state transition probabilities, values of β, and values of subsidy m; its proof remains an open problem. Note that this would imply that all Collapsing Bandits are indexable.

5 Experimental Evaluation

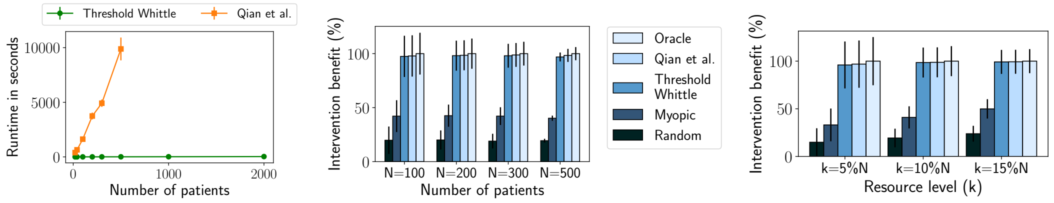
We evaluate our algorithm on several domains using both real and synthetic data distributions. We test the following algorithms: **Threshold Whittle** is the algorithm developed in this paper. **Qian et al. [27]**, a slow, but precise general method for computing the Whittle index, is our main baseline 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 Qian et al. [27] 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.

5.1 Real Data: Monitoring Tuberculosis Medication Adherence

We first test on tuberculosis medication adherence monitoring data, which contains daily adherence information recorded for each real patient in the system, as obtained from Killian et al. [18]. The "good" and "bad" states of the arm (patient) correspond to "Adhering" and "Not Adhering" to medication, respectively. State transition probabilities are estimated from the data. Because this data is noisy and contains only the adherence records and not the intervention (action) information (as the authors state), we perturb the computed average transition matrix by reducing (increasing) $P_{\omega,1}$ by a uniform random number between 0 and δ_1, δ_2 (δ_3, δ_4) then renormalizing to obtain $P^p_{\omega,1}$ ($P^a_{\omega,1}$) for the simulation. 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. All experiments in this section set all δ to 0.05.

In Fig. 4a, we plot the runtime in seconds vs the number of patients N. Fig. 4b 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. 4c 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. We also affirm the robustness of our algorithm to δ, the perturbation parameter used to approximate real-world $P^p_{\omega,1}$ and $P^a_{\omega,1}$ from the data, and present the extensive sensitivity analysis in Appendix 13. Finally, in Appendix 12 we couple our algorithm to a Thompson Sampling-based learning approach and show it performs well in the real-world case where transition probabilities would need to be learned online, supporting the deployability of our work.



(a) (b) (c)

Figure 4: (a) Threshold Whittle is several orders of magnitude faster than Qian et al. and scales to thousands of patients without sacrificing performance on realistic data (b). (c) Intervention benefit of Threshold Whittle is far larger than naive baselines and nearly as large as Oracle.

5.2 Synthetic Domains

We test our algorithm on four synthetic domains, that potentially characterize other healthcare or relevant domains, and highlight different phenomena. Specifically, we: (i) identify situations when Myopic fails completely while Whittle remains close to optimal, (ii) analyze the effect of latent state entropy on policy performance, (iii) identify limitations of Threshold Whittle by constructing processes for which Threshold Whittle shows separation from Oracle, and (iv) test robustness of our algorithm outside of the theoretically guaranteed conditions. To facilitate comparison with the real data distribution, we simulate trials for $T = 180$ rounds where reward is the undiscounted sum of arms in state 1 over all rounds. We consider the space of transition probabilities satisfying the assumed natural constraints, as outlined in Sec. 3.

Fig. 5a demonstrates a domain characterized by processes that are either self-correcting or non-recoverable. Self-correcting processes have a high probability of transitioning from state 0 to 1 regardless of the action taken, while non-recoverable processes have a low chance of doing so. We show that when the immediate reward is larger for the former than the latter, Myopic can perform even worse than Random. That is because a myopic policy always prefers to act on the self-correcting processes per their larger immediate reward, while Threshold Whittle, capable of long-term planning, looks to avoid spending resources on these processes. In this regime, the best long-term plan is to always act on the non-recoverable processes to keep them from failing. Analytical explanation of this phenomenon is presented in Appendix 11. We set the resource level, $k = 10\%N$ in our simulation for Fig. 5a. Note that performance of Myopic drops as the fraction of self-correcting processes becomes larger and reaches a minimum at $x = 100\% - k = 90\%$. Beyond this point, Threshold Whittle can no longer completely avoid self-correcting processes and the gap subsequently starts to decrease.

Fig. 5b explores the effect of uncertainty in the latent state on long-term planning. For each point on the x-axis, we draw all transition probabilities according to $P^p_{\omega,1}, P^a_{\omega,1} \sim [x, x + 0.1]$. The entropy of the state of a process is maximum near 0.5 making long term planning most uncertain and as a result, this point shows the biggest gap with Oracle, which can observe all the states in each round. Note that Myopic and Whittle policies perform similarly, as expected for (nearly) stochastically identical arms.

Fig. 5c studies processes that have a large propensity to transition to state 0 when passive and a corresponding low active action impact, but a significantly larger active action impact in state 1. This makes it attractive to exclusively act on processes in the 1 state. This simulates healthcare domains where a fraction of patients degrade rapidly, but can recover, and indeed respond very well to interventions if already in a good state. To simulate these, we draw transition matrices with $P^p_{0,1}, P^p_{1,1}, P^a_{0,1} \sim [0.3, 0.32]$ and $P^a_{1,1} \sim [0.7, 0.72]$ in varying proportions and sample the rest from the real TB adherence data. Because the best plan is to act on processes in state 1, both Myopic and Whittle act on the processes with the largest belief giving Oracle a significant advantage as it has perfect knowledge of states.

Although we provide theoretical guarantees on our algorithm for forward threshold optimal processes with non-increasing belief, Fig. 5d reveals that Alg. 1 performs well empirically even with these conditions relaxed. Here, we sample processes uniformly at random from the state transition probability space, and use rejection sampling to vary the proportion of threshold optimal processes. Threshold Whittle performs well even when as few as 20% of the processes are forward threshold optimal; we briefly analyze this phenomenon in Appendix 14.

