Risk-Aware Interventions in Public Health: Planning with Restless Multi-Armed Bandits

Aditya Mate
Harvard University
Cambridge, MA
aditya_mate@g.harvard.edu

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

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

ABSTRACT

Community Health Workers (CHWs) form an important component of health-care systems globally, especially in low-resource settings. CHWs are often tasked with monitoring the health of and intervening on their patient cohort. Previous work has developed several classes of Restless Multi-Armed Bandits (RMABs) that are computationally tractable and indexable, a condition that guarantees asymptotic optimality, for solving such health monitoring and intervention problems (HMIPs). However, existing solutions to HMIPs fail to account for risk-sensitivity considerations of CHWs in the planning stage and may run the danger of ignoring some patients completely because they are deemed less valuable to intervene on. Additionally, these also rely on patients reporting their state of adherence accurately when intervened upon. Towards tackling these issues, our contributions in this paper are as follows: (1) We develop an RMAB solution to HMIPs that allows for reward functions that are monotone increasing, rather than linear, in the belief state and also supports a wider class of observations. (2) We prove theoretical guarantees on the asymptotic optimality of our algorithm for any arbitrary reward function. Additionally, we show that for the specific reward function considered in previous work, our theoretical conditions are stronger than the state-of-the-art guarantees. (3) We show the applicability of these new results for addressing the three issues pertaining to: risk-sensitive planning, equitable allocation and reliance on perfect observations as highlighted above. We evaluate these techniques on both simulated as well as real data from a prevalent CHW task of monitoring adherence of tuberculosis patients to their prescribed medication in Mumbai, India and show improved performance over the state-of-the-art. Full paper and code is available at: https://github.com/AdityaMate/risk-aware-bandits.

KEYWORDS

Restless Multi-Armed Bandits; Whittle Index; Healthcare: Intervention Planning; Sequential Decision Making; MDP; POMDP

ACM Reference Format:
Aditya Mate, Andrew Perrault, and Milind Tambe. 2021. Risk-Aware Interventions in Public Health: Planning with Restless Multi-Armed Bandits. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 15 pages.

1 INTRODUCTION

Community Health workers (CHWs) play a key role in complementing the primary health facilities, and are critical to health care systems globally, and especially in low-resource countries [29]. CHWs are members of the local community who serve as frontline health workers and form the cornerstone of the bridge between the health resources and the local communities through building trust and a range of other activities such as outreach, providing health education, screening and basic emergency care [33, 34]. The effectiveness of CHWs in achieving desirable community health outcomes through the interventions they deliver has been recognized in the context of several domains such as achieving child survival goals [10], improving child and maternal health [20, 31], communicable and non-communicable diseases [5, 26], sexual and reproductive health [21], etc.

Figure 1: Community Health Worker delivering an intervention. Image source: Pippa Ranger

A key challenge that CHWs face in effective delivery of welfare activities is optimally managing their severely limited resources. In the global south, each CHW may routinely be responsible for managing the health outcomes of hundreds of patients. As a motivating example, we consider the real-world CHW HMIP of monitoring adherence for tuberculosis (TB) patients, who must complete a 6-month medication plan. Given the resource scarcity, the CHWs can only monitor and intervene on some k patients from their N-strong patient cohort ($k \ll N$) each day. In this situation, the CHWs must determine the best k candidates to intervene on each day, based on who would likely display the highest benefits of the intervention through improvement in their future adherence. While doing so, the CHWs must simultaneously juggle at least three real-world considerations, in addition to broadly maximizing the overall adherence of their cohort. These may include: incorporating risk-sensitive perspectives, ensuring no patients are left ignored for too long, or accounting for patients who may misrepresent their adherence status.

A naive planning approach typically implemented in practice is to intervene on patients in a round robin fashion. However, this strategy is likely sub-optimal because some patients may need interventions less often than others. Previous works in AI for health interventions [4, 16, 23] have largely focused on building assistants

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

that send personalized health reminders or recommendations to patients. However, these assume resource-rich environments in which interventions can be launched at will, and are thus irrelevant to the CHWs' intervention planning problem at hand. Some recent works in AI [17, 18, 25] have also explored intervention planning algorithms under limited resources using the Restless Multi-Armed Bandits (RMAB) framework. However, these are either slow or can only optimize for aggregate cohort-level health statistics weighing the adherence in all stages of the program equally and do not cater to the complicated patient-specific considerations of the CHWs.

In this paper, we tackle this issue of planning the limited CHW intervention resources in the HMIP while accommodating more complex objectives than past work. Our theoretical analysis identifies a wider class of indexable HMIPs even in the case of standard linear rewards. We leverage these results to construct tailor-made reward functions, designed to accommodate the real-world planner objectives outlined above. Further, we also develop additional techniques to solve the issue pertaining to patients incorrectly reporting/not reporting their true adherences.

Thus, our contributions in this paper are as follows: (1) We present an algorithm for the HMIP that can admit any arbitrary, monotonically increasing reward function and supports a wider class of observations. (2) We prove theoretical guarantees on the optimality of our algorithm. Further, we show that for the specific reward definition of average cohort adherence studied in previous work, our conditions are much wider (giving stronger results). For example, in the average reward case, the previous optimality guarantees become vacuous, while our theoretical guarantees hold for as much as 88% of the entire space of bandits. (3) We show the applicability of these results for catering to three real-world CHW considerations including: (i) risk-sensitive planning, (ii) fairness protection towards patients who may otherwise be completely ignored by the planning algorithms, and (iii) accounting for patients who may misrepresent their true adherences.

2 BACKGROUND

2.1 Restless Multi-Armed Bandits.

An RMAB consists of N independent arms, each consisting of an associated 2-action Markov Decision Process (MDP) [24]. An MDP is defined by the tuple $\{\mathcal{S}, \mathcal{A}, r, \mathcal{P}\}$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the set of possible actions, r is a state-dependent reward function $r : \mathcal{S} \to \mathbb{R}$ and $\mathcal{P}$ represents a transition function, with $P^a_{s,s'}$ representing the probability of transitioning from a current state s to a next state s' when an action a is taken. An MDP policy, $\pi : \mathcal{S} \to \mathcal{A}$ is a mapping from the state space to the action space specifying the action to be taken at a particular state. The reward accrued by a policy π can be measured either using the discounted reward or the average reward criterion. The discounted reward of a policy π starting from an initial state s_0 is defined as $R^\pi_\beta(s_0) = E\left[\sum_{t=0}^{\infty} \beta^t r(s_t)|\pi, s_0\right]$, where $\beta \in [0, 1)$ is the discount factor and actions are selected according to π. The average reward of a policy π can be defined (independent of the starting state) as: $\overline{R}^\pi = \sum_{s\in\mathcal{S}} f^\pi(s) r(s)$, where $f^\pi(s)$ represents the average visit frequency induced by the policy π, or the long term fraction of time spent in a state s when following π. The total reward accrued by the planner is the sum of the total individual rewards accrued by each of the arms (under either the discounted or average reward criteria). The planner's goal is to maximize her total reward summed up across all arms.

We model the intervention planning problem as an RMAB with each arm representing an agent (patient) with the planner (CHW) who must decide which arms to monitor and intervene upon.

2.2 Whittle Index solution technique

Computing the optimal policy for an RMAB has been shown to be PSPACE hard in general even when the transition dynamics are perfectly known [22]. However, Whittle proposed a heuristic [32], known today as the Whittle Index, that was later been shown to be asymptotically optimal for the time average reward problem [30], and also for other more general families of RMABs arising from stochastic scheduling problems [8].

The main idea of the Whittle Index technique is to compute an index for every arm at each time step that intuitively captures the value of pulling that arm at that timestep. Such an index is calculated for each arm independently, thus transforming the N-arm RMAB problem to N smaller problems each consisting of a single MDP. The Whittle Index policy for the RMAB is to pull the k arms with the highest Whittle indices.

The notion of the Whittle Index is centered around the concept of passive subsidy, m. Intuitively, passive subsidy is the amount one must pay the planner as compensation *not* to pull an arm. Formally, this can be expressed through a modified reward function for each arm, given as: $r_m : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, where $r_m(s, a = 0) = r(s) + m$ and $r_m(s, a = 1) = r(s)$, where $a = 1(a = 0)$ for the MDP corresponds to pulling (not pulling) an arm of the RMAB. The modified reward function induces a corresponding value function in each state, for each of the two actions: $V_m(s, a = 0)$ and $V_m(s, a = 1)$. The Whittle Index W is defined as the infimum subsidy m for which the planner is indifferent between either pulling or not pulling the arm. In other words, $W(s) = \inf_m\{m : V_m(s, a = 0) = V_m(s, a = 1)\}$.

A common challenge associated with the Whittle Index solution technique is establishing a technical condition, known as 'indexability' that guarantees the asymptotic optimality of the Whittle Index heuristic. This condition may not be satisfied by all RMABs and previous literature has established indexability only for specific problem instances. A second challenge is often computing the value of the Whittle Index itself, which can be computationally expensive or may often need numerical approximations.

2.3 Related Work

RMABs have proved to be a popular framework for modeling limited resource planning problems in a myriad of domains. Because establishing indexability for RMABs is very challenging, previous works have only explored the same for specialized problem structures. [8] prove indexability results for a family of RMABs that arise in machine maintenance and stochastic problems with switching penalties. However, they assume a deterministic action effect, whereas we do not. [12] and [27] augment the machine maintenance problem by introducing either i.i.d. or Markovian stochasticity in the reset action, and [28] study Whittle Index for general functions of states assuming a single, fixed, reset state. [19] explore Hidden Markov Bandits which consider partial observability with binary

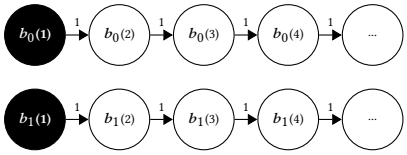


Figure 2: Belief states are arranged in two chains, one corresponding to each observation. Belief state deterministically transitions to the next belief state in the chain when passive. $b_0(1)$ and $b_1(1)$ (shown in black) are the reset states. [18].

state transitions, but don't accommodate state dependent rewards from passive arms.

Liu and Zhao [17] is a seminal work that builds off of the well-established 2-state Elliot-Gilbert channel model [7] and computes the Whittle Index efficiently along with a closed form. They assume that the state transitions are unaffected by the action taken and only accrue reward from the active arms. [1] is a recent work that considers RMABs with "controlled restarts" giving indexability results as well as a closed form for the Whittle Index, but they rely on state-independent restarts, which is narrower than the model of this paper. [25] present a more generic approach that relies on solving the MDP on each arm for the optimal action to compute the Whittle Index policy. While it can thus relax many of these constraints, this technique is very expensive computationally, and is thus very slow. [18] is a recent work that is orders of magnitude faster while also relaxing the restrictions of previous work. However, they fail to account for real world risk-sensitive and fairness related planning considerations. Additionally, they also assume perfect observability of the patient states when acted upon, which may be unrealistic. Despite these shortcomings, the performance guarantees only hold for a narrow range of RMABs.

A rich body of literature has also explored risk-sensitive and other similar learning-based perspectives to bandit planning. However, most of these consider risk and the risk-attitude in the learning stage while minimizing regret, and not in the planning stage [2, 13, 35]. [14] and [3] are other contemporaneous works that focus on RMAB planning with multiple available actions or model-free approaches to learning in RMABs.

3 PROBLEM FORMULATION

We define the health monitoring and intervention problem (HMIP) as follows. In this problem, the planner represents the community health worker responsible for managing the health outcomes for their patient cohort. The patient cohort is represented by a set of N agents (representing arms of the RMAB), $\mathcal{N} = \{1, 2, \dots N\}$, whose health outcomes are monitored by the planner. The planner must decide which arms to pull (which patients to intervene on) each day of the program. The health program lasts for T discrete days.

