HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Harvard John A. Paulson School of Engineering and Applied Sciences
have examined a dissertation entitled:

"Actualizing Impact of AI in Public Health: Optimization of Scarce Health Intervention Resources in the Real World"

presented by: Aditya Shrikant Mate

Signature Milind Tambe
Typed name: Professor M. Tambe

Signature Susan Murphy
Typed name: Professor S. Murphy

Signature Lucas Janson
Typed name: Professor L. Janson

May 23, 2023

Actualizing Impact of AI in Public Health: Optimization of Scarce Health Intervention Resources in the Real World

A DISSERTATION PRESENTED
BY
ADITYA S. MATE
TO
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2023

Thesis advisor: Professor Milind Tambe Aditya S. Mate

Actualizing Impact of AI in Public Health: Optimization of Scarce Health Intervention Resources in the Real World

Abstract

While AI is assuming omnipresence today more than ever, its adoption is still limited in solving challenges pertaining to socially-critical problem domains such as in public health, especially among low-resource and underserved communities. Motivated by the desire to solve impactful, real-world problems that involve reasoning, strategic decision-making, or planning in uncertain, stochastic or resource-limited settings, my thesis presents novel solutions designed for two such real-world public health challenges: tuberculosis prevention and improving maternal and child healthcare.

Building AI systems for realizing social impact in public health, demands solving a number of fundamental research questions. For instance, community health workers and NGOs operating with limited health resources face the challenge of optimally utilizing these resources to maximize their impact. In doing so, such NGOs must account for domain-specific considerations such as fairness or risk-averseness and plan the limited resources to serve beneficiaries at scale, in an uncertain and dynamically changing world. Even with new solution techniques for allocating limited resources in such socially critical domains now being built, their accurate evaluation through Randomized Controlled Trials (RCTs) remains difficult due to high sample variance in these settings.

Towards tackling these challenges, my thesis utilizes techniques such as Restless Multi-Armed Bandits (RMAB) to solve the sequential decision-making problem of allocating scarce health intervention resources. My thesis builds computationally efficient solution algorithms to this problem, that can be adopted by non-profits without needing access to heavy computing power. Next, I also propose techniques that allow the planner to accommodate real-world considerations such as risk-averseness or fairness in planning health interventions. Furthermore, my thesis also builds solutions that can plan such health interventions while accounting for dynamically changing patient cohorts and the finite stay of patients in such health programs. Transcending the boundaries of traditional research, I have transitioned this work from the blackboard to a first-of-its-kind field evaluation of the RMAB algorithm, involving 23,000 real-world mothers over a 7-week period, results of which show a $\sim$ 30% improvement in the performance metric of interest. Finally, my work mitigates the challenges faced in the evaluation of such resource allocation algorithms through RCTs. Using techniques from causal reasoning, I present a novel concept that retrospectively reassigns participants to experimental groups in a trial. Using this concept, I build a new estimator, that I show, can sharply reduce sample variance.

Contents

Listing of figures

Acknowledgments

I sincerely thank everyone who has helped me along my Ph.D. journey (and during the build-up to the Ph.D. journey). First and foremost I want to thank my advisor, Milind Tambe, for taking me on as his student and offering his excellent mentorship over the last five years! Working with him has proved deeply impactful for my growth as a researcher, professional and as a person. All the technical and non-technical skillsets, hunger for excellence and ambition I learned from you will benefit me throughout the rest of my career.

Thank you to my thesis defense committee members, Susan Murphy and Lucas Janson for your immense support and guidance as well as your kind willingness to discuss research and offer advice. Thank you also to my qualifying exam committee members Yiling Chen, Nick Menzies and Francis Doyle for useful advice and inputs to help me establish my research direction.

Thank you also to my amazing technical mentors in my lab, Andrew Perrault, Bryan Wilder and Kai Wang who I was fortunate to work with during my initial years. I benefited immensely from your patience and handholding in those early years when I was still learning how to even perform any research. I truly and deeply appreciate all your words of wisdom, thoughtful advice given during various stages of my Ph.D. and am grateful to have found incredible friends and mentors in you for life!

Thank you to all my amazing mentors and internship managers I've been fortunate to work with and learn from during my Ph.D., especially Bistra Dilkina, Kush Varshney, Philip Nelson, Aparna Taneja and Manish Jain. Thank you for being so generous to me with your advice, help and encouragement. I have learned a great deal from working with you.

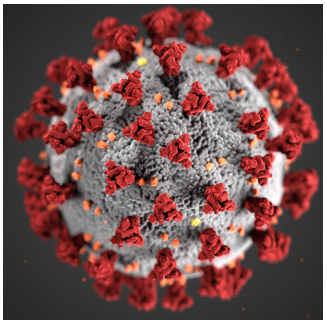
Thank you also to all my collaborators, co-authors, research partners and labmates and friends who have been a constant source of support, available to discuss research ideas or debug issues and challenges. Your intellectual company, enthusiasm and humility has been inspirational to me and provided me with an amazing work atmosphere last several years.

Finally, but most importantly, I want to deeply thank my family, to who I really owe most of my success and achievements. My Mom and Dad have been my constant source of encouragement, motivation and inspiration my entire life, pushing me out of my comfort zone and encouraging me to strive for excellence. I am deeply grateful to them for being my facilitators, being always available to lend an ear or offer sage advice and offering their steadfast support in every walk of my life. Thank you for ingraining in me the criticality of growing in humility, kindness, and generosity alongside professional growth and for teaching me to prioritize these qualities over all other academic and scholastic achievements.

0

Introduction

Public health is a grand global challenge today — with 6 out of 10 biggest causes of deaths being communicable, neonatal or nutritional diseases. This challenge is even more critical in the limited-resource settings of the global south, where half the world's population lacks access to essential health services [122]. This situation has only been made worse by the COVID-19 pandemic, further impacting the underserved populations and widening existing disparities [55]. With the belief that AI holds the potential to solve some of these most pressing problems faced by the world today, my thesis aims to unlock the power of AI to tackle such impactful real-world problems. My thesis focuses on challenges in public health, drawing on tools from AI — specifically, techniques such as Restless Multi-Armed Bandits, Probabilistic Modeling, Multi-agent Systems and Causal Reasoning to develop innovative solutions for two such motivating application domains: tuberculosis prevention and improving maternal and



(a) Picture credits: CDC

(b) Picture credits: WHO SEARO

(c) Picture credits: Pippa ranger

child healthcare.

My thesis describes work in partnership with 'ARMMAN'[11], an India-based non-profit, that endeavors to improve maternal and child health outcomes – and has served over 26 million women in India so far. My thesis also includes work in collaboration with the Government of Maharashtra and AI solutions designed for assisting TB prevention in India.

One common challenge in these public health contexts is that of monitoring and encouraging adherence to prescribed medication or ensuring engagement of beneficiaries with health information programs. For instance, Community Health Workers (CHWs) are often tasked with monitoring adherence of Tuberculosis patients to their prescribed 6-month-long treatment regimen and are expected to deliver interventions where necessary to encourage adherence. Similarly, support staff at non-profits like ARMMAN[11] running health information programs may deliver interventions to boost engagement with the goal of maximizing the positive health outcomes from the program.