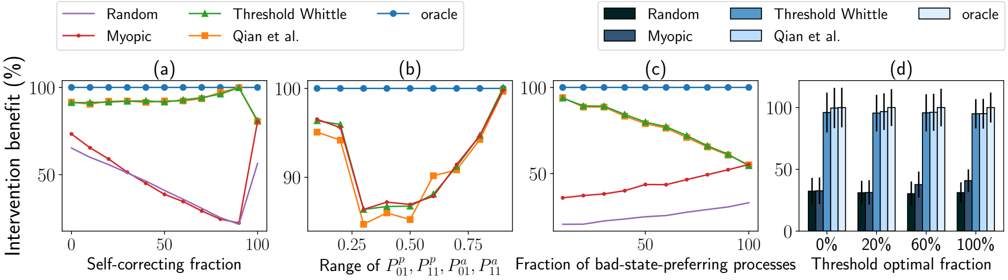


Figure 5: (a) Myopic can be trapped into performing even worse than Random while Threshold Whittle remains close to optimal. (b) Long-term planning is least effective when entropy of states is maximum. (c) Myopic and Whittle planning become similar when more processes are prone to failures. (d) Threshold Whittle is surprisingly robust to processes even outside of theoretically guaranteed conditions.

6 Conclusion

We open a new subspace of Restless Bandits, *Collapsing Bandits*, which applies to a broad range of real-world problems, especially in healthcare delivery. We give new theoretical results that cover a large portion of real-world data as well as an algorithm that runs thousands of times faster than the state of the art without sacrificing performance. We simultaneously also recognize limitations of our theoretical results, which become narrow in the average reward case. We envision several interesting avenues for future work, including techniques to incorporate the user/health worker inputs for planning, generalizing our inherently 2-state approach to allow for a multi-state model, and allowing multiple actions and/or more general reward functions.

Broader Impact

Our work is largely motivated by resource constrained health intervention delivery. This setting is common across low, middle, and high-income countries, in which community health workers (CHWs) are recruited to deliver basic care to a cohort of patients or benefactors. In fact, CHWs have been critical in achieving global health initiatives for over five decades, and evidence shows that CHWs have had a positive impact in myriad domains including maternal and newborn health [6, 9], (non-)communicable diseases [6, 29], and sexual/reproductive health [37] in low-resource communities across the world [7, 9, 29, 35]. Our modeling has the potential to improve the delivery of care in these highly resource-constrained settings.

Figure 6: CHW delivering vaccine. Credit: Pippa Ranger.

However, a deployment of our system to any setting must be done responsibly. For instance, we designed our system with the intention of *assisting* human CHWs plan resource-limited interventions. That said, we present results that highlight our algorithm's ability to plan for thousands of processes at a time, far more than for which a human could independently plan. Just making this capability available could encourage the automation of applicable interventions via automated calls or texts, potentially displacing CHW jobs, reducing human contact with patients, and unfairly limiting care for patients with limited access to technology.

Additionally, users of the system must be dutifully aware that its recommendations will be based solely on the data entered in the system. In the context of medication adherence monitoring, if the worker enters incorrect data, e.g., the patient was adhering ("good" state) but they instead mark the patient as not adhering ("bad" state), then the algorithm could make the wrong recommendation about the patient the next day, since its belief of the patient's adherence would also be wrong.

Finally, our AI system is inherently a blackbox which would likely be replacing an interpretable scheduling heuristic. This would limit any user or administrator's ability to audit decisions around why certain patients were recommended for intervention. As with any potential deployment of a blackbox system to a domain that affects the allocation of resources to humans, system designers should be acutely aware of the balance between their needs to be able to perform audits vs. their need for optimization.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Army Research Office by grant Multidisciplinary University Research Initiative (MURI) grant number W911NF1810208. J.A.K. was supported by an NSF Graduate Research Fellowship under Grant DGE1745303. A.P. was supported by the Harvard Center for Research on Computation and Society.

References

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

[2] N. Akbarzadeh and A. Mahajan. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *IEEE Conference on Decision and Control*, pages 7294–7300, 2019.

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

[4] J. N. Brownstein, F. M. Chowdhury, S. L. Norris, T. Horsley, L. Jack Jr, X. Zhang, and D. Satterfield. Effectiveness of community health workers in the care of people with hypertension. *American Journal of Preventive Medicine*, 32(5):435–447, 2007.

[5] A. H. Chang, A. Polesky, and G. Bhatia. 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):894, 2013.

[6] J. B. Christopher, A. Le May, S. Lewin, and D. A. Ross. Thirty years after Alma-Ata: a systematic review of the impact of community health workers delivering curative interventions against malaria, pneumonia and diarrhoea on child mortality and morbidity in sub-Saharan Africa. *Human Resources For Health*, 9(1), 2011. ISSN 1478-4491.

[7] Y. Dil, D. Strachan, S. Cairncross, A. Korkor, and Z. Hill. Motivations and challenges of community-based surveillance volunteers in the northern region of Ghana. *Journal of Community Health*, 37(6):1192–1198, 2012. ISSN 0094-5145.

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

[9] S. Elazan, A. Higgins-Steele, J. Fotso, M. Rosenthal, and D. Rout. Reproductive, maternal, newborn, and child health in the community: Task-sharing between male and female health workers in an Indian rural context. *Indian Journal of Community Medicine*, 41(1):34, 2016. ISSN 0970-0218.

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

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

[12] F. Iannello, O. Simeone, and U. Spagnolini. Optimality of myopic scheduling and Whittle indexability for energy harvesting sensors. In *2012 46th Annual Conference on Information Sciences and Systems*, pages 1–6. IEEE, 2012.

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

[14] Y. H. Jung and A. Tewari. Regret bounds for Thompson sampling in episodic restless bandit problems. In *Advances in Neural Information Processing Systems*, pages 9007–9016, 2019.

[15] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

[16] S. Kenya, N. Chida, S. Symes, and G. Shor-Posner. Can community health workers improve adherence to highly active antiretroviral therapy in the USA? A review of the literature. *HIV Medicine*, 12(9):525–534, 2011.

[17] S. Kenya, J. Jones, K. Arheart, E. Kobetz, N. Chida, S. Baer, A. Powell, S. Symes, T. Hunte, A. Monroe, et al. Using community health workers to improve clinical outcomes among people living with HIV: a randomized controlled trial. *AIDS and Behavior*, 17(9):2927–2934, 2013.

[18] J. A. Killian, B. Wilder, A. Sharma, V. Choudhary, B. Dilkina, and M. Tambe. Learning to prescribe interventions for tuberculosis patients using digital adherence data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2430–2438, 2019.

[19] Y. Lin, S. Liu, and S. Huang. Selective sensing of a heterogeneous population of units with dynamic health conditions. *IISE Transactions*, 50(12):1076–1088, 2018.

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

[21] B. Löwe, J. Unützer, C.M. Callahan, A.J. Perkins, and K. Kroenke. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical Care*, pages 1194–1201, 2004.