On each day of the program, each agent can be in one of two latent states, a 'good' state (1) and a 'bad' state (0)—denoted by $\mathcal{S} = \{1, 0\}$. In the context of tuberculosis adherence monitoring, this translates to each patient being in either the *adherent* or the *non-adherent* latent state respectively each day, for $T = 180$ days of the treatment program. Each agent follows an MDP, with states defined by the belief value, i.e. the probability that the agent is in the 'good' latent state at that time step. We assume such a belief-state MDP over states $b \in \mathcal{B}$ is fixed and known, but can be unique to every agent and have arbitrary transition dynamics.

The action space, $\mathcal{A}$ consists of two possible actions: passive (denoted as '0') and active (denoted as '1', representing an intervention). The planner can intervene on at most k agents each day (where $k \ll N$ because of scarce resources). Let $a_t \in \{0, 1\}^N$ denote the vector of actions chosen by the planner on a particular day. Then such an a_t must have $\|a_t\| \leq k$ because of the resource constraint. In case of passive actions, no observation about the agent is available and the belief state evolves according to the standard belief update: $b \rightarrow bP^p_{11} + (1-b)P^p_{01}$. When an active action is taken, the patient emits an observation ω from the observation set $\Omega = \{0, 1, \dots |\Omega| - 1\}$ and as a result of the intervention, transitions to a 'reset' belief state. The reset state engendered by the intervention, depends on whether precise observations are available. In case of precise observations, the planner can observe the agent's true latent state upon intervening, leading to $\Omega = \{0, 1\}$. In this case, the agent's belief state resets to a value $P^a_{\omega 1}$ depending on which $\omega \in \{0, 1\}$ was observed. In the context of TB however, assuming perfect, reliable observations may be unrealistic in some cases as patients may sometimes refuse to answer the CHWs' intervention phone calls or may not report their latent state truthfully. We cast these events as imprecise observations of the patient's latent state. When observations are imprecise, since true state of the patient is unobserved, the planner pre-defines a fixed reset belief state for every possible observation $\omega \in \Omega$. These imprecise observations are assumed to be emitted according to a fixed, known emission matrix, $E^{|\mathcal{S}|\times|\Omega|}$ unique to every patient. In our empirical analysis in Section 5, for simplicity, we assume two such possible imprecise observations—a positive shade and a negative shade of response (resetting to P^a_1 and P^a_0 respectively such that $P^a_0 \leq P^a_1$)—however, our algorithm is again amenable to a multiple-observation setting.

We impose two additional natural constraints on each arm as consistent with previous literature [17, 18] that closely simulate real settings: (1) $P^a_{0,1} < P^a_{1,1}; P^p_{0,1} < P^p_{1,1}$; (it is more likely for a patient to stay adhering than it is to switch from being non-adhering to adhering) and (2) $P^a > P^p; P^a_1 > P^p_{1,1}; P^a_0 > P^p_{0,1}$ (intervention effect is positive).

The planner's goal is to find an intervention policy that maximizes her utility measured according to her own yardstick, defined by the utility function $\mathcal{U}$. For each patient in a belief state b in the MDP, we assume the planner accrues a reward $\rho(b)$ for that patient at that time step, where ρ is chosen such that $\mathbb{E}[\mathcal{U}(b)] = \rho(b)$. The planner solves for a policy that maximizes the total reward accrued, $\sum_{t=1}^{T}\sum_{n=1}^{N}\rho(b_t)$ summed up over all agents over the entire time horizon, which is in effect tantamount to maximizing her expected utility.

Prior work in the context of TB such as [18] considers a planner with the goal of maximizing the overall average adherence of the patients. For such a planner, $\mathcal{U} = \begin{cases} 1 & \text{if patient adheres} \\ 0 & \text{if patient does not adhere} \end{cases}$. Thus $\mathbb{E}[\mathcal{U}] = \mathbb{P}[\text{patient adheres}] = b$. Thus setting $\rho(b) = b$ for each belief state optimizes for the average adherence objective. In this work, we allow the planner to have an arbitrary objective that

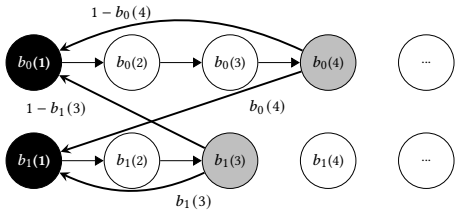


Figure 3: State transition diagram when a threshold policy with thresholds $u_0 = 4$ and $u_1 = 3$ is implemented. Belief stochastically resets to one of the reset states when active.

translates to the goal of maximizing the long term reward accrued, specified by an arbitrary, monotonically increasing $\rho(b)$.

4 INDEX POLICY COMPUTATION

Belief state MDP. Our analysis of the agents' behavior is centered around the belief state MDP that they follow. Let $b_\omega(u)$ denote a belief state, which is attained after being left passive for u time steps if the observation last received (when the arm was last pulled) was ω. Here the value $b_\omega(u)$ represents the belief, i.e., the probability that the agent is in the 'good' state. Let $\mathcal{B}$ denote the set of all possible belief states, which we organize into $\|\Omega\|$ chains, one chain for each possible observation as shown in Fig. 2. In this arrangement, when passive, the MDP transitions to the next belief state (on the right) in the same chain and when active, it jumps to one of the 'reset' states (shown in black). The MDP resets to the chain starting from the $b_\omega(1)$ state if an observation ω was observed as a result of the intervention. The reset probability is thus simply the probability of observing ω, which in turn, directly depends on the current belief state as shown in Fig. 3. The belief update when starting from an initial belief b and passive for u time steps, can be obtained via the standard belief update (as shown in [17])and is given by:

$$\tau_u(b) = \frac{P^p_{0,1} - (P^p_{1,1} - P^p_{0,1})^u (P^p_{0,1} - (1 + P^p_{0,1} - P^p_{1,1})b)}{(1 + P^p_{0,1} - P^p_{1,1})} \quad (1)$$

We use $\tau(b)$ to denote the passive belief update when $u = 1$.

The Whittle Index heuristic for RMABs has been shown to display strong performance, however it involves two challenges. First, the theoretical guarantees on the performance are valid only if a technical condition—referred to as indexability—holds good, which we prove for our problem in subsection 4.1. Second, computation of the index itself is challenging and can be computationally expensive. We use the theoretical results of subsection 4.1, to devise a fast algorithm to compute the Whittle Index efficiently, which we present in Subsection 4.2.

4.1 Indexability and Threshold Optimality

Definition 1 (Indexability)**.** *An RMAB is indexable if each arm of the RMAB is indexable. An arm is indexable if the set of passive-optimal states of the arm, given by $\mathcal{B}^*(m) = \{b : \exists\, \pi^* \in \Pi^*_m$, such that $\pi^*(b) = 0\}$ monotonically increases from $\emptyset$ to the entire state space as the subsidy, m increases from $-\infty$ to ∞.*

The optimal action is determined by comparing the passive and active value functions for a belief state b as given in Eq. 2 below and picking the action with a larger value.

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

A common strategy to proving indexability has been to first show that a special class of policies—'threshold policies'—are optimal for each arm under consideration. [18] has shown that if threshold policies are optimal (either forward or reverse threshold, defined below) then the RMAB is indexable; the same reasoning also applies to this work. This thus shifts the indexability heavy lifting to proving optimality of threshold policies for our problem.

Definition 2 (Threshold Policies)**.** *A policy π 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.*

Consider the reward of a belief state b to be given by a non-decreasing function, $\rho(b)$. Note that in a standard Collapsing Bandit [18], $\rho(b) = b$. Let $\Delta_a = (P^a_{11} - P^a_{01})$ and $\Delta_p = (P^p_{11} - P^p_{01})$ in all of the analysis in the rest of the paper. Let $\rho'_{max} = max_{b\in[0,1]} \frac{d(\rho(b))}{db}$, and $\rho'_{min} = min_{b\in[0,1]} \frac{d(\rho(b))}{db}$.

Theorem 1 (Forward Threshold Optimality). *Consider a belief-state MDP corresponding to an arm in an RMAB with some non-decreasing reward function given by $\rho(b)$ and transition matrix given by P. For any subsidy m, there is a* forward *threshold policy that is optimal if:*

$$\frac{\Delta_p(1 - \beta \max\{\Delta_p, \Delta_a\})}{\Delta_a(1 - \beta \min\{\Delta_p, \Delta_a\})} \geq \frac{\rho'_{max}}{\rho'_{min}} \quad (3)$$

Proof Sketch. Optimality of a forward threshold policy implies that if the optimal action at a belief b is passive, then it must be so for all $b' > b$. To accomplish this, we derive conditions which, if enforced, restrict the derivative of the passive action value function to be greater than the derivative of the active action value function w.r.t. b—thus implying forward threshold optimality. To arrive at such conditions, we first derive both upper and lower bounds on $V_m(b_1) - V_m(b_2)\ \forall\ b_1, b_2$. The key challenge is to then show that these bounds themselves imply tighter upper and lower bounds. We do this recursively for the new, tighter bounds and repeat this process an infinite number of times, arriving at tighter bounds each time and find that the bounds converge, which then leads us to the result. The full version of the proof is in Appendix A. □

Theorem 2 (Reverse Threshold Optimality). *Consider a belief-state MDP corresponding to an arm in an RMAB with some non-decreasing reward function given by $\rho(b)$ and transition matrix given by P. For any subsidy m, there is a* reverse *threshold policy that is optimal if:*

$$\frac{\Delta_p(1 - \beta \min\{\Delta_p, \Delta_a\})}{\Delta_a(1 - \beta \max\{\Delta_p, \Delta_a\})} \leq \frac{\rho'_{min}}{\rho'_{max}} \quad (4)$$

Proof Sketch. The proof follows similar reasoning as Thm.1. The final sufficiency condition obtained is such that when imposed, it restricts the derivative of the active action value function to

be always greater than the derivative of the passive action value function w.r.t. b. Complete proof is given under Appendix B. □

4.2 Fast Index Algorithm

Optimality of forward threshold policies forms the cornerstone of the fast Whittle Index computation algorithm. Recall that the Whittle Index of a belief state b is the infimum subsidy m such that the active and passive actions are both equally optimal to take at b. The key idea is to express the passive (active) action value function for a belief state b in a closed form by leveraging the forward threshold optimal structure.

The natural constraints imposed on the transition matrix at each arm (as mentioned in Sec. 3) ensure that $\tau_u(b)$ is a monotonic function of u. The fast algorithm presented below is guaranteed to be optimal for patients (RMAB arms) whose belief monotonically decreases with time (u) and for whom forward threshold policies are optimal. A forward threshold policy with a belief threshold of b_{th} induces a Markov chain over the belief states as shown in Fig. 3. Such a b_{th} determines a tuple of thresholds, $\overline{U}(b_{th}) = (u_0, u_1, \ldots u_{\|\Omega\|-1})$, where $b_\omega(u_\omega)$ specifies the threshold state for the chain corresponding to the observation ω. The threshold belief state is the first belief state of the chain where the optimal action is active. For the two-observation case, let (u_0, u_1) be the thresholds corresponding to the 0 and 1 chains respectively. A forward threshold policy with thresholds (u_0, u_1) induces a corresponding visit frequency $f^{(u_0,u_1)}(b)$ over the belief states. This $f^{(u_0,u_1)}(b)$ is the eigenvector solution for the equation $fM = f$, where M is the state transition matrix over the belief states. $M_{bb'}$ denotes the transition probability from belief state b to belief state b' and is completely determined by thresholds (u_0, u_1) as:

$$M_{bb'} = \begin{cases} 1 & \text{if } b' = \tau(b) \text{ and } b' \geq b_\omega(u_\omega) \text{ for } \omega \in \{0,1\} \\ b & \text{if } b' = b_1(1) \text{ and } b = b_\omega(u_\omega) \text{ for } \omega \in \{0,1\} \\ 1-b & \text{if } b' = b_0(1) \text{ and } b = b_\omega(u_\omega) \text{ for } \omega \in \{0,1\} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The visit frequencies $f^{(u_0,u_1)}(b)$ so determined, coupled with the known reward function $\rho(b)$, determine the overall reward of this threshold policy with a subsidy m, under the average reward criterion, given by $J_{m,\rho}^{(u_0,u_1)} = \sum_{b\in\mathcal{B}} f^{(u_0,u_1)}(b)\big(\rho(b) + m.\mathbb{1}_{\{b>b_{th}\}}\big)$.