Figure 1: During a field visit in Mumbai to understand the problem first-hand and to determine the most useful AI solutions.

In dealing with these challenges, my research thrust examines the entire data-to-impact pipeline and identifies several fundamental research questions consisting of three key components:

- **Optimization and Planning of Limited Resources**: Viewed algorithmically, one central question addressed in my work, is to decide how to allocate the limited health worker intervention resources opti-

mally. The objective is to maximize adherence or engagement with the program, but with limited intervention resources at hand. Thus, this intervention planning challenge entails determining a subset of k beneficiaries from the N-strong cohort to intervene on each timestep ($k \ll N$), so as to maximize the overall benefits of the interventions.

- **Deployment**: I have strived to actualize the impact of my technical work through deployment and real-world evaluations in the field. Building upon these evaluations and the lessons learned along the way, my work has culminated in a full-scale deployment of my algorithms and has served $> 100,000$ expectant and new mothers enrolled in the health program by the partner NGO so far.

- **Evaluation**: Finally, accurate evaluation and measurement of the performance of the developed methods is the key last step towards achieving measurable, social impact. I have proposed new techniques that improve the accuracy of this evaluation.

0.1 Problem Statement

To summarize the key research challenge addressed, in this thesis, I seek to answer the question: "How can we use AI to maximally utilize limited resources and design solutions that deliver measurable, real-world impact in public health?"

My key contributions towards answering this question are: (1) novel algorithms with relevant theoretical guarantees to solve the sequential decision-making challenges faced by stakeholders in public health settings, while accommodating a number of real-world considerations they may care about (2) first-of-its-kind field evaluation of these techniques in partnership with non-profits and finally (3) novel methods to improve evaluation through RCTs, of such resource allocation algorithms being increasingly developed for addressing critical issues of societal impact.

0.2 Summary of Contributions

0.2.1 Restless Multi-Armed Bandits for Optimizing Healthcare Interventions

My work is the first to cast this 'engagement monitoring and intervention planning' challenge as a Restless Multi-Armed Bandits (RMAB) problem. The RMAB framework – used typically for handling resource allocation problems [141] – models each TB patient or expectant mother as a Markov Decision Process representing one arm of the RMAB. Solving for the optimal RMAB policy is PSPACE hard in general. Existing solutions are computationally expensive needing a computing cluster to run, rendering them inaccessible to low-resource non-profits.

My work identifies 'Collapsing Bandits' (CoBs) [101] as a special RMAB subclass, with useful theoretical properties. Using these, I develop a fast algorithm exploiting the special structure of CoBs to accelerate computation. My algorithm achieves a 3-order-of-magnitude speedup while maintaining similar solution quality. This advance enables applying the RMAB techniques in the context of applications such as adherence monitoring for TB patients, without needing access to powerful computational resources.

0.2.2 Risk-Aware Bandits for Planning Public Health Interventions

While 'Collapsing Bandits' presents a useful tool for solving this problem, it fails to account for risk-sensitivity considerations of the real world. It also relies on the end beneficiaries reporting their state of adherence truthfully during each intervention.

To overcome these limitations, my thesis extends the collapsing bandits framework to allow for risk-aware planning that can account for real-world planner considerations such as risk-averseness or equitable allocation and can even counter imperfect observations characteristic of the real world. While these are useful solution techniques, the existence of a solution and the accompanying optimality guarantees hinge on the validity of a technical condition called 'indexability'. Unfortunately, there are no known results on indexability of RMABs in general. I also derive novel theoretical results on indexability for risk-aware bandits, that yield compact closed-form sufficient conditions for verifying indexability simply from the problem parameters [104].

0.2.3 'Streaming Bandits': Optimizing Interventions for Dynamically Changing Cohorts

Existing RMAB solutions (including [101] and [104]) assume all agents (such as mothers in a health program) start and end the program synchronously. However, in reality, the beneficiary cohort can be dynamic — new beneficiaries may join and existing enrolled beneficiaries may leave continually. This asynchronous and finite intermediate stay, unfortunately, prevents existing scalable techniques from working well out-of-the-box. Towards circumventing these issues in a scalable manner, I first showed that interpolating between the cheaply available solutions for the infinite- and small-horizon problems is nearly as effective as solving the finite-horizon problem exactly. Using this, I proposed 'Streaming Bandits' [100] that yields an interpolation-based algorithm that speeds up the planning process by 2-orders-of-magnitude, while preserving the performance quality of exact methods, even for dynamically changing cohorts.

0.2.4 Field Study in Deploying Restless Bandit Algorithms in Healthcare

My work [102] is the first to evaluate RMABs through a real-world field trial for any public health application. In partnership with ARMMAN [11], I ran a first-of-its-kind large-scale trial, extending over 7 weeks and involving 23,003 real mothers in India evaluating the RMAB algorithm. Real-world numbers from the trial show, with statistical significance, that my RMAB algorithm managed to cut engagement drops among mothers by $\sim$ 30% in comparison to other baselines.

0.2.5 Non-Stationary Restless Bandit Algorithms

While previous work shows the RMAB model to be useful, its practical utility is restrained by a key underlying model assumption: that each agent being catered to in the health program follows a Markov Decision Process (MDP). In reality, we find evidence from real-world data collected in the above trial suggesting that such real-world agents may not conform to the Markov behavior model.

To bridge this gap, my work casts this challenge as a non-stationary RMAB problem (RMAB-NS), that admits time-varying transition parameters, $\mathcal{P}(t)$ instead of a fixed point $\mathcal{P}$. Towards ensuring the practicability of our approach, this work presents a technique to infer the time-dependent parameters $\mathcal{P}(t)$, for real-world agents — a task exacerbated by the richer parameter space of the RMAB-NS model. The idea hinges on leveraging the small

number of unique behavior patterns displayed by agents, and pooling data by clustering agents according to these behavior patterns.

0.2.6 SAHELI

Building on the above technical advances and learnings from field testing, we build and deploy "Saheli", a system to efficiently utilize the limited availability of health workers for improving maternal and child health in India. Saheli is the *first deployed application* for RMABs in public health, and is already *in continuous use* by our partner NGO, ARMMAN. We have already reached $\sim 150K$ beneficiaries with Saheli, and are on track to serve 1 million beneficiaries by the end of 2023. This scale and impact has been achieved through multiple innovations in the RMAB model and its development, in preparation of real-world data, and in deployment practices; and through careful consideration of responsible AI practices. Specifically, this chapter describes our approach to learn from past data to improve the performance of Saheli's RMAB model, the real-world challenges faced during deployment and adoption of Saheli, and the end-to-end pipeline.

0.2.7 Improved Evaluation of Algorithmic Resource Allocation Policies

Evaluation of resource allocation policies through RCTs is significantly difficult because the individual outcomes are interlinked through the resource constraint. My work shows how this leads to significantly large variance, making the evaluation difficult. Despite the increasing use of algorithmic resource allocation, evaluation of such policies is a critical gap that no prior work has tried to mitigate. My thesis addresses this gap [106] by proposing a new estimator that reduces the sample variance.

I propose a novel concept that retrospectively reassigns participants to trial arms, thus generating additional samples of counterfactual trials. I prove theoretically that such reassignments are "allowed" and lead to unbiased estimate that is guaranteed to reduce sample variance. Empirically, I showed the statistical performance of my method to be equivalent to running anywhere between 2 *to* 13 *independent, full-sized* RCTs in parallel.