[22] R. Meshram, D. Manjunath, and A. Gopalan. On the whittle index for restless multiarmed hidden Markov bandits. *IEEE Transactions on Automatic Control*, 63(9):3046–3053, 2018.

[23] C. Mundorf, A. Shankar, T. Moran, S. Heller, A. Hassan, E. Harville, and M. Lichtveld. Reducing the risk of postpartum depression in a low-income community through a community health worker intervention. *Maternal and Child Health Journal*, 22(4):520–528, 2018.

[24] P.M. Newman, M.F. Franke, J. Arrieta, H. Carrasco, P. Elliott, H. Flores, A. Friedman, S. Graham, L. Martinez, L. Palazuelos, et al. 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*, 3(1):e000566, 2018.

[25] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

[26] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[27] Y. Qian, C. Zhang, B. Krishnamachari, and M. Tambe. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 123–131, 2016.

[28] J. Rahedi Ong'ang'o, C. Mwachari, H. Kipruto, and S. Karanja. 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.

[29] S. Shin, J. Furin, J. Bayona, K. Mate, J. Y. Kim, and P. Farmer. Community-based treatment of multidrug-resistant tuberculosis in Lima, Peru: 7 years of experience. *Social Science and Medicine*, 59(7):1529–1539, 2004. ISSN 0277-9536.

[30] B. Sombabu, A. Mate, D. Manjunath, and S. Moharir. Whittle index for AoI-aware scheduling. In *2020 12th International Conference on Communication Systems & Networks*. IEEE, 2020.

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

[32] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[33] S. S. Villar. Indexability and optimal index policies for a class of reinitialising restless bandits. *Probability in the Engineering and Informational Sciences*, 30(1):1–23, 2016.

[34] R. R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

[35] K. J. Wells, J. S. Luque, B. Miladinovic, N. Vargas, Y. Asvat, R. G. Roetzheim, and A. Kumar. Do community health worker interventions improve rates of screening mammography in the United States? A systematic review. *Cancer Epidemiology, Biomarkers & Prevention*, 20(8): 1580–1598, 2011. ISSN 1055-9965.

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

[37] WHO. *WHO Guideline on Health Policy and System Support to Optimize Community Health Worker Programmes*. WHO, 2018.

Appendix

7 Proof of Indexability

We give the proof assuming forward threshold policies are optimal, and note where relevant how the proof also works for reverse threshold optimal policies.

Fact 1. *For two non-concurrent, increasing, linear functions $f_1(m)$ and $f_2(m)$ and two points m_1, m_2, such that $m_1 \leq m_2$, if $f_1(m_1) \leq f_2(m_1)$ and $f_1(m_2) \geq f_2(m_2)$, then $\frac{df_1}{dm} \geq \frac{df_2}{dm}$. Additionally, if $f_1(m_1) < f_2(m_1)$ and $f_1(m_2) \geq f_2(m_2)$, then $\frac{df_1}{dm} > \frac{df_2}{dm}$.*

Proof. We now start proving the theorem by assuming that forward belief threshold policies are optimal. Let $b^*_{th}(m)$ denote the threshold corresponding to the optimal threshold policy for a given m. To show indexability, we must show that if a belief state b is passive, i.e., $b > b^*_{th}(m_1)$, for some m_1, then it is also passive, i.e., $b > b^*_{th}(m_2)$, for all $m_2 \geq m_1$.

In our problem, we have $2T$ belief states which, for a forward threshold policy, can be arranged in a descending order of their belief values: $\mathcal{B} := \{b_{2T}, b_{2T-1}, \dots, b_i, \dots, b_1\}$.[2] A forward threshold policy is then any real value b_{th} which splits $\mathcal{B}$ into a passive set $\mathcal{P} = \{b_i : b_i > b_{th}\ \forall b_i \in \mathcal{B}\}$ and active set $\mathcal{C} = \{b_i : b_{th} \geq b_i\ \forall b_i \in \mathcal{B}\}$. Note that all values of b_{th} such that $b_{i+1} \geq b_{th} > b_i$ $\forall i \in 1, \dots, 2T$ correspond to the same threshold policy. Thus there are only $2T + 1$ unique threshold policies possible corresponding to the $2T + 1$ such belief regions marked by points in $\mathcal{B}$. Let $\Pi = \{\pi_{2T+1}, \pi_{2T}, \dots, \pi_1\}$ denote these unique possible threshold policies arranged in a decreasing order, where $\pi_i \geq \pi_j$ implies $b^*_{th}(\pi_i) \geq b^*_{th}(\pi_j)$ where $b^*_{th}(\pi_i)$ is the optimal belief threshold associated with π_i.[3] Thus the threshold policy π_i would follow: $b_i > b^*_{th}(\pi_i) \geq b_{i-1}$ $\forall i \in 1, \dots, 2T + 1$, where $b_0 := -\infty$ and $b_{2T+1} := \infty$. Note that in a policy π_i, if for a belief state b, the optimal action is passive, then under a policy π_j, the optimal action at b is also passive $\forall j \leq i$ because $b^*_{th}(\pi_i) \geq b^*_{th}(\pi_j)$. Thus to prove indexability, it is sufficient to show that:

$$\begin{aligned} &\forall m_1, m_2 \text{ such that } m_1 \leq m_2, \\ &\text{if } \pi^*(m_1) = \pi_i \text{ and } \pi^*(m_2) = \pi_j, \text{ then} \\ &\implies i \geq j \end{aligned} \tag{9}$$

where $\pi^*(m)$ denotes the optimal threshold policy at subsidy m.

Lemma 1. Let m^*_i be the *infimum* among all m's for which $\pi^*(m) = \pi_i$. Then, the infimum is achievable (i.e., $\pi^*(m^*_i) = \pi_i$) and moreover $m^*_{2T+1} < m^*_{2T} < ... < m^*_1$.

Proof. We prove this using induction. Consider the base case: $m^*_{2T+1} < m^*_i\ \forall\ i < 2T + 1$. When $m \to -\infty$, the optimal action would clearly always be to act to avoid accruing large negative reward. So π_{2T+1} would be the optimal policy for $m \to -\infty$ and clearly the base case is true.

For the inductive case, assume the hypothesis, $m^*_{2T+1} < ... < m^*_{t+1} < m^*_i\ \forall i < t + 1$. Let m^*_t be the *infimum* among all m's for which $\pi^*(m) = \pi_t$. We must show: (1) $m^*_t < m^*_i\ \forall i < t$; (2) $\pi^*(m^*_t) = \pi_t$ (i.e., the infimum is achievable). For convenience, we denote $L = \{\pi_t, \pi_{t-1}, ...\pi_1\}$ as the set of "lower-side" polices and $U = \{\pi_{2T+1}, \pi_{2T}, ...\pi_{t+1}\}$ as the set of "upper-side" policies.