For a belief state b, the active and passive action value functions correspond to the average rewards of two threshold policies with thresholds of b and $b+\epsilon$ (where $\epsilon \to 0$) respectively. Thus, finding the Whittle Index for which the active and passive value functions are equal is same as finding the subsidy m that satisfies $J_{m,\rho}^{\overline{U}(b)} = J_{m,\rho}^{\overline{U}(b+\epsilon)}$. Note that changing the threshold to $b+\epsilon$ affects the threshold belief state only on the current chain. We use this idea to construct the fast Whittle Index computation algorithm (Alg.1).

4.3 Application to Collapsing Bandits

Our theoretical results also generalize and improve upon the current state-of-the-art guarantees explored for the HMIP, as we demonstrate in this section. Collapsing bandits (CoBs) [18] are a sub-case of the risk-sensitive bandits considered in this paper, with reward function $\rho(b) = b$. The conditions of Thms. 1 and 2 yield novel sufficiency conditions when $\rho(b) = b$, that are wider than those

Algorithm 1: Risk-sensitive Index Computation Algorithm

```
Initialize pointers to heads of chains, u_0 = 1, u_1 = 1.
while u_0 < T or u_1 < T do
    Compute m_1 := m such that J_{m,ρ}^{(u_0,u_1)} = J_{m,ρ}^{(u_0,u_1+1)}
    Compute m_0 := m such that J_{m,ρ}^{(u_0,u_1)} = J_{m,ρ}^{(u_0+1,u_1)}
    Set i = arg min{m_0, m_1} and W(b_i(u_i)) = m_i
    Increment u_i
end while
```

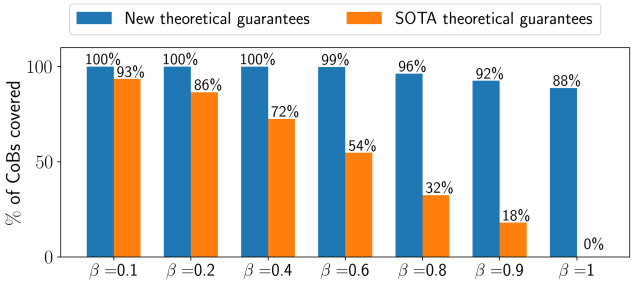


Figure 4: For $\rho(b) = b$, the theoretical guarantees presented in this paper hold for a wider range of processes as compared to the state-of-the-art conditions of Mate et al. [18].

presented in Mate et al. [18]. For example, under the average reward criterion (or $\beta = 1$), as shown in Fig. 4, the conditions of Mate et al. [18] become vacuous, whereas the new conditions derived here guarantee indexability for 88% of the entire space of CoBs.

Theorem 3. *Consider a belief-state MDP corresponding to an arm in a standard Collapsing Bandit. For any subsidy m, there is a* forward *threshold policy that is optimal if:*

$$\Delta_a \leq \Delta_p \text{ and } \Delta_a + \Delta_p \leq \frac{1}{\beta} \tag{6}$$

Intuitively, this condition requires that the action impact of both, passive and active actions in the "bad" state must not be too low (ensuring Δ_a and Δ_p are not too large) and further, the active action impact must be large (making Δ_a small). To prove the theorem, we show using simple algebraic manipulations that the condition of Eq. 6 satisfies the condition of Thm.1 when $\rho(b) = b$. Complete details of the proof are available in Appendix C.

Theorem 4. *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 if:*

$$\Delta_p \leq \Delta_a \text{ and } \Delta_p + \Delta_a \leq \frac{1}{\beta} \tag{7}$$

Intuitively, this condition requires that the action impact under both, passive and active actions in the "bad" state must not be too small (ensuring Δ_a and Δ_p are not too large) and further, the passive action impact must be large (making Δ_p smaller than Δ_a).

Note that both Thm. 3 and Thm. 4 define conditions for the discounted reward case, however, substituting $\beta = 1$ yields the sufficient conditions for the average reward criterion because the MDP is value-bounded (proof using Dutta's Theorem [6] is given in Appendix E).

Corollary 1. *Collapsing Bandits are indexable if:*

$$\Delta_p + \Delta_a \leq \frac{1}{\beta}. \tag{8}$$

From Thms. 3 and 4, we see that all CoBs satisfying the above condition have at least either a forward threshold policy or a reverse threshold policy as optimal. From Thm. 1 of Mate et al. [18], this implies that they must be indexable.

Corollary 2. *Collapsing bandits are indexable under either the average reward or the discounted reward criteria (for any β) if*

$$\Delta_p + \Delta_a \leq 1. \tag{9}$$

Remark 1. *Corollary 2 proves that Conjecture 1 of [18] must be true for at least 88% instances of Collapsing Bandits.*

Remark 2. *For $\beta < \frac{1}{2}$, the condition of Corollary 1 reduces to being "Always True", thus subsuming the previous results of an indexability guarantee for $\beta < \frac{1}{2}$ established by Qian et al. [25] and others.*

5 HANDLING IMPRECISE OBSERVATIONS

Real-world patients may misrepresent their adherence state or may sometimes simply not answer the CHW's calls, especially when not adhering to the prescribed dosage. In such cases, the intervention cannot be fully delivered, nor can the latent state be perfectly observed. We account for these uncertainties stemming from 'imprecise' observations by absorbing it in our RMAB planning framework, making it more amenable to real-world deployment.

5.1 Belief Dynamics

When precise binary observations of 'good' or 'bad' are unavailable, the planner may not get to directly observe and make confident conclusions about the latent state of the patient. Instead, the planner may only receive an observation $\omega \in \Omega$ that she associates uniquely with a corresponding likely belief about the patient's latent state in the next step using her previous historical experience and field expertise. For example, in practice, for $\|\Omega\| = 4$, these may correspond to either a confident positive, hesitant positive, a negative or no response from the patient. We remove the reliance on perfect observations from the patient, by including the human planner in the loop and allowing her to define her own belief state MDP for the patient, including the set of possible observations Ω as well as their respective reset belief states, P^a_ω. The observation probabilities and reset dynamics are explained further below.

We assume the planner observes an observation from the observation set $\Omega = \{0, 1, ...\|\Omega\| - 1\}$ every time a patient is intervened upon. We define the observation function, $\Theta_\omega(b)$ as the probability that the planner observes the evidence ω from the arm, when in a belief state b prior to the intervention. Thus, naturally the sum of the observation functions over all possible evidences must be equal to 1, giving: $\sum_{\omega=0}^{\omega=\|\Omega\|-1} \Theta_\omega(b) = 1$. Such an observation function can be either estimated by the planner directly or obtained via an emission matrix, either of which is specified by the planner from her historical experience. Such an emission matrix (and consequently the observation function) may be uniquely defined for each patient. Let $\mathbb{E}$ denote the emission matrix of a patient, as given by $\mathbb{E} = \begin{bmatrix} e_{00} & e_{01} \dots & e_{0\|\Omega\|-1} \\ e_{10} & e_{11} \dots & e_{1\|\Omega\|-1} \end{bmatrix}$ where $e_{s\omega}$ represents the probability of emitting the observation ω when the true state of the patient is s. For such an emission matrix $\mathbb{E}$, the corresponding observation function $\Theta_\omega(b)$, can then be obtained as: $\Theta_\omega(b) = \mathbb{P}(\omega|b) = be_{1\omega} + (1-b)e_{0\omega}$ Note that here $\Theta_\omega(b)$ is a linear in b and has a derivative independent of b, given by $\Theta'_\omega(b) = (e_{1\omega} - e_{0\omega}) = \Delta_{e\omega}$ (say).

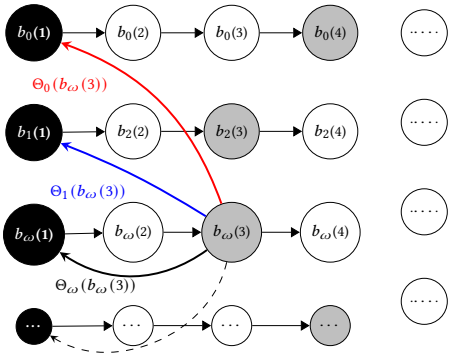


Figure 5: Multiple observations lead to a multiple-chain organization of belief states, with each observation having its corresponding reset state. An active action resets the belief state to $b_\omega(1)$ if observation ω is observed.

The planner defines a unique, fixed reset state P^a_ω for each observation, $\omega \in \Omega$. When the planner intervenes on a patient and receives an observation ω, the patient's belief state resets to P^a_ω, independent of the current belief. Further, given that the observation ω appears with a probability $\Theta_\omega(b)$ as established earlier, the passive and active action value functions can now be expressed as:

$$V_m(b) = \max \begin{cases} m + \rho(b) + \beta V_m(\tau(b)) \text{...passive} \\ \rho(b) + \beta\big(\sum_\omega \Theta_\omega(b).V_m(P^a_\omega)\big) \text{...active} \end{cases} \tag{10}$$

where $\sum_{\omega=0}^{\omega=\|\Omega\|-1} \Theta_\omega(b) = 1$

5.2 Threshold Optimality

For the setting with two possible observations ($\|\Omega\| = 2$), we derive conditions, which, if satisfied, guarantee the optimality of forward and reverse threshold policies as in previous sections. Let $\omega = 1$ ($\omega = 0$) be the observation corresponding to a positive (negative) response to the intervention and have a reset belief state of $P^a_1(P^a_0)$. The observation functions $\{\Theta_\omega(b)\}_{\omega=0,1}$ can be expressed using a single parameter and given by $\Theta_1(b) = \Theta(b)$ and $\Theta_0(b) = 1 - \Theta(b)$. We also let $\Delta_e = \Theta'(b) = (e_{11} - e_{01})$.

Theorem 5 (Forward Threshold Optimality). *Consider a belief-state MDP corresponding to an arm in an RMAB with some non-decreasing reward function given by $\rho(b)$, transition matrix given by P and an observation function, $\Theta(b)$ for a belief state b. For any subsidy m, there is a* forward *threshold policy that is optimal if:*

$$\frac{\Delta_p(1 - \beta \max\{\Delta_p, (\Delta_a.\Delta_e)\})}{\Delta_a(1 - \beta \min\{\Delta_p, (\Delta_a.\Delta_e)\})} \geq \frac{\rho'_{max}}{\rho'_{min}} \tag{11}$$

where $\Delta_e = \Theta'(b)$ for a linear $\Theta(b)$ such as in the example above.

Theorem 6 (Reverse Threshold Optimality). *Consider a belief-state MDP corresponding to an arm in an RMAB with some non-decreasing reward function given by $\rho(b)$, transition matrix given by P and an observation function, $\Theta(b)$ for a belief state b. For any subsidy m, there is a* reverse *threshold policy that is optimal if:*

$$\frac{\Delta_p(1-\beta\min\{\Delta_p,(\Delta_a.\Delta_e)\})}{\Delta_a(1-\beta\max\{\Delta_p,(\Delta_a.\Delta_e)\})} \leq \frac{\rho'_{min}}{\rho'_{max}} \tag{12}$$

where $\Delta_e = \Theta'(b)$ for a linear $\Theta(b)$ such as in the example above.