0.3 Thesis outline

The remainder of the thesis follows the outline of the summary. Chapters 1, 2, 3 pertain to the first theme (Optimization and Planning) and present novel algorithmic solutions and theoretical results to the resource allocation decision-making problem. Chapters 4, 5, 6 pertain to the second theme and present work on deploying and evaluating the RMAB algorithm in the field, in a maternal and child healthcare task. Chapter 7 (final theme) describes new methods to improve the evaluation of algorithmic resource allocation policies through RCTs. Related work and background information is provided in each chapter. Finally, Chapter 8 presents a concluding summary and directions for future research.

1

'Collapsing Bandits' for Optimizing Public Health Intervention Resources

1.1 Introduction

Motivation. This chapter 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[63,46,2,158], anti-poaching patrols[132], 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[116], hypertension[26], tuberculosis[134,28] and HIV[77,76]. 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[97,113] or Alzheimer's Disease[93].

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[125]. 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[160] and generally performs well empirically[8] 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[160]. 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.

1.2 Restless Multi-Armed Bandits

An RMAB consists of a set of N arms, each associated with a *two-action* MDP [131]. 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 [73]. 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 [163]. Because general RMABs are PSPACE-hard [125], 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 & Zhao [95] 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 & Mahajan [3] 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. [46] 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 [60] and Sombabu et al. [142] augment the machine maintenance problem from Glazebrook et al. [46] 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. [108] 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. [132] 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.

1.3 Collapsing Bandits

We introduce *Collapsing Bandits* (CoB) as a specially structured RMAB with partial observability. In CoB, each arm $n \in \{1, \ldots, N\}$ has binary latent states $\mathcal{S} = \{0, 1\}$, representing *bad* and *good* state, respectively. The agent acts during each of finite days $t \in 1, \ldots, T$. 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

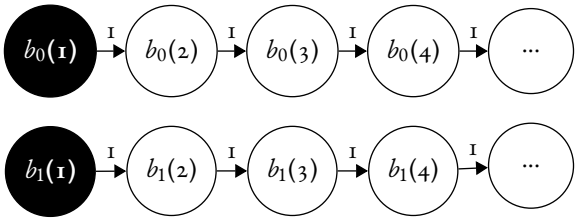


Figure 1.1: Belief-state MDP under the policy of always being passive. There is one chain for each observation $\omega \in \{0, 1\}$ with the head marked black. Belief states deterministically transition down the chains.

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. 1.2.

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. 2.2. 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. 1.2a for an illustration).

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

Formally,

$$b_\omega(u) = \tau_{u-1}(P^a_{\omega,1})\ \forall u \in [T] \text{ where } \tau_u(b) = \frac{P^p_{0,1} - (P^p_{1,1} - P^p_{0,1})^u (P^p_{0,1} - b(1 + P^p_{0,1} - P^p_{1,1}))}{(1 + P^p_{0,1} - P^p_{1,1})} \tag{1.1}$$

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 [163]. 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**):

$$V_m(b) = \max \begin{cases} m + b + \beta V_m(\tau_1(b)) & \text{passive} \\ b + \beta(b V_m(P^a_{1,1}) + (1-b) V_m(P^a_{0,1})) & \text{active} \end{cases} \tag{1.2}$$

where τ is defined in Eq. 1.1 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:

$$W(b) = \inf_m \{m : V_m(b; a=0) \geq V_m(b; a=1)\} \tag{1.3}$$

Critically, performance guarantees hold only if the problem satisfies *indexability* [160,163], 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[95]. Our proof features a sophisticated induction argument exploiting the finite size of $\mathcal{B}$ and relies on tools from real analysis for limit arguments.

□

All formal proofs can be found in the appendix. We remark that Thm. 1 generalizes the result in the seminal work by Liu & Zhao[95] who proved the indexability for a special class of CoB. In particular, the RMAB in Liu & Zhao[95] 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[163], 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 $X_\omega^{b_{th}} \in 1, \ldots, T$ is the index of the first

belief state in each chain where it is optimal to act (i.e., the belief is less than or equal to b_{tb}). We now drop the superscript b_{tb} for ease of exposition. See Fig. 1.2a 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:

$$f^{(X_0,X_1)}(b_\omega(u)) = \begin{cases} \alpha & \text{if } \omega = 0, u \leq X_0 \\ \beta & \text{if } \omega = 1, u \leq X_1 \\ 0 & \text{otherwise} \end{cases} \tag{1.4}$$

$$\alpha = \left(\frac{(X_1 b_0(X_0))}{1 - b_1(X_1)} + X_0\right)^{-1}, \beta = \left(\frac{X_1 b_0(X_0)}{1 - b_1(X_1)} + X_0\right)^{-1} \frac{b_0(X_0)}{1 - b_1(X_1)} \tag{1.5}$$

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

$$J_m^{(X_0,X_1)} = \sum_{b \in \mathcal{B}} b f^{(X_0,X_1)}(b) + m(1 - f^{(X_0,X_1)}(b_1(X_1)) - f^{(X_0,X_1)}(b_0(X_0))) \tag{1.6}$$

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 4 in Appendix A.1. 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.[132] (our main competitor) which has per-round complexity of $\approx \mathcal{O}(N \log(\frac{1}{\varepsilon})(|\mathcal{S}|T)^{2+\frac{1}{18}})$, where $\log(\frac{1}{\varepsilon})$ 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

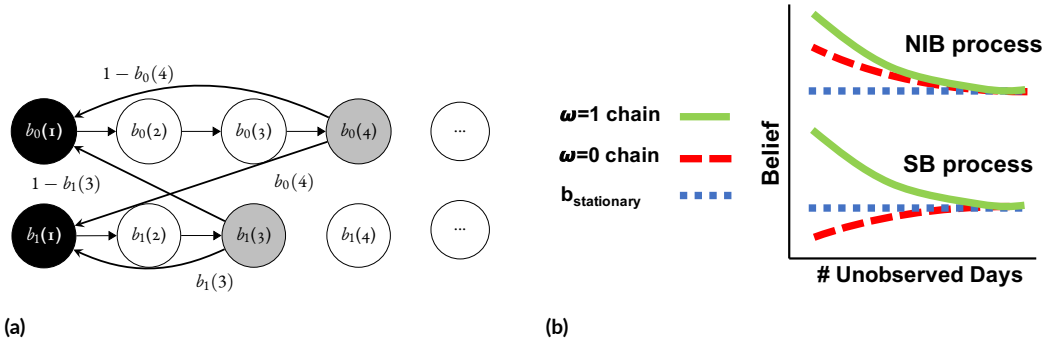


Figure 1.2: (a) Visualization of forward threshold policy ($X_0 = 4, X_1 = 3$). Black nodes are the head of each chain and grey nodes are the thresholds. (b) Non-increasing belief (NIB) process has non-increasing belief in both chains. A split belief process (SB) has non-increasing belief after being observed in state 1, but non-decreasing belief after being observed in state 0.

due to the best-known complexity of solving a linear program with $|\mathcal{S}|T$ variables[69].