As m is increased beyond m^*_{t+1}, let m' be the *infimum value* among all m's whose optimal policy is from $L = \{\pi_t, \pi_{t-1}, ...\pi_1\}$ (note, the definition of m' is different from m^*_t since at this point we do not know whether the smallest m's optimal policy is π_t or some π_i with $i < t$ yet). That is, either the optimal threshold policy at m' is from L (when the infimum is achievable) or there exists an infinite sequence $\{\bar{m}_l\}^{\infty}_{l=1}$ that converges *from the right side* to m' (i.e., $\bar{m}_l \geq m'$ for all s) and the optimal policy for any $\bar{m}_l$ is from policy set L (when the infimum is not achievable). For notational convenience, we will think of the former achievable case also as that there is a sequence $\{\bar{m}_l\}^{\infty}_{l=1}$

[2] For simplicity, this assumes the starting belief is equal to the belief at the head of one of the chains, i.e., $P^a_{1,1}$ or $P^a_{0,1}$. However, we could add to the set $\mathcal{B}$ another T belief states corresponding to a chain that starts from any arbitrary belief and evolves for T passive actions. These new states could be ordered appropriately within $\mathcal{B}$ and the rest of the proof would follow unchanged.

[3] For reverse threshold optimal processes, simply arrange $\mathcal{B}$ and Π in ascending order of belief. The rest of the proof follows similarly.

that converges to m' and the optimal policy for any $\bar{m}_l$ is from L (letting all $\bar{m}_l = m'$ will do). In fact, a stronger conclusion holds. That is, we can choose an infinite-length sequence $\{\bar{m}_l\}_{l=1}^{\infty}$ such that the optimal policy for each $\bar{m}_l$ will be the same. This simply follows from the fact that $\{\bar{m}_l\}_{l=1}^{\infty}$ has infinite length, and their optimal policy is from a finite set L. So some policy from L must be optimal for infinitely many of $\bar{m}_l$'s. Therefore, we shall assume that $\bar{m}_l \to m'$ from the right side and the optimal policy for each $\bar{m}_l$ is some $\bar{\pi} \in L$.

Our main claim is that for subsidy m', the passive action and active action must both be optimal at state b_t. Therefore, by definition, this implies the threshold policy π_t is optimal for m'. We thus have $m_t^* = m'$, $m_i^* > m_t^*\ \forall i < t$, and moreover π_t is indeed optimal for m_t^* (i.e., the infimum is achievable). This concludes the induction proof. The remainder of this proof will be devoted to prove this claim.

By definition of m', there exists a sequence $\{\underline{m}_u\}_{u=1}^{\infty}$ that converges to m' *from the left side* (i.e., $\underline{m}_u < m'$ for all t) and moreover the optimal policy for any $\underline{m}_u$ is from the policy set $U = \{\pi_{2T+1}, \pi_{2T}, ...\pi_{t+1}\}$. Similar to the above reasoning, we shall choose the sequence $\{\underline{m}_u\}_{u=1}^{\infty}$ such that their optimal policy is the same $\underline{\pi} \in U$.

We now prove that the passive action and active action must both be optimal at state b_t for m'. Assume, for the sake of contradiction, that the optimal action at b_t for subsidy m' is passive and that the active action is not optimal (the other case where the optimal action is active follows a similar contradiction argument). That means the optimal policy for m' has a threshold $b_{th}^*(m') < b_t$ and thus $\pi^*(m') \in L$. Moreover, since the active action is not optimal for b_t, $\underline{\pi}$ must not be optimal for m' and thus achieves strictly less reward than $\pi^*(m')$. Since $\underline{m}_u \to m'$, we thus have

$$\lim_{u \to \infty} V_{\underline{m}_u}(\underline{\pi}) = V_{m'}(\underline{\pi}) < V_{m'}(\pi(m')),$$

where the last inequality uses the fact that $\underline{\pi}$ is sub-optimal for m' because the active action is strictly sub-optimal for b_t. On the other hand,

$$V_{m'}(\pi(m')) = \lim_{u \to \infty} V_{\underline{m}_u}(\pi(m')) \le \lim_{u \to \infty} V_{\underline{m}_u}(\underline{\pi})$$

These two inequalities contradict each other. This concludes our proof of the lemma. □

Let π_i be the optimal policy at some m_1.

$$\implies m_i^* \le m_1$$
$$\implies m_j^* < m_i^* \le m_1\ \forall j > i \text{ using Lemma 1}$$

Let $V_\pi(m, b)$ be the discounted reward of policy π at arbitrary state b as defined in Eq. 2 of the main text. Then for any $V_{\pi_i}(m, b)$ and $V_{\pi_j}(m, b)$ such that $j > i$ we have:

$$V_{\pi_i}(m_j^*, b) < V_{\pi_j}(m_j^*, b) \ (\pi_j \text{ is optimal at } m_j^*) \tag{10}$$
$$V_{\pi_i}(m_i^*, b) \ge V_{\pi_j}(m_i^*, b) \ (\pi_i \text{ is optimal at } m_i^*) \tag{11}$$
$$m_j^* < m_i^* \text{ if } j > i \tag{12}$$
$$\implies \frac{dV_{\pi_i}}{dm} > \frac{dV_{\pi_j}}{dm} \forall j > i \tag{13}$$

Where Eq. 10 is a strict inequality as implied by Lemma 1 and Eq. 13 follows from Fact 1 and the value function's linear dependence on m (whether discounted or average reward criterion). We now claim that $\forall m_j > m_i^*$, if π_j is optimal for m_j then we must have $j \le i$. Towards a contradiction, assume $j > i$. Then similar to the above equations, we have the following:

$$V_{\pi_i}(m_j, b) \le V_{\pi_j}(m_j, b) \ (\pi_j \text{ is optimal at } m_j) \tag{14}$$
$$V_{\pi_i}(m_i^*, b) \ge V_{\pi_j}(m_i^*, b) \ (\pi_i \text{ is optimal at } m_i^*) \tag{15}$$
$$m_i^* < m_j \tag{16}$$
$$\implies \frac{dV_{\pi_i}}{dm} \le \frac{dV_{\pi_j}}{dm} \forall j > i \tag{17}$$

Where Eq. 17 follows from Fact 1 and the value function's linear dependence on m (whether discounted or average reward criterion). which contradicts Eq. 13. Therefore, our claim holds. From 9, that implies indexability. □

8 Technical Condition for Forward Threshold Policies to be Optimal

We restate Eq. 2 here:

$$V_m(b) = max \begin{cases} m + b + \beta V_m(\tau(b)) & \text{passive} \\ b + \beta(bV_m(P^a_{1,1}) + (1-b)V_m(P^a_{0,1})) & \text{active} \end{cases}$$

where $\tau(b) := \tau_1(b)$ from Eq. 1. Simplified, $\tau(b)$ is simply a linear function of b given by the expression

$$\begin{aligned} \tau(b) &= bP^p_{1,1} + (1-b)P^p_{0,1} \\ &= (P^p_{1,1} - P^p_{0,1})b + P^p_{0,1} \end{aligned} \tag{18}$$

We will start by stating two facts, then proving three useful technical lemmas.