6 EXPERIMENTAL EVALUATION

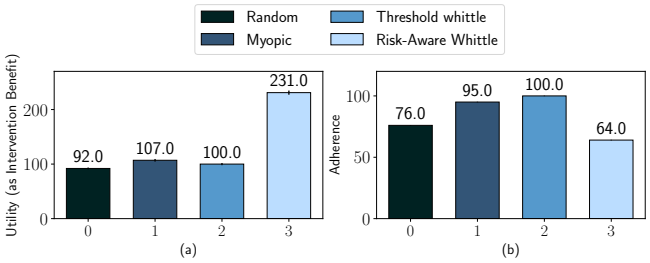


Figure 6: Risk-Aware Whittle optimizes for the objectives the planner cares about, and avhieves much higher utility than Threshold Whittle, even while scoring lower on average adherence—a metric that previous approaches to the HMIP focus on.

We explore several suitable reward functions $\rho(b)$, tailor-made for each of the specific CHW planning considerations at hand. We demonstrate the effectiveness of our approach for addressing at least three real-world objectives evaluating our algorithm on both real and simulated data. We use real tuberculosis medication adherence monitoring data, consisting of records of patients in Mumbai, India, obtained from [15] and run simulations following the same data imputation steps as [18] for consistency. We compare the following algorithms: **Risk-Aware Whittle** is the algorithm presented in this paper. **Threshold Whittle** is the SOTA fast algorithm presented in [18], our primary baseline, which has been shown to display near-optimal performance. **Random** selects k patients to call at random. **Myopic** calls the k patients that maximize the expected adherence at the immediate next time step. **'Everybody'** is an unattainable upper baseline that simulates the effect of intervening on everybody everyday. Wherever applicable, we measure performance using 'intervention benefit', which scales the reward from 0% (corresponding to no interventions) to 100% (corresponding to Threshold Whittle unless indicated otherwise) and is given by $I.B.(ALG) = \frac{\overline{R}^{ALG} - \overline{R}^{\text{No intervention}}}{\overline{R}^{\text{Threshold Whittle}} - \overline{R}^{\text{No intervention}}}$ where $\overline{R}$ is the average reward of the algorithm. All results are measured over 50 independent trials.

6.1 Risk-sensitive planning

Real-world health workers may be risk-averse and prefer to consolidate the well-being of at least some of their patients rather than being unsure about the health outcomes of the entire patient cohort. For example, in case of the TB treatment, the medication program may be effective only if completed with a high degree of adherence. In such a case, the CHW may want to prioritize maximizing the number of patients who complete the program with a high adherence rate. To account for risk-averseness, we employ a convex reward function, $\rho(b) = e^{\lambda b}$ for $\lambda = 20$ in our algorithm and measure its impact. We run a simulation for $N = 100$ patients and $k = 20$ calls per day, with patient transition matrices drawn from a fixed simulated distribution.

Fig. 6 shows the tradeoff between the utility to the planner and the average adherence of the patient cohort. Algorithms studied in previous work only focus on maximizing the average patient adherence, which unfortunately may not be perfectly aligned with the objectives the CHWs value the most. Our algorithm, on the other hand directly optimizes for the CHW's objectives, and achieves a much higher utility than the state-of-the-art, Threshold Whittle even while yielding a lower average adherence, which is less valuable to the planner, and is thus a bad yardstick to measure performance.

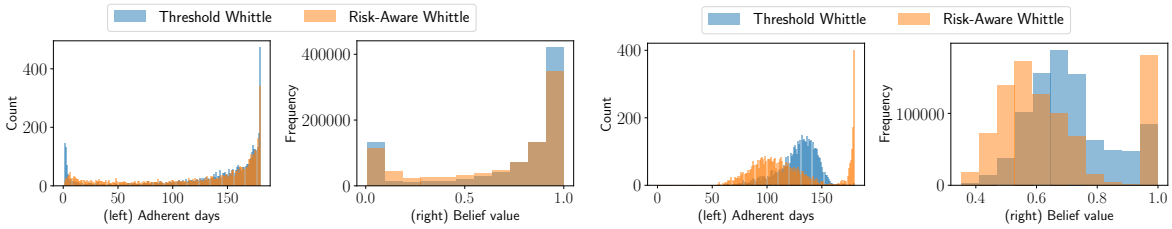
Fig. 7b(right) shows the histogram of time spent by patients in a belief state over the duration of the program. The convex reward function imposed by Risk-Aware Whittle "scoops out" patients from the moderate belief zone, pushing part of these towards the high-belief zone, boosting the number of patients adhering with high confidence, towards realizing the objectives the planner cares about. This effect is also manifested in the adherence histogram of Fig. 7b(left), which shows the total days adhered to on the x-axis and the corresponding number of patients with that score on the y-axis. Fig. 8(b) plots the number of patients completing the program high degree of adherence (defined as adherent for > 90% days in the program). Risk-aware Whittle shows a steep increase over Threshold Whittle in the number high-adherence patients.

6.2 Fairness towards Patients: Real Data

A specific fairness concern faced by CHW planning algorithms is that some patients may be completely ignored by the algorithm because it deems them less valuable to intervene on. Even though it may be optimal when measured with the yardstick of average cohort outcome, such an algorithm may be socially unacceptable.

To address this issue, we use a concave reward function soliciting risk-seeking behavior through which the planner intervenes on patients that may be sub-optimal in expectation. Such a reward function imposes a large negative reward on lower belief values, making the algorithm intervene on these patients in a bid to bring them to moderate belief states. We employ $\rho(b) = -e^{(\lambda(1-b))}$ with $\lambda = 20$ as the concave reward function. We use the real TB adherence data from Mumbai to draw patient transition matrices for $N = 100$ patients and a budget $k = 20$ calls per day to run the simulation.

Fig. 7a(right) shows the histogram of time spent by patients in possible belief states. The effect of the risk-seeking reward function is to transfer patients from very low and very high belief values and to spread them over the moderate belief values. Fig. 7a(left) plots the histogram of adherence of patients and shows the effectiveness of this algorithm in nearly wiping out the spike at $x = 0$, representing the patients who never interact with the CHW. This is corroborated by Fig. 8(a) which plots the number of patients with



(a) Handling Fairness Concerns **(b) Handling Risk-Sensitivity**

Figure 7: (a) Threshold Whittle ignores many patients leaving them at a very low adherence (see blue spike at $x = 0$). Risk-Aware Whittle removes the blue spike, redistributing these patients towards moderate belief values. (b–left:) Risk-Aware Whittle boosts the number of patients completing treatment with high adherence rates. (b–right:) Risk-Aware Whittle better caters to risk-averse planners, who prefer having patients in the high belief zone.

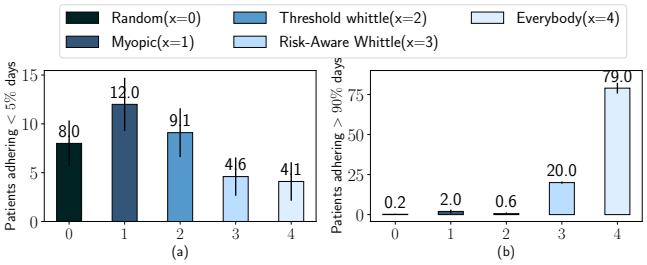


Figure 8: Risk-Aware Whittle is significantly better at tackling the specific concerns of the CHW. (a) shows a sharp decrease in the number of patients with a severely low adherence rate. (b) shows a significant jump in the number of patient finishing the treatment with a high adherence score.

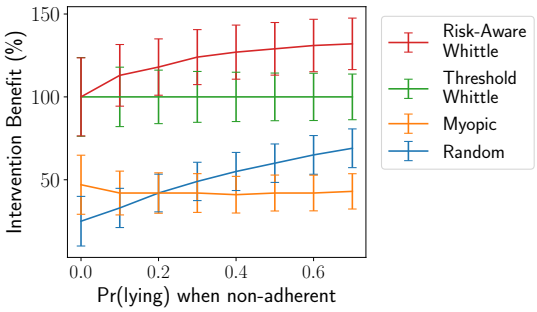


Figure 9: Risk-Aware Whittle beats Threshold Whittle when patients misrepresent their adherence states.

very low adherence (defined as $< 5\%$ days of adherence) and shows substantial decrease under the Risk-Aware Whittle algorithm as against the Threshold Whittle algorithm.

6.3 Imprecise Observations

We next evaluate empirically, the performance of our algorithm when precise observations of their latent states are not available from patients like in real-world. To model this, we assume patients emit two possible observations: '0' (denoting a negative response such as not answering the CHW's call at all or responding prevaricatively) and '1' indicating a positive response to the intervention. We simulate using an emission matrix given by $\mathbb{E} = \begin{bmatrix} e_{00} = 1 - P_{lie0} & e_{01} = P_{lie0} \\ e_{10} = P_{lie1} & e_{11} = 1 - P_{lie1} \end{bmatrix}$ parameterized by $P_{lie0(1)}$, capturing the probability that patients misrepresent when in a true latent state of $0(1)$. In Fig. 9 we fix $p_{lie1} = 0.01$ as the small probability that the intervention goes unanswered when the patient is adherent and vary p_{lie0}, from $[0, 0.7]$ the probability of giving a false observation when non-adherent . We measure the performance on the y-axis, as improvement in the overall adherence in terms of "intervention benefit" (defined previously), normalized w.r.t 'Threshold Whittle' as the baseline fixed at $y = 100\%$. Figure shows, our algorithm outperforms Threshold Whittle, which doesn't account for imprecise observations and thus grapples with incorrect belief values.

7 DISCUSSION AND CONCLUSION

Mitigating bias in socio-technical systems such as ours, is an important issue [9, 11]. We rely on the human in the loop to ensure that more complex human objectives can be addressed, and provide flexibility to admit other objectives, which for example, may be more ethical or fair as against the specific examples considered here. The human-in-the-loop and other stakeholders situated in the community may be able to better assess the needs of the community and may collectively provide a better perspective on the objective.

To conclude, we propose a new RMAB-based planning framework that allows for planning health interventions while accommodating the real-world objectives of the health workers effectively. We prove theoretical guarantees on the performance of our algorithm that apply to a more general class and are stronger than the guarantees for the specific sub-case studied previously. Through empirical results, we demonstrate the effectiveness of our algorithm in achieving improved health outcomes, addressing three real-world planning challenges faced by the health workers.

ACKNOWLEDGMENTS

This work was supported in part by the Army Research Office by MURI grant number W911NF1810208. A.P. was supported by the Harvard Center for Research on Computation and Society. We also thank Haipeng Cheng for suggesting edits to the initial draft of the paper and Shahin Jabbari for helping address the ethical implications of the work.

REFERENCES

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

[2] Guido Biele, Ido Erev, and Eyal Ert. 2009. Learning, risk attitude and hot stoves in restless bandit problems. *Journal of mathematical psychology* 53, 3 (2009), 155–167.

[3] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. 2021. Learning Index Policies for Restless Bandits with Application to Maternal Healthcare. *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)* (2021).

[4] R. Chen and I. Paschalidis. 2018. Learning Optimal Personalized Treatment Rules Using Robust Regression Informed K-NN. In *NIPS Machine Learning for Health Workshop*.

[5] J. B. Christopher, A. Le May, S. Lewin, and D. A. Ross. 2011. 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).

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

[7] Edgar N Gilbert. 1960. Capacity of a burst-noise channel. *Bell system technical journal* 39, 5 (1960), 1253–1265.