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 A.6 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. 1.2b. 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. 1.2 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[145], and the latter is linear in b. Thus, as shown in Fig. 1.3, 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.

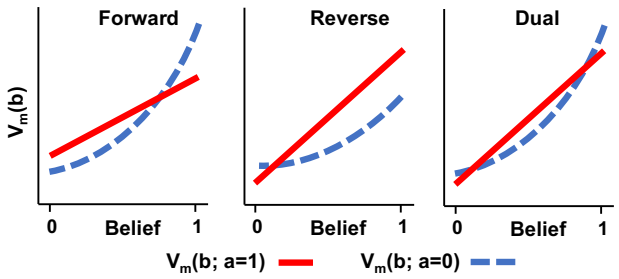


Figure 1.3: Components of $V_m(b)$ in Eq. 1.2. Since the passive action is convex in b, active action is linear in b, and value function is a max over these, at most three optimal policy types are possible.

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:*

$$(P^p_{1,1} - P^p_{0,1})(1 + \beta(P^a_{1,1} - P^a_{0,1}))(1 - \beta) \geq P^a_{1,1} - P^a_{0,1} \tag{1.7}$$

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 & Zhao[95] consider the case where $P^a_{1,1} = P^p_{1,1}$ and $P^a_{0,1} = P^p_{0,1}$, which

makes Eq. 1.7 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:*

$$(P^p_{1,1} - P^p_{0,1})\left(1 + \frac{\beta(P^a_{1,1} - P^a_{0,1})}{1-\beta}\right) \leq P^a_{1,1} - P^a_{0,1} \tag{1.8}$$

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. 6 and Thm. 7 also serve as conditions for the average reward case as $\beta \to 1$ (a proof based on Dutta's Theorem [36] is given in Appendix A.4).

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.

1.5 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 chapter. **Qian et al.** [132], 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. [132] 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.

1.5.1 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. [79]. 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 $P^{\phi}_{\omega,1}$ ($P^{a}_{\omega,1}$) 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. 1.4a, we plot the runtime in seconds vs the number of patients N. Fig. 1.4b 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. 1.4c 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 $P^{\phi}_{\omega,1}$ and $P^{a}_{\omega,1}$ from the data, and present the extensive sensitivity analysis in Appendix A.7. Finally, in Appendix A.6 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.

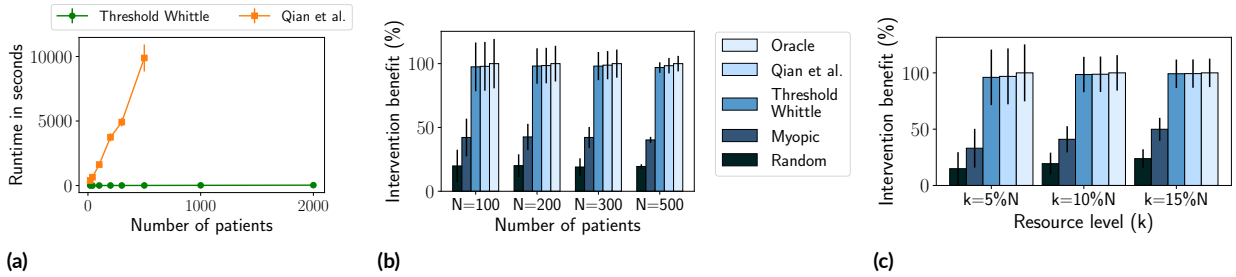


Figure 1.4: (a) Threshold Whittle is several orders of magnitude faster than Qian et al. and scales to thousands of patients without sacrificing performance on realistic data (b). (c) Intervention benefit of Threshold Whittle is far larger than naive baselines and nearly as large as Oracle.

1.5.2 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. 1.3.

Fig. 1.5a 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 A.5. We set the resource level, $k = 10\%N$ in our simulation for Fig. 1.5a. 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. 1.5b 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. 1.5c 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. 1.5d 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 A.8.

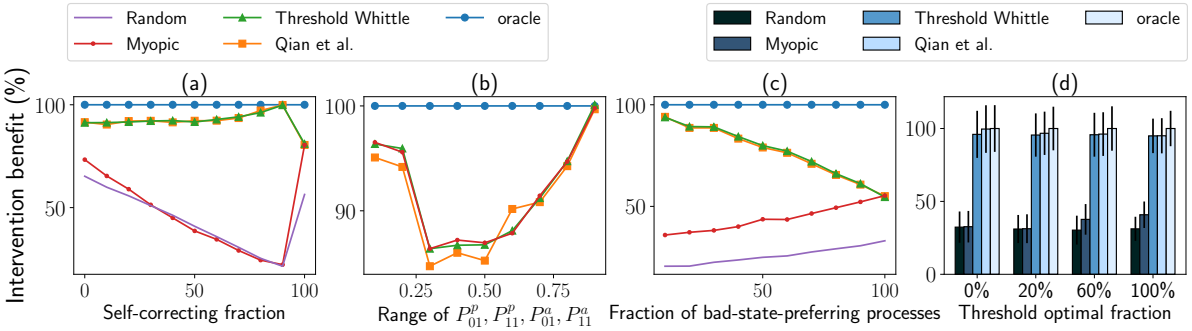


Figure 1.5: (a) Myopic can be trapped into performing even worse than Random while Threshold Whittle remains close to optimal. (b) Long-term planning is least effective when entropy of states is maximum. (c) Myopic and Whittle planning become similar when more processes are prone to failures. (d) Threshold Whittle is surprisingly robust to processes even outside of theoretically guaranteed conditions.

1.6 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 [32,38], (non-)communicable diseases [32,140], and sexual/reproductive health [165] in low-resource communities across the world [35,38,140,162]. Our modeling has the potential to improve the delivery of care in these highly resource-constrained settings.

Figure 1.6: 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.

2

Risk-aware Bandits for Risk-sensitive Healthcare Intervention Planning

2.1 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 [155]. 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 [167,165]. The effectiveness of CHWs in achieving de-

Figure 2.1: Community Health Worker delivering an intervention. Image source: Pippa Ranger

sirable community health outcomes through the interventions they deliver has been recognized in the context of several domains such as achieving child survival goals [52], improving child and maternal health [162,119], communicable and non-communicable diseases [32,140], sexual and reproductive health [121], etc.

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 [129,91,30] have largely focused on building assistants that send personalized health reminders or recommendations to patients. However, these assume resource-rich environ-

ments 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[95,132,101] 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 chapter, 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 chapter 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.

2.2 Background

2.2.1 Restless Multi-Armed Bandits.

An RMAB consists of N independent arms, each consisting of an associated 2-action Markov Decision Process (MDP)[131]. 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, $\pi : \mathcal{S} \rightarrow \mathcal{A}$ 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 $R_\beta^\pi(s_0) = E\left[\sum_{t=0}^{\infty} \beta^t r(s_t) | \pi, s_0\right]$, 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.

2.2.2 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 [125]. However, Whittle proposed a heuristic [163], known today as the Whittle Index, that was later been shown to be asymptotically optimal for the time average reward problem [160], and also for other more general families of RMABs arising from stochastic scheduling problems [46].

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: $r_m : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, 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.