Fact 2. $\frac{d(\tau(b))}{db} = (P^p_{1,1} - P^p_{0,1}) \leq 1.$

Fact 3. $\forall b, b'$ s.t. $b \geq b'$, $\tau(b) \geq \tau(b')$.

Facts 2 and 3 follow from Eq 18.

Lemma 2. $V_m(b_1) - V_m(b_2) \geq b_1 - b_2, \forall b_1, b_2$ s.t. $b_1 > b_2$

Proof. We will proceed via induction, where the base case will be a one-step value function. Then we will show that the t-step value function assumption implies the t+1-step inductive value function hypothesis. In the base case the value function equals only the one-step immediate reward. It is sufficient to compare the value functions $V^1_m(b_1)$ and $V^1_m(b_2)$ element-wise, since if the true optimal action for one of the value functions is passive and the other active, the bound can still be established by flipping the action of one of the value functions as needed. This gives:

Base case $V^1_m(b_1) - V^1_m(b_2) =$

$$m + b_1 - (m + b_2) = b_1 - b_2 \qquad \text{passive} \tag{19}$$

$$b_1 - b_2 = b_1 - b_2 \qquad \text{active} \tag{20}$$

is clearly satisfied. Now assume $V^t_m(b_1) - V^t_m(b_2)) \geq b_1 - b_2$. Then $V^{t+1}_m(b_1) - V^{t+1}_m(b_2)$

Case 1 (both passive):

$$\begin{aligned} &= m + b_1 + \beta V^t_m(\tau(b_1)) - (m + b_2 + \beta V^t_m(\tau(b_2))) \\ &= b_1 - b_2 + \beta\Big(V^t_m(\tau(b_1)) - V^t_m(\tau(b_2))\Big) \\ &\geq b_1 - b_2 + \beta(\tau(b_1) - \tau(b_2)) \\ &\geq b_1 - b_2 \end{aligned} \tag{21}$$

Case 2 (both active):

$$\begin{aligned} &= b_1 - b_2 + \beta\Big((b_1 - b_2)V^t_m(P^a_{1,1}) + (b_2 - b_1)V^t_m(P^a_{0,1})\Big) \\ &= b_1 - b_2 + \beta\Big((b_1 - b_2)(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1}))\Big) \\ &= (b_1 - b_2)(1 + \beta(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1})) \\ &\geq (b_1 - b_2)(1 + \beta * 0) \\ &= (b_1 - b_2) \end{aligned} \tag{22}$$

□

Corollary 1. $V_m(b)$ is an increasing function in b, i.e., $V_m(b) \geq V_m(b')$, $\forall b, b'$ s.t. $b \geq b'$.

Proof. The proof follows from Lemma 2 by setting $b_1 = b$ and $b_2 = b'$. □

Lemma 3. $V_m(b_1) - V_m(b_2) \leq \frac{b_1 - b_2}{1-\beta}, \forall b_1, b_2$ s.t. $b_1 > b_2$

Proof. Proceed by induction again. The base case $V_m(b_1) - V_m(b_2) =$

$$m + b_1 - (m + b_2) = b_1 - b_2 \leq \frac{b_1 - b_2}{1 - \beta} \qquad \text{both passive} \tag{23}$$

$$b_1 - b_2 = b_1 - b_2 \leq \frac{b_1 - b_2}{1 - \beta} \qquad \text{both active} \tag{24}$$

which are both clearly satisfied. Now assume $V_m^t(b_1) - V_m^t(b_2) \leq \frac{b_1 - b_2}{1-\beta}$. Then, $V_m^{t+1}(b_1) - V_m^{t+1}(b_2)$

Case 1 (both passive):

$$\begin{aligned}
&= \big(m + b_1 + \beta V_m^t(\tau(b_1))\big) - \big(m + b_2 + \beta V_m^t(\tau(b_2))\big) \\
&= (b_1 - b_2) + \beta\big(V_m^t(\tau(b_1)) - V_m^t(\tau(b_2))\big) \\
&\leq (b_1 - b_2) + \beta\left(\frac{\tau(b_1) - \tau(b_2)}{1 - \beta}\right) \\
&\leq (b_1 - b_2) + \beta\left(\frac{(b_1 - b_2)}{1 - \beta}\right) \text{ by Fact 3} \\
&= \frac{b_1 - b_2}{1 - \beta}
\end{aligned} \tag{25}$$

Case 2 (both active):

$$\begin{aligned}
&= \Big(b_1 + \beta\big(b_1 V_m^t(P_{1,1}^a) + (1 - b_1) V_m^t(P_{0,1}^a)\big)\Big) - \\
&\Big(b_2 + \beta\big(b_2 V_m^t(P_{1,1}^a) + (1 - b_2) V_m^t(P_{0,1}^a)\big)\Big) \\
&= (b_1 - b_2) + \beta\Big(\big(b_1 - b_2\big)\big(V_m^t(P_{1,1}^a) - V_m^t(P_{0,1}^a)\big)\Big) \\
&\leq (b_1 - b_2) + \beta\left((b_1 - b_2).\frac{P_{1,1}^a - P_{0,1}^a}{1 - \beta}\right) \\
&\leq (b_1 - b_2) + \beta\left(\frac{(b_1 - b_2)}{1 - \beta}\right) \text{ by Fact 2} \\
&= \frac{b_1 - b_2}{1 - \beta}
\end{aligned} \tag{26}$$

□

Lemma 4. $\frac{d(V_m(b))}{db} \geq 1 + \beta\alpha$
where, $\alpha = \min\{P_{1,1}^p - P_{0,1}^p, P_{1,1}^a - P_{0,1}^a\}$

Proof. Using Eq. 2, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} 1 + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))} \frac{d(\tau(b))}{db} & \text{passive} \\ 1 + \beta(V_m(P_{1,1}^a) - V_m(P_{0,1}^a)) & \text{active} \end{cases} \tag{27}$$