[8] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride. 2006. Some indexable families of restless bandit problems. *Adv. Appl. Probab.* 38, 3 (2006), 643–672.

[9] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 90–99.

[10] Andy Haines, David Sanders, Uta Lehmann, Alexander K Rowe, Joy E Lawn, Steve Jan, Damian G Walker, and Zulfiqar Bhutta. 2007. Achieving child survival goals: potential contribution of community health workers. *The lancet* 369, 9579 (2007), 2121–2131.

[11] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[12] Y. Hsu. 2018. Age of Information: Whittle Index for Scheduling Stochastic Arrivals. In *IEEE International Symposium on Information Theory*.

[13] Kia Khezeli and Eilyan Bitar. 2017. Risk-sensitive learning and pricing for demand response. *IEEE Transactions on Smart Grid* 9, 6 (2017), 6000–6007.

[14] Jackson A Killian, Andrew Perrault, and Milind Tambe. 2021. Beyond "To Act or Not to Act": Fast Lagrangian Approaches to General Multi-Action Restless Bandits. *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (2021).

[15] 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*.

[16] P. Liao, K. Greenewald, P. Klasnja, and S. Murphy. 2019. Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity. In *JSM*.

[17] 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.

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

[19] Rahul Meshram, D Manjunath, and Aditya Gopalan. 2018. On the Whittle index for restless multiarmed hidden Markov bandits. *IEEE Trans. Automat. Control* 63, 9 (2018), 3046–3053.

[20] Siddharth Nishtala, Harshavardhan Kamarthi, Divy Thakkar, Dhyanesh Narayanan, Anirudh Grama, Ramesh Padmanabhan, Neha Madhiwalla, Suresh Chaudhary, Balaraman Ravindra, and Milind Tambe. 2020. Missed calls, Automated Calls and Health Support: Using AI to improve maternal health outcomes by increasing program engagement. *arXiv preprint arXiv:2006.07590* (2020).

[21] World Health Organization et al. 2013. *Using lay health workers to improve access to key maternal and newborn health interventions in sexual and reproductive health*. Technical Report. World Health Organization.

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

[23] M. E. Pollack, C. E. McCarthy, S. Ramakrishnan, I. Tsamardinos, L. Brown, S. Carrion, D. Colbry, C. Orosz, and B. Peintner. 2002. Autominder: A planning, monitoring, and reminding assistive agent. In *7th International Conference on Intelligent Autonomous Systems*.

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

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

[26] S. Shin, J. Furin, J. Bayona, K. Mate, J. Y. Kim, and P. Farmer. 2004. Community-based treatment of multidrug-resistant tuberculosis in Lima, Peru: 7 years of experience. *Soc. Sci. Med.* 59, 7 (2004), 1529–1539.

[27] B. Sombabu, A. Mate, D. Manjunath, and S. Moharir. 2020. Whittle Index for AoI-Aware Scheduling. In *2020 12th International Conference on Communication Systems & Networks (COMSNETS)*. IEEE.

[28] Vishrant Tripathi and Eytan Modiano. 2019. A whittle index approach to minimizing functions of age of information. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 1160–1167.

[29] UNICEF and WHO. 1978. Declaration of Alma Ata. *International Conference on Primary Health Care, Alma Ata, USSR* (1978).

[30] R. R. Weber and G. Weiss. 1990. On an index policy for restless bandits. *J. Appl. Probab.* 27, 3 (1990), 637–648.

[31] K. J. Wells, J. S. Luque, B. Miladinovic, N. Vargas, Y. Asvat, R. G. Roetzheim, and A. Kumar. 2011. Do Community Health Worker Interventions Improve Rates of Screening Mammography in the United States? A Systematic Review. *Cancer Epidem. Biomar.* 20, 8 (2011), 1580–1598.

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

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

[34] Anne Witmer, Sarena D Seifer, Leonard Finocchio, Jodi Leslie, and Edward H O'Neil. 1995. Community health workers: integral members of the health care work force. *American journal of public health* 85, 8_Pt_1 (1995), 1055–1058.

[35] Jianyu Xu, Lujie Chen, and Ou Tang. 2020. An Online Algorithm for the Risk-Aware Restless Bandit. *European Journal of Operational Research* (2020).

SUPPLEMENTARY MATERIAL FOR: "RISK-AWARE INTERVENTIONS IN PUBLIC HEALTH: PLANNING WITH RESTLESS MULTI-ARMED BANDITS", AAMAS 2021

A PROOF OF THEOREM 1

Theorem 1 (Forward Threshold Optimality). *Consider a belief-state MDP corresponding to an arm in an RMAB with some non-decreasing reward function given by $\rho(b)$ and transition matrix given by P. For any subsidy m, there is a* forward *threshold policy that is optimal if:*

$$\frac{\Delta_p(1-\beta\max\{\Delta_p,\Delta_a\})}{\Delta_a(1-\beta\min\{\Delta_p,\Delta_a\})} \geq \frac{\rho'_{max}}{\rho'_{min}} \tag{3}$$

Proof. We start with presenting three facts and proving several lemmas that underpin the proofs of Thms. 1 and 2.

Fact 1. $\frac{d(\tau(b))}{db} = P^p_{11} - P^p_{01}$

Fact 2. $\forall b, b'$ s.t. $b \geq b'$, $\tau(b) \geq \tau(b')$.

Fact 3. $\forall b, b'$ s.t. $b \geq b'$, $\tau(b) - \tau(b') = (P^p_{11} - P^p_{01})(b - b')$.

Lemma 1. $V_m(b_1) - V_m(b_2) \geq \rho'_{min}(b_1 - b_2) \forall\ b_1, b_2$ *s.t.* $b_1 \geq b_2$

Proof. We will prove this via induction, where the base case will be a one-step value function. For the iterative case, we will show that the t-step value function assumption implies the t+1-step inductive value function hypothesis. It is sufficient to compare the value functions for each case corresponding to each action being the optimal. If the true optimal action for one of the value functions is passive and the other active, then the bound can still be established by flipping the action of one of the value functions as needed. This gives:

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

$$m + \rho(b_1) - (m + \rho(b_2)) = \rho(b_1) - \rho(b_2) \qquad \text{passive} \tag{13}$$

$$\rho(b_1) - \rho(b_2) = \rho(b_1) - \rho(b_2) \qquad \text{active} \tag{14}$$

is clearly satisfied. Now assume $V^t_m(b_1) - V^t_m(b_2)) \geq \rho'_{min}(b_1 - b_2)$. Then $V^{t+1}_m(b_1) - V^{t+1}_m(b_2)$
Case 1 (both passive):

$$\begin{aligned}
&= m + \rho(b_1) + \beta V^t_m(\tau(b_1)) - (m + \rho(b_2) + \beta V^t_m(\tau(b_2))) \\
&= \rho(b_1) - \rho(b_2) + \beta\Big(V^t_m(\tau(b_1)) - V^t_m(\tau(b_2))\Big) \\
&\geq \rho(b_1) - \rho(b_2) + \beta\rho'_{min}(\tau(b_1) - \tau(b_2)) \\
&\geq \rho(b_1) - \rho(b_2)
\end{aligned} \tag{15}$$

Case 2 (both active):

$$\begin{aligned}
&= \rho(b_1) - \rho(b_2) + \beta\Big((b_1 - b_2)V^t_m(P^a_{1,1}) + (b_2 - b_1)V^t_m(P^a_{0,1})\Big) \\
&= \rho(b_1) - \rho(b_2) + \beta\Big((b_1 - b_2)(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1}))\Big) \\
&\geq \rho'_{min}(b_1 - b_2) + \beta\Big((b_1 - b_2)(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1}))\Big) \\
&\geq \rho'_{min}(b_1 - b_2)
\end{aligned} \tag{16}$$

□

Lemma 2. *If* $\forall\ b_1, b_2$ *s.t.* $b_1 \geq b_2$, $\exists\ \kappa$ *such that* $V_m(b_1) - V_m(b_2) \geq \kappa\rho'_{min}(b_1 - b_2)$, *then, for* $\alpha = \min\{\Delta_a, \Delta_p\}$*:*

$$V_m(b_1) - V_m(b_2) \geq \rho'_{min}(1 + \beta\alpha\kappa)(b_1 - b_2) \tag{17}$$

Proof. Using Eq. 2, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}\frac{d(\tau(b))}{db} & \text{passive} \\ \rho'(b) + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) & \text{active} \end{cases} \tag{18}$$

Case 1 (passive):

$$= \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}(P^p_{1,1} - P^p_{0,1}) \tag{19}$$

$$= \rho'(b) + \beta\lim_{\delta\to 0}\frac{V_m(\tau(b)+\delta) - V_m(\tau(b))}{\tau(b)+\delta-\tau(b)}(P^p_{1,1} - P^p_{0,1}) \tag{20}$$

$$\geq \rho'(b) + \beta\kappa\rho'_{min}(P^p_{1,1} - P^p_{0,1}) \tag{21}$$

$$\geq \rho'_{min} + \beta\kappa\rho'_{min}(P^p_{1,1} - P^p_{0,1}) \tag{22}$$

$$\geq \rho'_{min}(1 + \beta\alpha\kappa) \tag{23}$$

Case 2 (active):

$$= \rho'(b) + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \tag{24}$$

$$\geq \rho'(b) + \beta\kappa\rho'_{min}(P^a_{1,1} - P^a_{0,1}) \tag{25}$$

$$\geq \rho'_{min} + \beta\kappa\rho'_{min}(P^a_{1,1} - P^a_{0,1}) \tag{26}$$

$$\geq \rho'_{min}(1 + \beta\alpha\kappa) \tag{27}$$

Thus,

$$\begin{aligned}
&\implies \frac{d(V_m(b))}{db} \geq \rho'_{min}(1 + \beta\alpha\kappa) \\
&\implies \int_{b_2}^{b_1}\frac{d(V_m(b))}{db}db \geq \int_{b_2}^{b_1}\rho'_{min}\Big(1 + \beta\alpha\kappa\Big)db \\
&\implies V_m(b_1) - V_m(b_2) \geq \rho'_{min}\big(1 + \beta\alpha\kappa\big)(b_1 - b_2)
\end{aligned} \tag{28}$$

□

Lemma 3. $V_m(b_1) - V_m(b_2) \geq \frac{\rho'_{min}(b_1-b_2)}{1-\beta\alpha}\forall\ b_1, b_2$ *s.t.* $b_1 \geq b_2$

Proof. Consider the function, $f(x) = 1 + \beta\alpha x$ and let $f^n(x) \coloneqq \underbrace{f(f(...f(x)))}_{f(.)\text{ applied }n\text{ times}}$. We show using induction that:

$$V_m(b_1) - V_m(b_2) \geq f^n(1)\,\rho'_{min}(b_1 - b_2) \forall n \in \mathbb{W},\ \forall\ b_1, b_2 \text{ s.t. } b_1 \geq b_2 \tag{29}$$

Consider the base case, $n = 0$. Eq. 29 reduces to the statement of Lemma 1 with $f^0(1) = 1$, and thus holds true. For the inductive case, we assume Eq.29 to be true for some n and then show that it must also be true for $n + 1$, as follows:

$$\begin{aligned}
&\text{If } V_m(b_1) - V_m(b_2) \geq f^n(1).\rho'_{min}(b_1 - b_2), \text{ then} \\
&\implies V_m(b_1) - V_m(b_2) \geq f\Big(f^n(1)\Big).\rho'_{min}(b_1 - b_2). \text{ using Lemma 2} \\
&\implies V_m(b_1) - V_m(b_2) \geq f^{n+1}(1)\rho'_{min}(b_1 - b_2)
\end{aligned} \tag{30}$$