2.2.3 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. [46] 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. [60] and [142] augment the machine maintenance problem by introducing either i.i.d. or Markovian stochasticity in the reset action, and [152] study Whittle Index for general functions of states assuming a single, fixed, reset state. [108] explore Hidden Markov Bandits which consider partial observability with binary state transitions, but don't accommodate state dependent rewards from passive arms.

Liu & Zhao [95] is a seminal work that builds off of the well-established 2-state Elliot-Gilbert channel model [44] 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. [3] 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 chapter. [132] 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. [101] 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

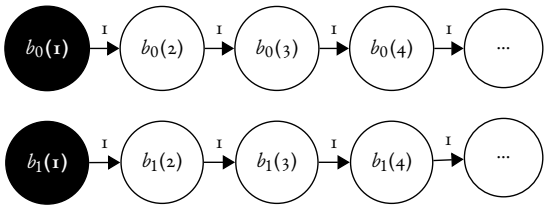


Figure 2.2: Belief states are arranged in two chains, one corresponding to each observation. Belief state deterministically transitions to the next belief state in the chain when passive. $b_0(1)$ and $b_1(1)$ (shown in black) are the reset states.[101].

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[78,171,20]. [81] and [23] are other contemporaneous works that focus on RMAB planning with multiple available actions or model-free approaches to learning in RMABs.

2.3 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), $\mathcal{N} = \{1, 2, \ldots N\}$, 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: $b \to bP^p_{11} + (1-b)P^p_{01}$. When an active action is taken, the patient emits an observation ω from the observation set $\Omega = \{0, 1, \ldots |\Omega| - 1\}$ 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 2.5, 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[95,101] 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[101] 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 objec-

tive. In this work, we allow the planner to have an arbitrary objective that translates to the goal of maximizing the long term reward accrued, specified by an arbitrary, monotonically increasing $\rho(b)$.

2.4 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. 2.2. 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. 2.3. 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 [95])and is given by:

$$\tau_u(b) = \frac{P^p_{0,1} - (P^p_{1,1} - P^p_{0,1})^u (P^p_{0,1} - (1 + P^p_{0,1} - P^p_{1,1})b)}{(1 + P^p_{0,1} - P^p_{1,1})} \tag{2.1}$$

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 2.4.1. Second, computation of the index itself is challenging and can be computationally expensive. We use the theoretical results of subsection 2.4.1, to devise a fast algorithm to compute the Whittle Index efficiently, which we present in Subsection 2.4.2.

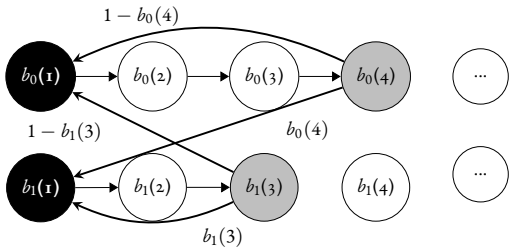


Figure 2.3: State transition diagram when a threshold policy with thresholds $u_0 = 4$ and $u_1 = 3$ is implemented. Belief stochastically resets to one of the reset states when active.

2.4.1 Indexability and Threshold Optimality

Definition 4 (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. 2.2 below and picking the action with a larger value.

$$V_m(b) = \max \begin{cases} m + \rho(b) + \beta V_m(\tau(b)) \text{...passive} \\ \rho(b) + \beta\Big(b.V_m(P^a_{1,1}) + (1-b)V_m(P^a_{0,1})\Big) \text{...active} \end{cases} \tag{2.2}$$

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. [101] 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 5 (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 [101], $\rho(b) = b$. 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 chapter. 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 4 (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:*

$$\frac{\Delta_p(1-\beta\max\{\Delta_p,\Delta_a\})}{\Delta_a(1-\beta\min\{\Delta_p,\Delta_a\})} \geq \frac{\rho'_{max}}{\rho'_{min}} \tag{2.3}$$

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 proof is in Appendix B.1 of the chapter. □

Theorem 5 (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:*

$$\frac{\Delta_p(1-\beta\min\{\Delta_p,\Delta_a\})}{\Delta_a(1-\beta\max\{\Delta_p,\Delta_a\})} \leq \frac{\rho'_{min}}{\rho'_{max}} \tag{2.4}$$

Proof Sketch. The proof follows similar reasoning as Thm.4. 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.2 of the chapter. □

2.4.2 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. 2.3) 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. 2.3. 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:

$$M_{bb'} = \begin{cases} 1 & \text{if } b' = \tau(b) \text{ and } b' \geq b_\omega(u_\omega) \text{ for } \omega \in \{0,1\} \\ b & \text{if } b' = b_1(1) \text{ and } b = b_\omega(u_\omega) \text{ for } \omega \in \{0,1\} \\ 1-b & \text{if } b' = b_0(1) \text{ and } b = b_\omega(u_\omega) \text{ for } \omega \in \{0,1\} \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

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 $J_{m,\rho}^{(u_0,u_1)} = \sum_{b\in\mathcal{B}} f^{(u_0,u_1)}(b)\big(\rho(b) + m.1_{\{b>b_{th}\}}\big)$.

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 $b+\varepsilon$ (where $\varepsilon \to 0$) 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 $J_{m,\rho}^{\overline{U}(b)} = J_{m,\rho}^{\overline{U}(b+\varepsilon)}$. Note that changing the threshold to $b+\varepsilon$ affects the threshold belief state only on the current chain. We use this idea to construct the fast Whittle Index computation algorithm (Alg.1).

Algorithm 2: Risk-sensitive Index Computation Algorithm

1: Initialize pointers to heads of chains, $u_0 = 1, u_1 = 1$.
2: **while** $u_0 < T$ or $u_1 < T$ **do**
3: Compute $m_1 := m$ such that $J_{m,\rho}^{(u_0,u_1)} = J_{m,\rho}^{(u_0,u_1+1)}$
4: Compute $m_0 := m$ such that $J_{m,\rho}^{(u_0,u_1)} = J_{m,\rho}^{(u_0+1,u_1)}$
5: Set $i = \arg\min\{m_0, m_1\}$ and $W(b_i(u_i)) = m_i$
6: Increment u_i
7: **end while**

2.4.3 Application to Collapsing Bandits

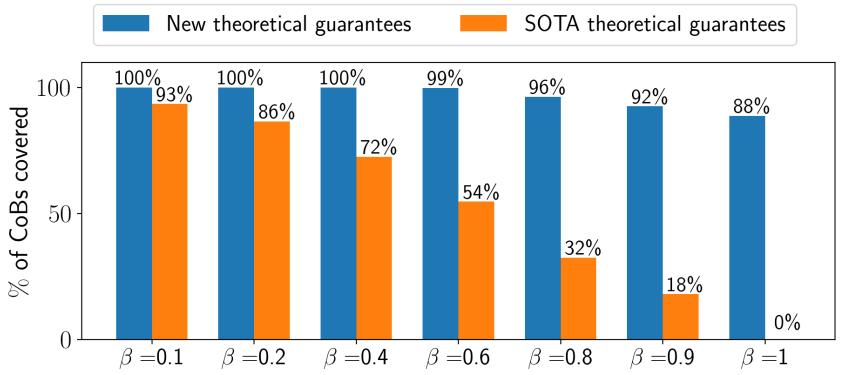


Figure 2.4: For $\rho(b) = b$, the theoretical guarantees presented in this chapter hold for a wider range of processes as compared to the state-of-the-art conditions of Mate et al. [101].