Case 1 (passive):

$$= 1 + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))}(P_{1,1}^p - P_{0,1}^p) \tag{28}$$

$$= 1 + \beta \lim_{\delta \to 0} \frac{V_m(\tau(b) + \delta) - V_m(\tau(b))}{\tau(b) + \delta - \tau(b)}(P_{1,1}^p - P_{0,1}^p) \tag{29}$$

$$\geq 1 + \beta(P_{1,1}^p - P_{0,1}^p) \text{ by Lemma 2} \tag{30}$$

$$\geq 1 + \beta\alpha \tag{31}$$

Case 2 (active):

$$\begin{aligned}
&= 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) && (32)\\
&\geq 1 + \beta(P^a_{1,1} - P^a_{0,1}) \text{ by Lemma 2} && (33)\\
&\geq 1 + \beta\alpha && (34)
\end{aligned}$$

□

Now we derive the technical condition for **Theorem 2**. In this case, proving that threshold policies are optimal is equivalent to proving that, if it is optimal to act now, then it is optimal to act for all later beliefs. Formally, if for a belief b, the optimal action is to act, then we must show that for a lower $b' < b$, the optimal action is also to act. To do this, we show that Theorem 2 implies that the derivative wrt b of the passive action value function is greater than the derivative wrt b of the active action value function:

$$(P^p_{1,1} - P^p_{0,1})(1 + \beta(P^a_{1,1} - P^a_{0,1}))(1 - \beta) \geq P^a_{1,1} - P^a_{0,1} \tag{35}$$

Note that since $(P^a_{1,1} - P^a_{0,1}) \leq 1, \implies (1 + \beta(P^a_{1,1} - P^a_{0,1}))(1 - \beta) \leq 1$, Eq.35 itself implies that $\alpha = P^a_{1,1} - P^a_{0,1}$. Thus, it becomes:

$$\begin{aligned}
&(P^p_{1,1} - P^p_{0,1})(1 + \beta\alpha)(1 - \beta) \geq P^a_{1,1} - P^a_{0,1} && (36)\\
\implies &(P^p_{1,1} - P^p_{0,1})(1 + \beta\alpha) \geq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma 3} && (37)\\
\implies &(P^p_{1,1} - P^p_{0,1})\frac{d(V_m(b))}{db} \geq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma. 4} && (38)\\
\implies &1 + \beta\frac{d(V_m(\tau(b)))}{d(\tau b)}\frac{d(\tau(b))}{db} \geq 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \text{ by Fact 2} && (39)\\
\implies &\frac{d(V_m(b|a=0))}{d(b)} \geq \frac{d(V_m(b|a=1))}{d(b)} && (40)\\
& && (41)
\end{aligned}$$

9 Technical Condition for Reverse Threshold Policies to be Optimal

Now we derive a technical condition for a reverse threshold policy. That is, a threshold policy in which if it is optimal to be passive in the current state, then it must also be optimal to act in all later states in the order. First we prove one more technical Lemma.

Lemma 5. $\frac{d(V_m(b))}{db} \leq 1 + \frac{\beta\gamma}{1-\beta}$
where, $\gamma = \max\{P^p_{1,1} - P^p_{0,1}, P^a_{1,1} - P^a_{0,1}\}$

Proof. Using Equation 8, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} 1 + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}\frac{d(\tau(b))}{db} & \text{passive} \\ 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) & \text{active} \end{cases} \tag{42}$$

Case 1 (passive):

$$\begin{aligned}
&= 1 + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}(P^p_{1,1} - P^p_{0,1}) && (43)\\
&= 1 + \beta\lim_{\delta\to 0}\frac{V_m(\tau(b) + \delta) - V_m(\tau(b))}{\tau(b) + \delta - \tau(b)}(P^p_{1,1} - P^p_{0,1}) && (44)\\
&\leq 1 + \frac{\beta}{1-\beta}(P^p_{1,1} - P^p_{0,1}) \text{ by Lemma 3} && (45)\\
&\leq 1 + \frac{\beta\gamma}{1-\beta} && (46)
\end{aligned}$$

Case 2 (active):

$$
\begin{aligned}
&= 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) && (47)\\
&\leq 1 + \frac{\beta}{1-\beta}(P^a_{1,1} - P^a_{0,1}) \text{ by Lemma 3} && (48)\\
&\leq 1 + \frac{\beta\gamma}{1-\beta} && (49)
\end{aligned}
$$

□

Now to give a condition under which reverse threshold policies are optimal. Formally, if for a belief b, the optimal action is to be passive, then we must show that for a lower $b' < b$, the optimal action is also to be passive. We do this by showing that the Theorem 3 statement implies that the derivative wrt b of the passive value function is less than the derivative wrt b of the active action value function:

$$
(P^p_{1,1} - P^p_{0,1})(1 + \frac{\beta(P^a_{1,1} - P^a_{0,1})}{1-\beta}) \leq P^a_{1,1} - P^a_{0,1} \tag{50}
$$

Note that the Eq. 50 itself implies that $\gamma = P^a_{1,1} - P^a_{0,1}$, thus giving:

$$
\begin{aligned}
&(P^p_{1,1} - P^p_{0,1})(1 + \frac{\beta\gamma}{1-\beta}) \leq P^a_{1,1} - P^a_{0,1} && (51)\\
\implies &(P^p_{1,1} - P^p_{0,1})(1 + \frac{\beta\gamma}{1-\beta}) \leq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma 2} && (52)\\
\implies &(P^p_{1,1} - P^p_{0,1})\frac{d(V_m(b))}{db} \leq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma 5} && (53)\\
\implies &1 + \beta\frac{d(V_m(\tau(b)))}{d(\tau b)}\frac{d(\tau(b))}{db} \leq 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \text{ by Fact 2} && (54)\\
\implies &\frac{d(V_m(b|a=0))}{d(b)} \leq \frac{d(V_m(b|a=1))}{d(b)} && (55)\\
& && (56)
\end{aligned}
$$

10 Threshold Conditions for Average Reward Case

First we define the concept of *value boundedness* [8]:

Definition 4 (Value Boundedness)**.** *For a given MDP, consider a value function $V_\beta(b)$, states $b \in \mathcal{B}$ and some index state $z \in \mathcal{B}$. Then an MDP is value bounded if for a constant M and function $M(b)$:*

$$
M(b) < V_\beta(b) - V_\beta(z) < M \tag{57}
$$

We now prove that Thm. 2 and Thm. 3 hold respectively under the average reward criterion as $\beta \to 1$ using Dutta's Theorem as follows [8]. Consider an MDP that is *value bounded.* Let $\pi_\beta(\cdot)$ be a stationary optimal policy for the discounted MDP. (1) Suppose $\pi_\beta(\cdot) \to \pi$ pointwise, as $\beta \to 1$. Then π is a stationary optimal policy for the average reward criterion. (2) Furthermore, given state ordering O, if for all discounted optimal policies $\pi_\beta(b)$, $O(b') \geq O(b)$ implies $\pi_\beta(b') \geq \pi_\beta(b)$ (i.e., threshold policies are optimal), then any sequence of discounted optimal policies converge to an average optimal policy as $\beta \to 1$.