Thus we show Eq. 29 to be true for all n. We note that the sequence $\{f^n(1)\}_{n=0}^{\infty}$ is strictly increasing and bounded, and thus the sequence converges. The point of convergence can be obtained as follows:

$$\begin{aligned}
&\text{Let the sequence converge to } f^\infty = \lim_{n\to\infty} f^n(1) \\
\implies & f\big(f^\infty\big) = f\big(\lim_{n\to\infty} f^n(1)\big) = \lim_{n\to\infty} f^{n+1}(1) = \lim_{n\to\infty} f^n(1) = f^\infty \\
\implies & 1 + \beta\alpha f^\infty = f^\infty \\
\implies & 1 = f^\infty(1-\beta\alpha) \\
\implies & f^\infty = \frac{1}{(1-\beta\alpha)}
\end{aligned} \tag{31}$$

Resubstituting f^∞ in place of $f^n(1)$ in Eq.29 finally gives us the required result. □

Corollary 3. $\frac{d(V_m(b))}{db} \geq \frac{\rho'_{min}}{1-\beta\alpha}$

Proof. This follows from Lemma 3 by setting $b_1 = b+\delta, b_2 = b$ under the limit $\delta \to 0$. □

Lemma 4. $V_m(b_1) - V_m(b_2) \leq \frac{\rho'_{max}(b_1-b_2)}{1-\beta} \forall\ b_1, b_2\ s.t.\ b_1 \geq b_2$

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

$$m + \rho(b_1) - (m + \rho(b_2)) = \rho(b_1) - \rho(b_2) \leq \frac{\rho'_{max}(b_1-b_2)}{1-\beta}$$

both passive

$$\rho(b_1) - \rho(b_2) \leq \frac{\rho'_{max}(b_1-b_2)}{1-\beta} \text{ both active}$$

which are both clearly satisfied. Now assume $V_m^t(b_1) - V_m^t(b_2) \leq \frac{\rho'_{max}(b_1-b_2)}{1-\beta}$. Then, $V_m^{t+1}(b_1) - V_m^{t+1}(b_2)$
Case 1 (both passive):

$$\begin{aligned}
&= \big(m + \rho(b_1) + \beta V_m^t(\tau(b_1))\big) - \big(m + \rho(b_2) + \beta V_m^t(\tau(b_2))\big) \\
&= (\rho(b_1) - \rho(b_2)) + \beta\big(V_m^t(\tau(b_1)) - V_m^t(\tau(b_2))\big) \\
&\leq (\rho(b_1) - \rho(b_2)) + \beta\left(\frac{\rho'_{max}(\tau(b_1) - \tau(b_2))}{1-\beta}\right) \\
&\leq \rho'_{max}(b_1 - b_2) + \beta\left(\frac{\rho'_{max}(b_1-b_2)}{1-\beta}\right) \text{ by Fact 3} \\
&= \frac{\rho'_{max}(b_1-b_2)}{1-\beta}
\end{aligned} \tag{32}$$

Case 2 (both active):

$$\begin{aligned}
&= \Big(\rho(b_1) + \beta\big(b_1 V_m^t(P_{1,1}^a) + (1-b_1)V_m^t(P_{0,1}^a)\big)\Big) - \\
&\Big(\rho(b_2) + \beta\big(b_2 V_m^t(P_{1,1}^a) + (1-b_2)V_m^t(P_{0,1}^a)\big)\Big) \\
&= (\rho(b_1) - \rho(b_2)) + \beta\Big((b_1-b_2)\big(V_m^t(P_{1,1}^a) - V_m^t(P_{0,1}^a)\big)\Big) \\
&\leq (\rho(b_1) - \rho(b_2)) + \beta\left((b_1-b_2).\frac{\rho'_{max}(P_{1,1}^a - P_{0,1}^a)}{1-\beta}\right) \\
&\leq \rho'_{max}(b_1-b_2) + \beta\left(\frac{\rho'_{max}(b_1-b_2)}{1-\beta}\right) \\
&= \frac{\rho'_{max}(b_1-b_2)}{1-\beta}
\end{aligned} \tag{33}$$

□

Lemma 5. *If* $\forall\ b_1, b_2\ s.t.\ b_1 \geq b_2$, $\exists\ \kappa$ *such that* $V_m(b_1) - V_m(b_2) \leq \kappa\rho'_{max}(b_1-b_2)$, *then, for* $\gamma = \max\{\Delta_a, \Delta_p\}$:

$$V_m(b_1) - V_m(b_2) \leq \rho'_{max}(1+\beta\gamma\kappa)(b_1-b_2) \tag{34}$$

Proof. Using Equation 2, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}\frac{d(\tau(b))}{db} & \text{passive} \\ \rho'(b) + \beta(V_m(P_{1,1}^a) - V_m(P_{0,1}^a)) & \text{active} \end{cases} \tag{35}$$

Case 1 (passive):

$$\begin{aligned}
&= \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}(P_{1,1}^p - P_{0,1}^p) && (36)\\
&= \rho'(b) + \beta\lim_{\delta\to 0}\frac{V_m(\tau(b)+\delta) - V_m(\tau(b))}{\tau(b)+\delta-\tau(b)}(P_{1,1}^p - P_{0,1}^p) && (37)\\
&\leq \rho'(b) + \beta\kappa\rho'_{max}(P_{1,1}^p - P_{0,1}^p) && (38)\\
&\leq \rho'_{max}(1+\beta\gamma\kappa) && (39)
\end{aligned}$$

Case 2 (active):

$$\begin{aligned}
&= \rho'(b) + \beta(V_m(P_{1,1}^a) - V_m(P_{0,1}^a)) && (40)\\
&\leq \rho'(b) + \beta\kappa\rho'_{max}(P_{1,1}^a - P_{0,1}^a) && (41)\\
&\leq \rho'_{max} + \beta\rho'_{max}\gamma\kappa && (42)\\
&\leq \rho'_{max}(1+\beta\gamma\kappa) && (43)
\end{aligned}$$

Thus,

$$\begin{aligned}
&\implies \frac{d(V_m(b))}{db} \leq \rho'_{max}(1+\beta\gamma\kappa) \\
&\implies \int_{b_2}^{b_1}\frac{d(V_m(b))}{db}db \leq \int_{b_2}^{b_1}\rho'_{max}\Big(1+\beta\gamma\kappa\Big)db \\
&\implies V_m(b_1) - V_m(b_2) \leq \rho'_{max}\big(1+\beta\gamma\kappa\big)(b_1-b_2)
\end{aligned} \tag{44}$$

□

Lemma 6. $V_m(b_1) - V_m(b_2) \leq \frac{\rho'_{max}(b_1-b_2)}{1-\beta\gamma} \forall\ b_1, b_2\ s.t.\ b_1 \geq b_2$

Proof. We use an approach similar to the proof of Lemma 3. Consider the function, $g(x) = 1+\beta\gamma x$ and let $g^n(x) := \underbrace{g(g(...g(x)))}_{g(.)\text{ applied }n\text{ times}}$.

We show using induction that, $\forall n \in \mathbb{W},\ \forall\ b_1, b_2$ s.t. $b_1 \geq b_2$:

$$V_m(b_1) - V_m(b_2) \leq g^n\Big(\frac{1}{1-\beta}\Big)\rho'_{max}(b_1 - b_2) \tag{45}$$

Consider the base case, $n = 0$. Eq. 45 reduces to the statement of Lemma 11 with $g^0\big(\frac{1}{1-\beta}\big) = \frac{1}{1-\beta}$, and is thus true. For the inductive case, we assume Eq.45 to be true for some n and then show that it must also be true for $n+1$, as follows:

$$\begin{aligned}
&\text{If } V_m(b_1) - V_m(b_2) \leq g^n\Big(\frac{1}{1-\beta}\Big)\rho'_{max}(b_1 - b_2), \text{ then}\\
&\implies V_m(b_1) - V_m(b_2) \leq g\Big(g^n\Big(\frac{1}{1-\beta}\Big)\Big)\rho'_{max}(b_1 - b_2). \text{ by Lemma 12}\\
&\implies V_m(b_1) - V_m(b_2) \leq g^{n+1}\Big(\frac{1}{1-\beta}\Big)\rho'_{max}(b_1 - b_2)
\end{aligned} \tag{46}$$

Thus we show Eq. 45 to be true for all n. We note that the sequence $\{g^n(\frac{1}{1-\beta})\}_{n=0}^{\infty}$ is strictly decreasing and bounded, and thus the sequence converges. The point of convergence can be obtained as follows:

$$\begin{aligned}
&\text{Let the sequence converge to } g^\infty = \lim_{n\to\infty} g^n(\frac{1}{1-\beta})\\
&\implies g(g^\infty) = g\Big(\lim_{n\to\infty} g^n(\frac{1}{1-\beta})\Big) = \lim_{n\to\infty} g^{n+1}(\frac{1}{1-\beta}) = g^\infty\\
&\implies 1 + \beta\gamma g^\infty = g^\infty\\
&\implies 1 = g^\infty(1-\beta\gamma)\\
&\implies g^\infty = \frac{1}{(1-\beta\gamma)}
\end{aligned} \tag{47}$$

Resubstituting g^∞ in place of $g^n(\frac{1}{1-\beta})$ in Eq.45 finally gives us the required result. □

Corollary 4. $\frac{d(V_m(b))}{db} \leq \frac{\rho'_{max}}{1-\beta\gamma}$

Proof. This follows from Lemma 13 by setting $b_1 = b+\delta, b_2 = b$ under the limit $\delta \to 0$. □

Now we complete the proof for Thm.1 as follows:

$$Eq.3 \implies \Delta_p \geq \frac{\Delta_a(1-\beta\alpha)\rho'_{max}}{(1-\beta\gamma)\rho'_{min}} \tag{48}$$

$$\implies \frac{(P^p_{1,1} - P^p_{0,1})\rho'_{min}}{(1-\beta\alpha)} \geq \frac{\rho'_{max}(P^a_{1,1} - P^a_{0,1})}{(1-\beta\gamma)} \tag{49}$$

$$\implies \frac{(P^p_{1,1} - P^p_{0,1})\rho'_{min}}{(1-\beta\alpha)} \geq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma 13} \tag{50}$$

$$\implies (P^p_{1,1} - P^p_{0,1})\frac{d(V_m(b))}{db} \geq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Cor. 5} \tag{51}$$

$$\implies \frac{d(\tau(b))}{db}\frac{d(V_m(b))}{db} \geq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Fact 1} \tag{52}$$

$$\implies \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau b)}\frac{d(\tau(b))}{db} \geq \rho'(b) + \tag{53}$$

$$\beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \tag{54}$$

$$\implies \frac{d(V_m(b|a=0))}{d(b)} \geq \frac{d(V_m(b|a=1))}{d(b)} \tag{55}$$

□

B PROOF OF THEOREM 2

Theorem 2 (Reverse Threshold Optimality). *Consider a belief-state MDP corresponding to an arm in an RMAB with some non-decreasing reward function given by $\rho(b)$ and transition matrix given by P. For any subsidy m, there is a* reverse *threshold policy that is optimal if:*

$$\frac{\Delta_p(1-\beta\min\{\Delta_p,\Delta_a\})}{\Delta_a(1-\beta\max\{\Delta_p,\Delta_a\})} \leq \frac{\rho'_{min}}{\rho'_{max}} \tag{4}$$