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) [101] are a sub-case of the risk-sensitive bandits considered in this chapter, with reward function $\rho(b) = b$. The conditions of Thms. 4 and 5 yield novel sufficiency conditions when $\rho(b) = b$, that are wider than those presented in Mate et al. [101]. For example, under the average reward criterion (or $\beta = 1$), as shown in Fig. 2.4, the conditions of Mate et al. [101] become vacuous, whereas the new conditions derived here guarantee indexability for 88% of the entire space of CoBs.

Theorem 6. *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:*

$$\Delta_a \leq \Delta_p \text{ and } \Delta_a + \Delta_p \leq \frac{1}{\beta} \tag{2.6}$$

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. 2.6 satisfies the condition of Thm.4 when $\rho(b) = b$. Complete details of the proof are available in Appendix B.3 of the chapter.

Theorem 7. *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:*

$$\Delta_p \leq \Delta_a \text{ and } \Delta_p + \Delta_a \leq \frac{1}{\beta} \tag{2.7}$$

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. 6 and Thm. 7 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 [36] is given in Appendix B.4 in the full version of the chapter).

Corollary 1. *Collapsing Bandits are indexable if:*

$$\Delta_p + \Delta_a \leq \frac{1}{\beta}. \tag{2.8}$$

From Thms. 6 and 7, 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. [101], 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*

$$\Delta_p + \Delta_a \leq 1. \tag{2.9}$$

Remark 1. *Corollary 2 proves that Conjecture 1 of [101] 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.* [132] *and others.*

2.5 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.

2.5.1 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

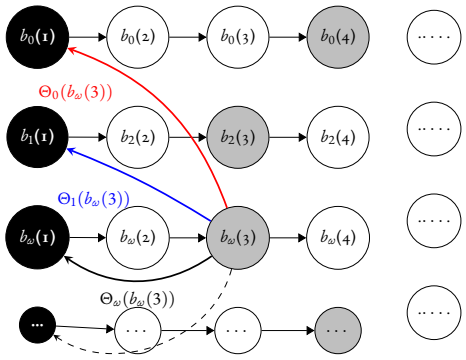


Figure 2.5: Multiple observations lead to a multiple-chain organization of belief states, with each observation having its corresponding reset state. An active action resets the belief state to $b_\omega(1)$ if observation ω is observed.

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).

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:

$$V_m(b) = \max \begin{cases} m + \rho(b) + \beta V_m(\tau(b)) \text{...passive} \\ \rho(b) + \beta\big(\sum_\omega \Theta_\omega(b).V_m(P^a_\omega)\big) \text{...active} \end{cases} \tag{2.10}$$

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

2.5.2 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 8 (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:*

$$\frac{\Delta_p(1-\beta\max\{\Delta_p,(\Delta_a.\Delta_e)\})}{\Delta_a(1-\beta\min\{\Delta_p,(\Delta_a.\Delta_e)\})} \geq \frac{\rho'_{max}}{\rho'_{min}} \tag{2.11}$$

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

Theorem 9 (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:*

$$\frac{\Delta_p(1-\beta\min\{\Delta_p,(\Delta_a.\Delta_e)\})}{\Delta_a(1-\beta\max\{\Delta_p,(\Delta_a.\Delta_e)\})} \leq \frac{\rho'_{min}}{\rho'_{max}} \tag{2.12}$$

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

2.6 Experimental Evaluation

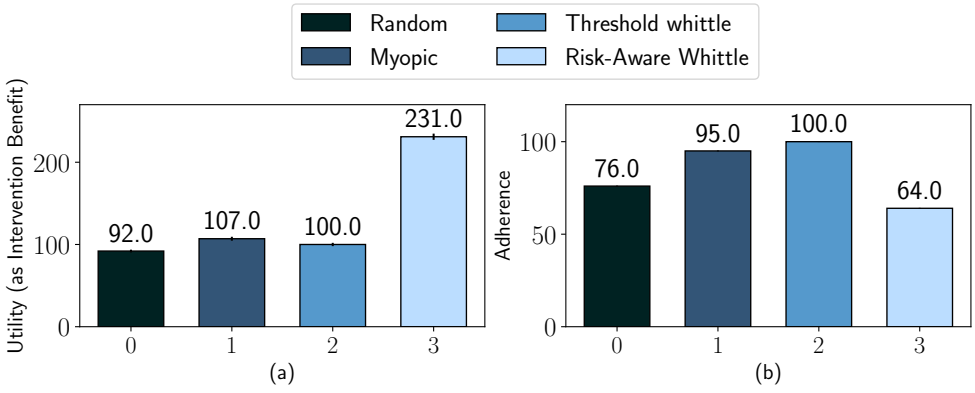
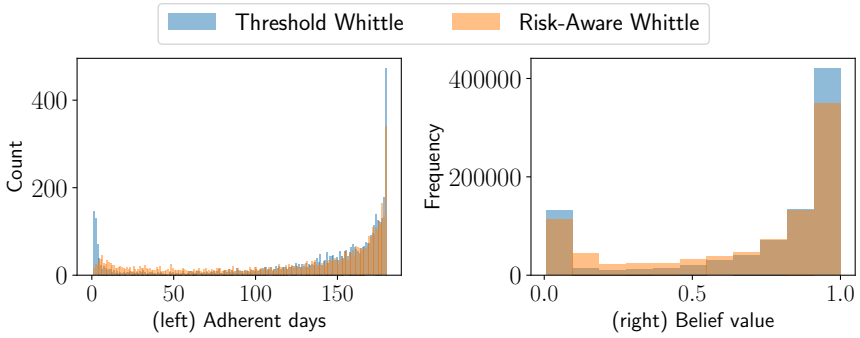
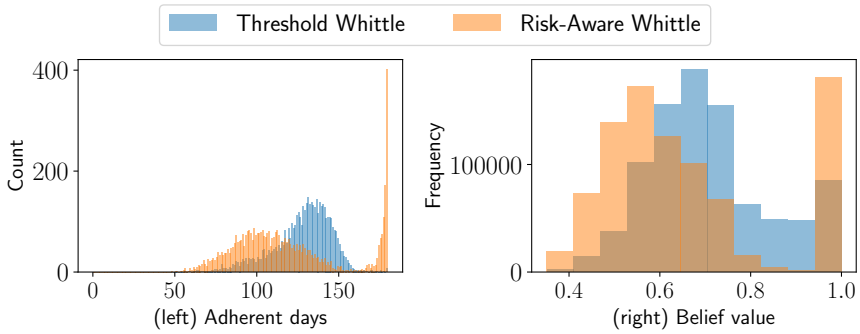


Figure 2.6: Risk-Aware Whittle optimizes for the objectives the planner cares about, and avhieves much higher utility than Threshold Whittle, even while scoring lower on average adherence—a metric that previous approaches to the HMIP focus on.



(a) Handling Fairness Concerns



(b) Handling Risk-Sensitivity

Figure 2.7: (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.