(2) and (1) together imply that any MDP that admits threshold optimal policies under discounted reward criteria also admits threshold optimal policies under average reward criteria. By construction, any MDP that satisfies Thm. 2 or Thm. 3 admits threshold optimal policies under the discounted reward criterion. Therefore, to prove that those conditions hold under the average reward criterion as $\beta \to 1$, we need only prove that any CoB is value bounded.

Theorem 4. *Any Collapsing Bandit is value bounded.*

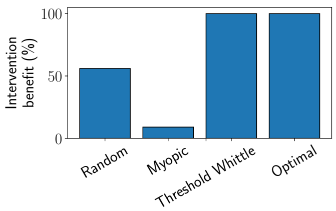
11 Example When the Myopic Policy Fails

We present an example in which the myopic baseline is barely better than No Calls, while Threshold Whittle is *optimal*. Consider the system with $N = 2$ and $k = 1$ and the transition probabilities shown in Fig. 7a.

$$P^{p,1} = \begin{bmatrix} 0.97 & 0.03 \\ 0.03 & 0.97 \end{bmatrix} \quad P^{a,1} = \begin{bmatrix} 0.96 & 0.04 \\ 0.01 & 0.99 \end{bmatrix}$$

$$P^{p,2} = \begin{bmatrix} 0.25 & 0.75 \\ 0.03 & 0.97 \end{bmatrix} \quad P^{a,2} = \begin{bmatrix} 0.23 & 0.77 \\ 0.01 & 0.99 \end{bmatrix}$$

(a)



(b)

Figure 7: For the example transition matrices, Myopic performs worse than random, while Threshold Whittle is nearly optimal.

Fig. 7b shows how various policies perform on these two processes. The myopic policy is worse than random and threshold Whittle is nearly optimal. The myopic policy always acts on process 2 because the immediate reward it considers, $(b_{t+1}|a = 1) - (b_{t+1}|a = 0)$ is marginally higher for process 2 than process 1. However, process 1 is better to pull in the long run because process 2 has a large $P^p_{0,1}$, making it self-correcting, meaning the process is likely to become adhering quickly even without an intervention. However, process 1 has a very small $P^a_{0,1}$ and $P^p_{0,1}$ and is thus difficult to revive from the bad state even with an intervention, making it important to keep intervening to stop the process from ever entering the bad state.

The following analysis shows that the myopic policy always prefers to pull arm 2:

For process 1:

$$\begin{aligned} (b_{t+1}|a = 0) &= 0.97.b_t + 0.03.(1 - b_t) && = 0.94.b_t + 0.03 \\ (b_{t+1}|a = 1) &= 0.99.b_t + 0.01.(1 - b_t) && = 0.95.b_t + 0.04 \\ \text{Thus, } \Delta b_t &= (b_{t+1}|a = 1) - (b_{t+1}|a = 0) && = 0.01 + 0.01.b_t < 0.02 \end{aligned}$$

Similarly, for process 2:

$$\Delta b_t = 0.02$$

The myopic policy chooses the arm with the greater Δb_t.

12 Learning Online

So far we assumed that *all transition probabilities are known.* However, in a real deployment, the transition probabilities of processes would be unknown at the start, and it would be desirable to learn the transition probabilities online in tandem with planning. To develop an online planning regime for our algorithm, we use the tuberculosis medication adherence monitoring domain from the main text as a case study and motivating example.

We implement a Thompson sampling-based learning method [32], which is a heuristic which has been shown to work well in practice and has been frequently used in the bandit literature [14]. In Thompson sampling, we sample from a posterior distribution over the estimated parameters and use the samples for planning. This allows for “sub-optimal" actions to be taken periodically, building exploration implicitly into planning. Then, as arms are pulled we use the observations to update our posterior distribution. We maintain a Beta distribution posterior over the parameters of each row of a patient’s transition matrix and sample from it each day to generate a matrix with which the system can plan for that round.

Additionally, we consider two specific features of the TB medication adherence monitoring domain that can be used to accelerate learning with Thompson sampling. First, it is reasonable to assume that patients (processes) might remember some number of previous days of their medication adherence behavior. Thus, when the agent pulls an arm, the arm may reveal state observations for some number of previous days which we call *buffer length*. The larger the buffer length, the faster learning will converge since more observations are obtained for updating the posterior. We parameterize buffer length and evaluate its effect on learning and planning in experiments. Second, we verify with real data that virtually all patients adhere to the natural constraints on the transition probabilities given in Section 3. We exploit this known structure on the transition probabilities – i.e., that processes tend to degrade when passive and that interventions must have positive effect – to identify a constrained probability space from which we would like to sample when learning online. We implement a version of Thompson sampling called *constrained* Thompson sampling which samples from this joint, constrained probability space via rejection sampling.

On-demand index computation algorithm. When we learn online, the transition matrices for a process change every day, and thus pre-computing the Whittle indices for every belief state as in Alg. 1 is inefficient. We can address this by identifying and solving only the indifference equation that is relevant to the current state of the process. We use the insight that for a threshold of X_i on the current chain i, the corresponding threshold X_j on chain j would be the state with the largest belief lower than $b(X_i)$, i.e., $X_j = \min_u\{u : b_j(u) < b(X_i)\}$. The Whittle index for X_i is then obtained by solving for $m : J_m^{(X_i,X_j)} = J_m^{(X_i+1,X_j)}$. These computations are repeated every day yielding overall complexity of $\mathcal{O}(|\Omega|T^2)$ per process.