Proof. Optimality of a reverse threshold policy implies that if the optimal action at a belief b is active, then it must be so for all $b' > b$. Similar to proof of Theorem 1, we approach this by deriving conditions which if imposed, restrict the derivative of the active action value function to be greater than the derivative of the passive action value function w.r.t. b — thus implying reverse threshold optimality. We show that the conditions of Theorem 2 satisfy this required property:

$$Eq.4 \implies \Delta_p \leq \frac{\Delta_a(1-\beta\gamma)\rho'_{min}}{(1-\beta\alpha)\rho'_{max}} \tag{56}$$

$$\implies \frac{(P^p_{1,1} - P^p_{0,1})\rho'_{max}}{(1-\beta\gamma)} \leq \frac{\rho'_{min}(P^a_{1,1} - P^a_{0,1})}{(1-\beta\alpha)} \tag{57}$$

$$\implies \frac{(P^p_{1,1} - P^p_{0,1})\rho'_{max}}{(1-\beta\alpha)} \leq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma 3} \tag{58}$$

$$\implies (P^p_{1,1} - P^p_{0,1})\frac{d(V_m(b))}{db} \leq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Cor. 5} \tag{59}$$

$$\implies \frac{d(\tau(b))}{db}\frac{d(V_m(b))}{db} \leq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Fact 1} \tag{60}$$

$$\implies \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau b)}\frac{d(\tau(b))}{db} \leq \rho'(b) + \tag{61}$$

$$\beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \tag{62}$$

$$\implies \frac{d(V_m(b|a=0))}{d(b)} \leq \frac{d(V_m(b|a=1))}{d(b)} \tag{63}$$

□

C PROOF OF THEOREM 3

Theorem 3. *Consider a belief-state MDP corresponding to an arm in a standard Collapsing Bandit. For any subsidy m, there is a* forward *threshold policy that is optimal if:*

$$\Delta_a \leq \Delta_p \text{ and } \Delta_a + \Delta_p \leq \frac{1}{\beta} \tag{6}$$

Proof. To prove this thoerem, we show that the condition of Eq. 6 satisfies the condition of Thm.1 when $\rho(b) = b$. Note that

$\rho'_{max} = \rho'_{min} = 1$.

$$
\begin{aligned}
&Eq.6 \implies (\Delta_p - \Delta_a)(\frac{1}{\beta} - (\Delta_p + \Delta_a)) \geq 0 \\
&\implies (\Delta_p - \Delta_a) - \beta(\Delta_p - \Delta_a)(\Delta_p + \Delta_a) \geq 0 \\
&\implies \Delta_p - \beta\Delta_p^2 - \Delta_a + \beta\Delta_a^2 \geq 0 \\
&\implies \Delta_p(1 - \beta\Delta_p) \geq \Delta_a(1 - \beta\Delta_a) \\
&\implies \frac{\Delta_p(1 - \beta\max\{\Delta_p, \Delta_a\})}{\Delta_a(1 - \beta\min\{\Delta_p, \Delta_a\})} \geq 1 (\because \Delta_p \geq \Delta_a)
\end{aligned}
$$

□

D PROOF OF THEOREM 4

Theorem 4. *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 if:*

$$\Delta_p \leq \Delta_a \text{ and } \Delta_p + \Delta_a \leq \frac{1}{\beta} \tag{7}$$

Proof. To prove this thoerem, we show that the condition of Eq. 7 satisfies the condition of Thm.2 when $\rho(b) = b$. Note that $\rho'_{max} = \rho'_{min} = 1$.

$$
\begin{aligned}
&Eq.7 \implies (\Delta_p - \Delta_a)(\frac{1}{\beta} - (\Delta_p + \Delta_a)) \leq 0 && (64)\\
&\implies (\Delta_p - \Delta_a) - \beta(\Delta_p - \Delta_a)(\Delta_p + \Delta_a) \leq 0 && (65)\\
&\implies (\Delta_p - \Delta_a) - \beta(\Delta_p^2 - \Delta_a^2) \leq 0 && (66)\\
&\implies \Delta_p - \beta\Delta_p^2 - \Delta_a + \beta\Delta_a^2 \leq 0 && (67)\\
&\implies \Delta_p(1 - \beta\Delta_p) \geq \Delta_a(1 - \beta\Delta_a) && (68)\\
&\implies \Delta_p \leq \frac{\Delta_a(1 - \beta\Delta_a)}{(1 - \beta\Delta_p)} && (69)\\
&\implies \frac{\Delta_p(1 - \beta\min\{\Delta_p, \Delta_a\})}{\Delta_a(1 - \beta\max\{\Delta_p, \Delta_a\})} \leq 1 (\because \Delta_p \leq \Delta_a) && (70)
\end{aligned}
$$

□

E VALUE BOUNDEDNESS THEOREM

Definition 3 (Value Boundedness)**.** *For a given belief state MDP, with a value function $V_\beta(b)$, states $b \in \mathcal{B}$ and some index state $z \in \mathcal{B}$, an MDP is value bounded if for a constant U_0 and function $L(b)$:*

$$L(b) < V_\beta(b) - V_\beta(z) < U_0 \tag{71}$$

We use Dutta's Theorem [6] to prove that Thm. 1 and Thm. 2 hold respectively under the average reward criterion as $\beta \to 1$.

To prove that the conditions of these theorems hold under the average reward criterion as $\beta \to 1$, we need to prove that any Collapsing Bandit is value bounded.

Theorem 7. *Any Collapsing Bandit is value bounded.*

Proof. Set the index state to be the head of the $\omega = 1$ chain, i.e., $z = P^a_{1,1}$. Since $P^a_{1,1}$ is the maximum possible belief, $V_\beta(P^a_{1,1})$ is the largest possible value function according to Corollary 5. Therefore we can set $U_0 = 0$. Now according to Lemmas 3 and 13, we have:

$$V_{m,\beta}(P^a_{1,1}) - V_{m,\beta}(b) \leq \frac{P^a_{1,1} - b}{1 - \beta\gamma} \leq \frac{P^a_{1,1} - b}{1 - \gamma} \forall \beta \in [0, 1) \tag{72}$$

$$V_{m,\beta}(b) - V_{m,\beta}(P^a_{1,1}) \geq \frac{b - P^a_{1,1}}{1 - \beta\alpha} \geq \frac{b - P^a_{1,1}}{1} \forall \beta \in [0, 1) \tag{73}$$

Thus $L(b) = \frac{b - P^a_{1,1}}{1-\gamma}$, where $\gamma = \max\{P^a_{1,1} - P^a_{0,1}, P^p_{1,1} - P^p_{0,1}\}$, thus completing the proof. □

F PROOF OF THEOREM 5

Theorem 5 (Forward Threshold Optimality). *Consider a belief-state MDP corresponding to an arm in an RMAB with some non-decreasing reward function given by $\rho(b)$, transition matrix given by P and an observation function, $\Theta(b)$ for a belief state b. For any subsidy m, there is a* forward *threshold policy that is optimal if:*

$$\frac{\Delta_p(1 - \beta\max\{\Delta_p, (\Delta_a.\Delta_e)\})}{\Delta_a(1 - \beta\min\{\Delta_p, (\Delta_a.\Delta_e)\})} \geq \frac{\rho'_{max}}{\rho'_{min}} \tag{11}$$

where $\Delta_e = \Theta'(b)$ for a linear $\Theta(b)$ such as in the example above.

Proof. We start with re-deriving the difference bound lemmas for the imprecise observations case. Recall that the value function for the active and passive actions is now given by:

$$V_m(b) = \max \begin{cases} m + \rho(b) + \beta V_m(\tau(b)) \text{...passive} \\ \rho(b) + \beta\big(\sum_\omega \Theta_\omega(b).V_m(P^a_\omega)\big) \text{...active} \end{cases} \tag{10}$$

Lemma 7.

$$\sum_{\omega=1}^{\|\Omega\|-1} \Theta'_\omega(b)\Big(V_m(P^a_\omega) - V_m(P^a_0)\Big) = \sum_\omega \Big(\Theta'_\omega(b)V_m(P^a_\omega)\Big) \tag{74}$$

Proof.

$$
\begin{aligned}
R.H.S. &= \sum_\omega \Big(\Theta'_\omega(b)V_m(P^a_\omega)\Big) \\
&= \sum_{\omega=1}^{\|\Omega\|-1} \Big(\Theta'_\omega(b)V_m(P^a_\omega)\Big) + \Theta'_0(b)V_m(P^a_0) \\
&= \sum_{\omega=1}^{\|\Omega\|-1} \Big(\Theta'_\omega(b)V_m(P^a_\omega)\Big) + \\
&\quad \Big(1 - \sum_{\omega=1}^{\|\Omega\|-1} \Theta_\omega(b)\Big)' V_m(P^a_0) \\
&= \sum_{\omega=1}^{\|\Omega\|-1} \Big(\Theta'_\omega(b)V_m(P^a_\omega)\Big) + \\
&\quad \Big(- \sum_{\omega=1}^{\|\Omega\|-1} \Theta'_\omega(b)\Big) V_m(P^a_0) \\
&= \sum_{\omega=1}^{\|\Omega\|-1} \Theta'_\omega(b)\Big(V_m(P^a_\omega) - V_m(P^a_0)\Big) \\
&= L.H.S.
\end{aligned}
$$

□

Lemma 8. *$V_m(b_1) - V_m(b_2) \geq \rho'_{min}(b_1 - b_2) \forall\ b_1, b_2$ s.t. $b_1 \geq b_2$*

Proof. The proof follows the same procedure as the precise observations case. We get:

Base case $V_m^1(b_1) - V_m^1(b_2) =$

$$m + \rho(b_1) - (m + \rho(b_2)) = \rho(b_1) - \rho(b_2) \quad \text{passive} \tag{75}$$

$$\rho(b_1) - \rho(b_2) = \rho(b_1) - \rho(b_2) \quad \text{active} \tag{76}$$

is clearly satisfied. Now assume $V_m^t(b_1) - V_m^t(b_2)) \geq \rho'_{min}(b_1 - b_2)$. Then $V_m^{t+1}(b_1) - V_m^{t+1}(b_2)$
Case 1 (both passive):

$$\begin{aligned}
&= m + \rho(b_1) + \beta V_m^t(\tau(b_1)) - (m + \rho(b_2) + \beta V_m^t(\tau(b_2))) \\
&= \rho(b_1) - \rho(b_2) + \beta\Big(V_m^t(\tau(b_1)) - V_m^t(\tau(b_2))\Big) \\
&\geq \rho(b_1) - \rho(b_2) + \beta\rho'_{min}(\tau(b_1) - \tau(b_2)) \\
&\geq \rho(b_1) - \rho(b_2)
\end{aligned} \tag{77}$$

Case 2 (both active):

$$\begin{aligned}
&= \rho(b_1) - \rho(b_2) + \beta\Big((\Theta(b_1) - \Theta(b_2))V_m^t(P_{1,1}^a) + \\
&(\Theta(b_2) - \Theta(b_1))V_m^t(P_{0,1}^a)\Big) \\
&= \rho(b_1) - \rho(b_2) + \beta\Big((\Theta(b_1) - \Theta(b_2))(V_m^t(P_{1,1}^a) - V_m^t(P_{0,1}^a))\Big) \\
&\geq \rho'_{min}(b_1 - b_2) + \beta\Big((\Theta(b_1) - \Theta(b_2))(V_m^t(P_{1,1}^a) - V_m^t(P_{0,1}^a))\Big) \\
&\geq \rho'_{min}(b_1 - b_2)
\end{aligned} \tag{78}$$

□

Lemma 9. *If $\forall\ b_1, b_2$ s.t. $b_1 \geq b_2$, $\exists\ \kappa$ such that $V_m(b_1) - V_m(b_2) \geq \kappa\rho'_{min}(b_1 - b_2)$, then, for $\alpha = \min\{\Delta_a, \sum_\omega \Delta_{e\omega}\Delta_{p\omega}\}$:*

$$V_m(b_1) - V_m(b_2) \geq \rho'_{min}(1 + \beta\alpha\kappa)(b_1 - b_2) \tag{79}$$