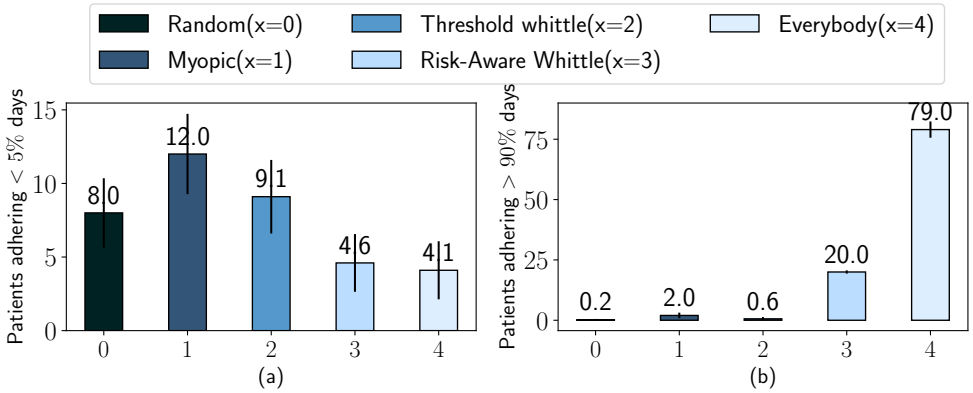


Figure 2.8: Risk-Aware Whittle is significantly better at tackling the specific concerns of the CHW. (a) shows a sharp decrease in the number of patients with a severely low adherence rate. (b) shows a significant jump in the number of patient finishing the treatment with a high adherence score.

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 objec-

tives 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[79] and run simulations following the same data imputation steps as[101] for consistency. We compare the following algorithms: **Risk-Aware Whittle** is the algorithm presented in this chapter. **Threshold Whittle** is the SOTA fast algorithm presented in[101], 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.

2.6.1 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. 2.6 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. 2.7b(right) shows the histogram of time spent by patients in a belief state over the duration of the pro-

gram. 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. 2.7b(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. 2.8(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.

2.6.2 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. 2.7a(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. 2.7a(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. 2.8(a) which plots the number of patients with 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.

2.6.3 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. 2.9 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.

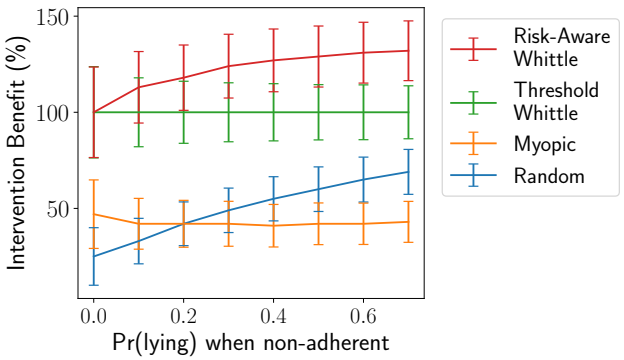


Figure 2.9: Risk-Aware Whittle beats Threshold Whittle when patients misrepresent their adherence states.

2.7 Discussion and Conclusion

Mitigating bias in socio-technical systems such as ours, is an important issue[49,58]. 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.

3

‘Streaming Bandits’: Optimizing Interventions for Dynamically Changing Cohorts

3.1 Introduction

In community healthcare settings, adherence of patients to prescribed health programs, that may involve taking regular medication or periodic health checkups, is critical to their well-being. One way to improve patients’ health outcomes is by tracking their health or monitoring their adherence to such programs. Such health monitoring programs combined with suitably designed intervention schemes help patients alleviate health issues such as diabetes [117], hypertension [27], tuberculosis [29,135], depression [98,114], etc. However, health interventions often

require dedicated time of healthcare workers, which is a severely scarce resource, grossly inadequate to meet the total demand. This issue is especially more severe in the global south. Moreover, planning interventions with these limited resources is made more challenging due to the fact that the extent of adherence of patients may be both, uncertain as well as transient. Consequently, the healthcare workers have to grapple with this sequential decision making problem of deciding which patients to intervene on, with limited resources, in an uncertain environment. Existing literature on healthcare monitoring and intervention planning (HMIP)[4,101,18,105,102] casts this as a *restless multi-armed bandit* (RMAB) planning problem. In this setup, the patients are typically represented by the arms of the bandit and the planner must decide which arms to pull (which patients to intervene on) under a limited budget. The RMAB problem formalizes the (restless) behavioral dynamics of the patients both in the presence and in the absence of interventions.

In addition to healthcare, RMABs have caught traction as solution techniques in a myriad of other domains involving limited resource planning for applications such as anti-poaching patrol planning[132], multi-channel communication systems[95], sensor monitoring tasks[45], UAV routing[85] etc. For ease of presentation, we consider the HMIP problem for motivation but our approach is relevant and can be extended to other real-world domains.

The existing literature on RMABs for intervention planning, however, has mainly focused on problems involving an infinite time horizon (i.e., the health programs are assumed to run forever) and, moreover, the results are limited to settings where no new patients (or bandit arms) arrive midway during the health program. We consider a general class of RMABs, which we call *streaming restless multi-armed bandits*, or S-RMAB. In an S-RMAB instance, the arms of the bandit are allowed to arrive asynchronously, that is, the planner observes an incoming and outgoing stream of bandit arms. The classic RMAB (both with infinite and finite horizon) is a special case of the S-RMAB where all arms appear (leave) at the same time. Additionally, each arm of an S-RMAB is allowed to have its own transition probabilities, capturing the potentially heterogeneous nature of patient cohorts. S-RMABs display a special structure in the presence of streaming arms and a finite horizon, which the existing methods fail to utilize. Our approach exploits this structure to arrive at approaches that perform better in the streaming bandit setting.

A fairly general approach, proposed by[132] may be applied even when patients arrive and leave asynchronously after staying for a finite duration. The method allows to approximate the exact solution arbitrarily well, but it

is computationally expensive as the number of patients or arms increases. A more recent approach, proposed by [101], exploits the structure of the HMIP and is considerably faster, but the method relies on the assumption of an infinite planning horizon. This algorithm suffers a severe deterioration in performance when employed on shorter horizon settings.

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

3.2 Related work

The RMAB problem was introduced by [163]. The paper studied the RMAB problem with the goal of maximizing the average reward in a dynamic programming framework. Whittle formulated a relaxation of the problem and provided a heuristic called *Whittle Index policy*. This policy is optimal when the underlying Markov Decision Processes satisfy indexability, which is computationally intensive to verify. Later, [124] established that solving RMAB is PSPACE hard, even when the transition rules are known. Since then, specific classes of RMABs have been studied extensively. [132] studied the infinite horizon RMAB problem and proposed a binary search based algorithm to find Whittle index policy. However, the algorithm becomes computationally expensive as the number of arms grows. [18] models the problem of maximizing health information coverage as an RMAB problem and proposes a hierarchical policy which leverages the structural assumptions of the RMAB model. [4] provide a solution for the class of bandits with "controlled restarts" and state-independent policies, possessing the indexability property. [101] model a health intervention problem, assuming that the uncertainty about the state collapses when an intervention is provided. They provide an algorithm called *Threshold Whittle* to compute the Whittle indices

for infinite horizon RMAB. There are many other papers that provide Whittle indexability results for different subclasses of Partially Observable Markovian bandits[45,59,143,96]. However, these papers focus on infinite horizon, whereas we focus on the more challenging setting when there is a fixed finite horizon.