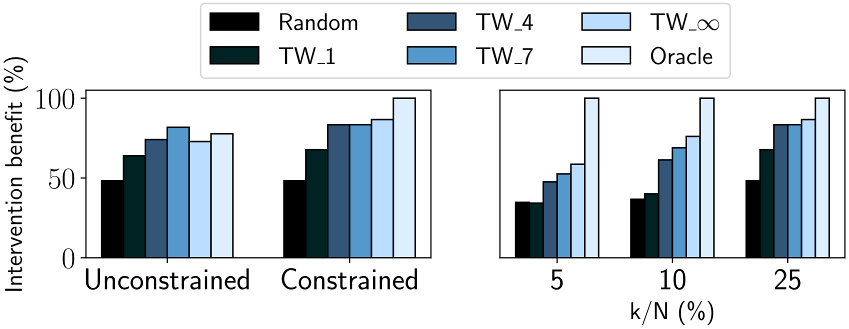


Figure 8: (left) Constrained Thompson sampling improves learning. (right) buffer lengths of 4–7 perform well for various values of k/N, using constrained Thompson sampling. TW_X is the on-demand index algorithm run in tandem with Thompson sampling and a buffer length of X.

Fig. 8 (right) evaluates the impact of varying buffer lengths for various ratios of k/N. Note that in these experiments, Oracle fully observes states, but must still learn transition probabilities online. Critically, we see that even when simulated patients report 4–7 observations per arm-pull, the performance is close to that of the non-Oracle learning upper bound (buffer length=∞) for any k/N. This is a key consideration for deployment in a medication adherence context: patients need only remember their last 4–7 doses on average for our approach to be nearly effective as possible in the TB context.

Fig. 8 (left) compares the performance of learning policies with and without constrained Thompson sampling for $k/N = 25\%$. All policies benefit from the constrained sampling approach, suggesting that imposing our knowledge of the transition probability constraints was beneficial to learning.

13 Sensitivity Analysis

In Fig. 9, we investigate Threshold Whittle's performance relative to the choice of parameters used to perturb the real data from the TB medication adherence domain. All the plots show that Threshold Whittle's performance is robust to the choice of parameters.

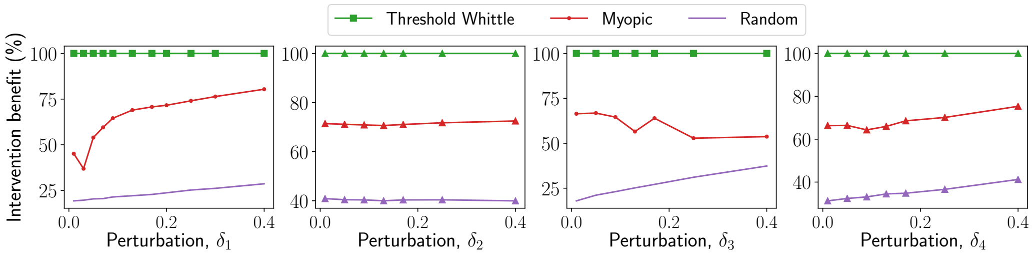
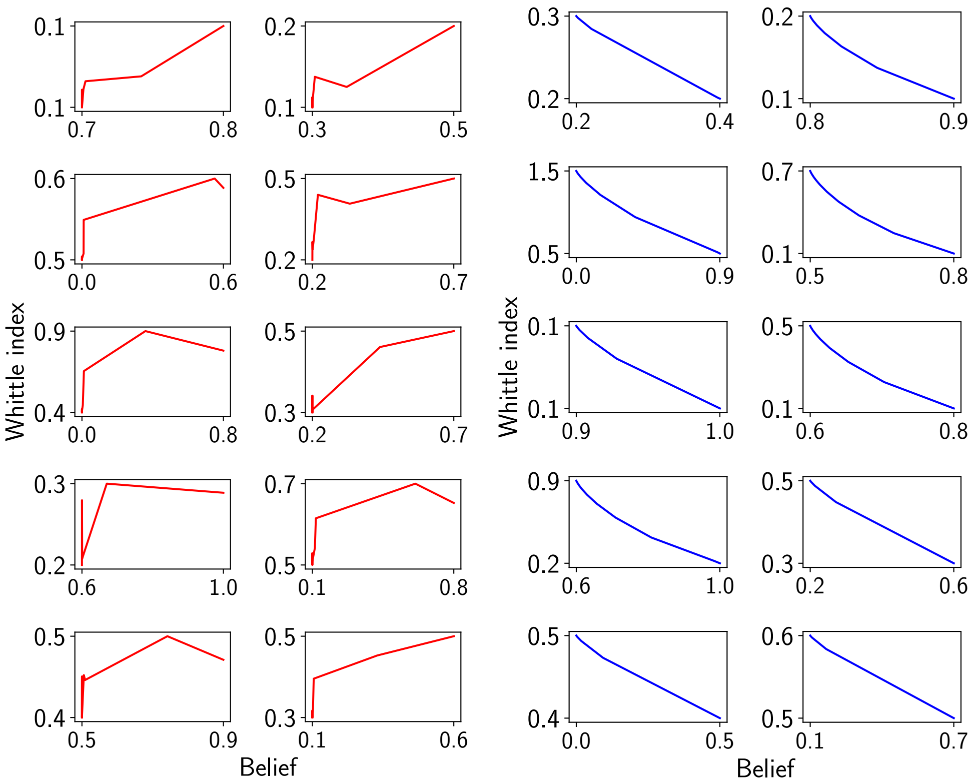


Figure 9: Performance of Threshold Whittle is robust to perturbation of the transition matrix parameters. Note that 100% corresponds to the performance of Threshold Whittle for this plot only.



(a)

(b)

Figure 10: (a) Threshold Whittle-computed indices vs. reachable beliefs for 10 randomly sampled reverse threshold optimal processes (one line per process). These indices tend to increase in belief, as expected for reverse threshold optimal processes according to the proof in Appendix 7. (b) Threshold Whittle-computed indices vs. reachable beliefs for 10 randomly sampled forward threshold optimal processes (one line per process). These indices always decrease in belief, as expected for forward threshold optimal processes according to the proof in Appendix 7.

14 Threshold Whittle's Performance on Reverse Threshold Optimal Processes

Here we investigate why Threshold Whittle demonstrates near-optimal performance even on reverse-threshold-optimal processes. We randomly sample forward and reverse threshold optimal processes, checked with Thm. 2 and Thm. 3, respectively, using $\beta = 0.95$, then compute their indices with the Threshold Whittle algorithm. Figures. 10a and 10b show a few samples of these trajectories for reverse and forward threshold optimal processes, respectively. Via similar arguments from the proof in Appendix 7, it can be shown that the true Whittle indices for reverse (forward) threshold optimal processes should always be increasing (decreasing) in belief. Fig. 10a shows that for such reverse threshold optimal processes, the indices computed by Threshold Whittle do tend to increase in belief as expected, which may lead to Threshold Whittle's good performance even though it is not guaranteed to be optimal on those processes. (And for completeness, Fig. 10b shows that for forward threshold optimal policies, the indices computed by Threshold Whittle always decrease in belief as expected.)