Proof. Using Eq. 10, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}\frac{d(\tau(b))}{db} & \text{passive} \\ \rho'(b) + \beta(\sum_\omega \Theta_\omega(b) V_m(P_\omega^a)) & \text{active} \end{cases} \tag{80}$$

Case 1 (passive):

$$= \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}(P_{1,1}^p - P_{0,1}^p) \tag{81}$$

$$= \rho'(b) + \beta \lim_{\delta \to 0} \frac{V_m(\tau(b) + \delta) - V_m(\tau(b))}{\tau(b) + \delta - \tau(b)}(P_{1,1}^p - P_{0,1}^p) \tag{82}$$

$$\geq \rho'(b) + \beta\kappa\rho'_{min}(P_{1,1}^p - P_{0,1}^p) \tag{83}$$

$$\geq \rho'_{min} + \beta\kappa\rho'_{min}(P_{1,1}^p - P_{0,1}^p) \tag{84}$$

$$\geq \rho'_{min}(1 + \beta\alpha\kappa) \tag{85}$$

Case 2 (active):

$$= \rho'(b) + \beta(\sum_{\omega=1}^{\|\Omega\|-1} \Theta'_\omega(b)\Big(V_m(P_\omega^a) - V_m(P_0^a)\Big)) \tag{86}$$

$$\geq \rho'(b) + \beta\kappa\rho'_{min}\Big(\sum_{\omega=1}^{\|\Omega\|-1} \Theta'_\omega(b)(P_\omega^a - P_0^a)\Big) \tag{87}$$

$$\geq \rho'_{min} + \beta\kappa\rho'_{min} \sum_{\omega=1}^{\|\Omega\|-1} (\Delta_{e\omega}\Delta_{p\omega}) \tag{88}$$

$$\geq \rho'_{min}(1 + \beta\alpha\kappa) \tag{89}$$

Thus,

$$\begin{aligned}
&\implies \frac{d(V_m(b))}{db} \geq \rho'_{min}(1 + \beta\alpha\kappa) \\
&\implies \int_{b_2}^{b_1} \frac{d(V_m(b))}{db} db \geq \int_{b_2}^{b_1} \rho'_{min}\Big(1 + \beta\alpha\kappa\Big) db \\
&\implies V_m(b_1) - V_m(b_2) \geq \rho'_{min}\big(1 + \beta\alpha\kappa\big)(b_1 - b_2)
\end{aligned} \tag{90}$$

□

Lemma 10. *$V_m(b_1) - V_m(b_2) \geq \frac{\rho'_{min}(b_1 - b_2)}{1 - \beta\alpha} \forall\ b_1, b_2$ s.t. $b_1 \geq b_2$*

Proof. This proof is exactly same as the proof for Lemma. 3. □

Corollary 5. *$\frac{d(V_m(b))}{db} \geq \frac{\rho'_{min}}{1 - \beta\alpha}$*

Proof. This follows from Lemma 10 by setting $b_1 = b + \delta, b_2 = b$ under the limit $\delta \to 0$. □

Lemma 11. *$V_m(b_1) - V_m(b_2) \leq \frac{\rho'_{max}(b_1 - b_2)}{1 - \beta} \forall\ b_1, b_2$ s.t. $b_1 \geq b_2$*

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

$$m + \rho(b_1) - (m + \rho(b_2)) = \rho(b_1) - \rho(b_2) \leq \frac{\rho'_{max}(b_1 - b_2)}{1 - \beta}$$

both passive

$$\rho(b_1) - \rho(b_2) \leq \frac{\rho'_{max}(b_1 - b_2)}{1 - \beta} \text{ both active}$$

which are both clearly satisfied. Now assume $V_m^t(b_1) - V_m^t(b_2) \leq \frac{\rho'_{max}(b_1 - b_2)}{1 - \beta}$. Then, $V_m^{t+1}(b_1) - V_m^{t+1}(b_2)$
Case 1 (both passive):

$$\begin{aligned}
&= \big(m + \rho(b_1) + \beta V_m^t(\tau(b_1))\big) - \big(m + \rho(b_2) + \beta V_m^t(\tau(b_2))\big) \\
&= (\rho(b_1) - \rho(b_2)) + \beta\big(V_m^t(\tau(b_1)) - V_m^t(\tau(b_2))\big) \\
&\leq (\rho(b_1) - \rho(b_2)) + \beta\left(\frac{\rho'_{max}(\tau(b_1) - \tau(b_2))}{1 - \beta}\right) \\
&\leq \rho'_{max}(b_1 - b_2) + \beta\left(\frac{\rho'_{max}(b_1 - b_2)}{1 - \beta}\right) \text{ by Fact 3} \\
&= \frac{\rho'_{max}(b_1 - b_2)}{1 - \beta}
\end{aligned} \tag{91}$$

Case 2 (both active):

$$
\begin{aligned}
&= \Big(\rho(b_1) + \beta\big(\Theta(b_1)V_m^t(P_{1,1}^a) + (1-\Theta(b_1))V_m^t(P_{0,1}^a)\big)\Big) - \\
&\Big(\rho(b_2) + \beta\big(\Theta(b_2)V_m^t(P_{1,1}^a) + (1-\Theta(b_2))V_m^t(P_{0,1}^a)\big)\Big) \\
&= (\rho(b_1) - \rho(b_2)) + \beta\Big(\big(\Theta(b_1) - \Theta(b_2)\big)\big(V_m^t(P_{1,1}^a) - V_m^t(P_{0,1}^a)\big)\Big) \\
&\leq (\rho(b_1) - \rho(b_2)) + \beta\left((\Theta(b_1) - \Theta(b_2)).\frac{\rho'_{max}(P_{1,1}^a - P_{0,1}^a)}{1-\beta}\right) \\
&\leq \rho'_{max}(b_1 - b_2) + \beta\left(\frac{\rho'_{max}(\Theta(b_1) - \Theta(b_2))}{1-\beta}\right) \\
&\leq \rho'_{max}(b_1 - b_2) + \beta\left(\frac{\rho'_{max}(b_1 - b_2)}{1-\beta}\right) \\
&= \frac{\rho'_{max}(b_1 - b_2)}{1-\beta}
\end{aligned} \tag{92}
$$

□

Lemma 12. *If* $\forall\, b_1, b_2$ *s.t.* $b_1 \geq b_2$, $\exists\, \kappa$ *such that* $V_m(b_1) - V_m(b_2) \leq \kappa\rho'_{max}(b_1 - b_2)$, *then, for* $\gamma = \max\{\Delta_p, \sum_\omega \Delta_{a\omega}\Delta_{e\omega}\}$:

$$V_m(b_1) - V_m(b_2) \leq \rho'_{max}(1 + \beta\gamma\kappa)(b_1 - b_2) \tag{93}$$

Proof. Using Equation 10, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}\frac{d(\tau(b))}{db} & \text{passive} \\ \rho'(b) + \beta(V_m(P_{1,1}^a) - V_m(P_{0,1}^a)) & \text{active} \end{cases} \tag{94}$$

Case 1 (passive):

$$
\begin{aligned}
&= \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}(P_{1,1}^p - P_{0,1}^p) && (95)\\
&= \rho'(b) + \beta\lim_{\delta\to 0}\frac{V_m(\tau(b)+\delta) - V_m(\tau(b))}{\tau(b)+\delta-\tau(b)}(P_{1,1}^p - P_{0,1}^p) && (96)\\
&\leq \rho'(b) + \beta\kappa\rho'_{max}(P_{1,1}^p - P_{0,1}^p) && (97)\\
&\leq \rho'_{max}(1 + \beta\gamma\kappa) && (98)
\end{aligned}
$$

Case 2 (active):

$$
\begin{aligned}
&= \rho'(b) + \beta(\sum_\omega \Theta'_\omega V_m(P_\omega^a)) && (99)\\
&= \rho'(b) + \beta(\sum_\omega \Theta'_\omega (V_m(P_\omega^a) - V_m(P_0^a))) && (100)\\
&\leq \rho'(b) + \beta\kappa\rho'_{max}(\sum_\omega \Theta'_\omega (P_\omega^a - P_0^a)) && (101)\\
&\leq \rho'_{max} + \beta\rho'_{max}\gamma\kappa && (102)\\
&\leq \rho'_{max}(1 + \beta\gamma\kappa) && (103)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\implies \frac{d(V_m(b))}{db} \leq \rho'_{max}(1 + \beta\gamma\kappa) \\
&\implies \int_{b_2}^{b_1}\frac{d(V_m(b))}{db}db \leq \int_{b_2}^{b_1}\rho'_{max}\Big(1 + \beta\gamma\kappa\Big)db && (104)\\
&\implies V_m(b_1) - V_m(b_2) \leq \rho'_{max}\big(1 + \beta\gamma\kappa\big)(b_1 - b_2)
\end{aligned}
$$

□

Lemma 13. $V_m(b_1) - V_m(b_2) \leq \frac{\rho'_{max}(b_1 - b_2)}{1-\beta\gamma} \forall\, b_1, b_2$ *s.t.* $b_1 \geq b_2$

Proof. The proof is same as the proof of Lemma 13. □

Corollary 6. $\frac{d(V_m(b))}{db} \leq \frac{\rho'_{max}}{1-\beta\gamma}$

Proof. This follows from Lemma 13 by setting $b_1 = b+\delta, b_2 = b$ under the limit $\delta \to 0$. □

Now we complete the proof for Thm.5 as follows:

$$
\begin{aligned}
&Eq.11 \implies \Delta_p \geq \frac{\Delta_a(1-\beta\alpha)\rho'_{max}}{(1-\beta\gamma)\rho'_{min}} && (105)\\
&\implies \frac{(P_{1,1}^p - P_{0,1}^p)\rho'_{min}}{(1-\beta\alpha)} \geq \frac{\rho'_{max}(P_{1,1}^a - P_{0,1}^a)}{(1-\beta\gamma)} && (106)\\
&\implies \frac{(P_{1,1}^p - P_{0,1}^p)\rho'_{min}}{(1-\beta\alpha)} \geq V_m(P_1^a) - V_m(P_0^a) \text{ by Lemma 13} && (107)\\
&\implies (P_{1,1}^p - P_{0,1}^p)\frac{d(V_m(b))}{db} \geq V_m(P_1^a) - V_m(P_0^a) \text{ by Cor. 5} && (108)\\
&\implies \frac{d(\tau(b))}{db}\frac{d(V_m(b))}{db} \geq V_m(P_1^a) - V_m(P_0^a) \text{ by Fact 1} && (109)\\
&\implies \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau b)}\frac{d(\tau(b))}{db} \geq \rho'(b) + && (110)\\
&\quad \beta\Big(V_m(P_1^a) - V_m(P_0^a)\Big) && (111)\\
&\implies \frac{d(V_m(b|a=0))}{d(b)} \geq \frac{d(V_m(b|a=1))}{d(b)} && (112)
\end{aligned}
$$

□

G PROOF OF THEOREM 6

Theorem 6 (Reverse Threshold Optimality). *Consider a belief-state MDP corresponding to an arm in an RMAB with some non-decreasing reward function given by* $\rho(b)$, *transition matrix given by* P *and an observation function,* $\Theta(b)$ *for a belief state* b. *For any subsidy* m, *there is a* reverse *threshold policy that is optimal if:*

$$\frac{\Delta_p(1 - \beta\min\{\Delta_p, (\Delta_a.\Delta_e)\})}{\Delta_a(1 - \beta\max\{\Delta_p, (\Delta_a.\Delta_e)\})} \leq \frac{\rho'_{min}}{\rho'_{max}} \tag{12}$$

where $\Delta_e = \Theta'(b)$ *for a linear* $\Theta(b)$ *such as in the example above.*

Proof. This proof follows along the same lines as Proof of Thm. 2 using the value function bounds for imprecise observations. □