The RMAB problem with finite horizon has been comparatively less studied. [118] provided solutions to the one-armed restless bandit problem, where only one arm is activated at each time before a time horizon T. Their solution do not directly extend to the scenario when multiple arms can be pulled at each time step. [61] considered finite horizon multi-armed restless bandits with identically distributed arms. They show that an index based policy based on the Lagrangian relaxation of the RMAB problem, similar to the infinite horizon setting, provides a near-optimal solution. [86] study the problem of selecting patients for early-stage cancer screening, by formulating it as a very restricted subclass of RMAB. All these works consider that all the arms are available throughout T time steps. Some other works, such as [109,53] also adopt different approaches to decomposing the bandit arms, which may be applicable to finite horizon RMABs. These techniques to solving weakly coupled Markov Decision Processes are more general, but consequently less efficient than the Whittle Index approach in settings where indexability assumption holds.

The S-RMAB problem has been studied in a more restricted setting by [173]. They assume that, at each time step, arms may randomly arrive and depart due to random abandonment. However, the main limitation of their solution is the assumption that all arms have the same state-transition dynamics. This assumption does not hold in most of the real-world instances and thus, in this chapter, we consider heterogeneous arms—arms are allowed to have their own transition dynamics. We show empirically that our algorithms perform well even with heterogeneous arms.

Another related category of work studied *sleeping arms* for the *stochastic multi-armed bandits* (SMAB) problem, where the arms are allowed to be absent at any time step [74,83,24]. However, the SMAB is different from RMAB because, in the former, when an arm is activated, a reward is drawn from a Bernoulli reward distribution (and not dependent on any state-transition process). Thus, the algorithms and analysis of SMABs do not translate to the RMAB setting.

3.3 Streaming bandits

The *streaming restless multi-armed bandit* (S-RMAB) problem is a general class of RMAB problem where a stream of arms arrive over time (both for finite and infinite-horizon problems). Similar to RMAB, at each time step, the decision maker is allowed to take active actions on at most k of the available arms. Each arm i of the S-RMAB is a Partially Observable Markov Decision Process (POMDP)—represented by a 4-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$. $\mathcal{S} = \{0, 1\}$ denotes the state space of the POMDP, representing the "bad" state (say, patient not adhering to the health program) and "good" state (patient adhering), respectively. $\mathcal{A}$ is the action space, consisting of two actions $\mathcal{A} = \{a, p\}$ where an action a (or, p), denotes the active (or, passive) action. The state $s \in \mathcal{S}$ of the arm, transitions according to a known transition function, $P^{a,i}_{s,s'}$ if the arm is pulled and according to the known function, $P^{p,i}_{s,s'}$ otherwise. We also assume the transition function to conform to two natural constraints often considered in existing literature[95,101]: (i) Interventions should positively impact the likelihood of arms being in the good state, i.e. $P^a_{01} > P^p_{01}$ and $P^a_{11} > P^p_{11}$ and (ii) Arms are more likely to remain in the good state than to switch from the bad state to the good state, i.e. $P^a_{11} > P^a_{01}$ and $P^p_{11} > P^p_{01}$. Though the transitions probabilities are known to the planner, the actual state change is stochastic and is only partially observable—that is, when an arm is pulled, the planner discovers the true state of the arm; however, when the arm is not pulled, uncertainty about the true state persists. Under such uncertainties, it is customary to analyze the POMDP using its equivalent belief state MDP representation instead[101]. The state space of this MDP is defined by a set of all possible "belief" values that the arm can attain, denoted by $\mathcal{B}_i$. Each belief state $b \in \mathcal{B}_i$ represents the likelihood of the arm being in state 1 (good state). This likelihood is completely determined by the number of days passed since that arm was last pulled and the last observed state of the arm[95]. At each time step t, the planner accrues a state-dependent reward r_t from an active arm i, defined as:

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

The total reward* of $R_t = \sum_{i \in [N]} (r_t(i))$ is accrued by the planner at time t, which is the sum of individual rewards obtained from the available arms. The planner's goal is to maximize her total reward collected, $\bar{R} :=$

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

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

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

$$\bar{R} := \sum_{t\in[T]} \sum_{i\in[N(t)]} r_t(i). \tag{3.1}$$

3.4 Methodology

The dominant paradigm for solving RMAB problems is the Whittle index approach. The central idea of the Whittle approach is to decouple the RMAB arms and then compute indices for each arm that capture the "value" of pulling that arm. The Whittle Index policy then proceeds by pulling the k arms with the largest values of Whittle Index. This greedy approach makes the time complexity linear in the number of arms, as indices can be computed independently for each arm. The computation of the index hinges on the notion of a "passive subsidy" m, which is the amount rewarded to the planner for each arm kept passive, in addition to the usual reward

collected from the arm. The Whittle Index for an arm is defined as the infimum value of subsidy, m that must be offered to the planner, so that the planner is indifferent between pulling and not pulling the arm. To formalize this notion, consider an arm of the bandit in a belief state b. Its active and passive value functions, under a discount factor of β, and when operating under a passive subsidy m, can be written as:

$$V^{p}_{m,T}(b) = b + m + \beta V_{m,T-1}(bP^{p}_{11} + (1-b)P^{p}_{01}) \tag{3.2}$$

$$V^{a}_{m,T}(b) = b + \beta b V_{m,T-1}(P^{a}_{11}) + \beta(1-b)V_{m,T-1}(P^{a}_{01}) \tag{3.3}$$

The value function for the belief state b is $V_{m,T}(b) = \max\{V^{p}_{m,T}(b),$

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

In this section we first show theoretically that the Streaming Bandit setup is indexable (subsection 3.4.1). Next, in subsection 3.4.2, we observe and formalize a useful phenomenon about the Whittle Index in the finite horizon setting. We use this phenomenon to design fast algorithms for S-RMABs in subsection 3.4.3 and we provide runtime complexity analysis for the same in subsection 3.4.4. Finally in subsection 3.4.5 we identify cases beyond those identified by previous work to which our efficient algorithm extends.

3.4.1 Conditions for indexability of streaming bandits

In this section, we extend the conditions for indexability that [101] originally established for infinite horizon, to the finite horizon setting of Streaming bandits. To show indexability, we first show in Theorem 10, that S-RMABs can be reduced to a standard RMAB with augmented belief states. We build on this result and prove another useful Lemma, both of which combined can be used to show that indexability holds for this augmented RMAB instance, and ultimately for S-RMABs (Theorem 20).

Definition 6 (Threshold Optimality [101])**.** *An RMAB instance is called threshold optimal if either a forward*

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

First, we show that the belief state MDP of a Streaming Bandit arm with deterministic arrival and departure time can be formulated as an augmented belief state MDP of the same instance with infinite horizon. Using this, we prove that, whenever the infinite horizon problem satisfies threshold optimality for a passive subsidy m, then the augmented belief state MDP for finite horizon also satisfies threshold optimality. Using the result that indexability holds whenever threshold optimality is satisfied [101], we imply that the Streaming Bandits problem is indexable whenever threshold optimality on the underlying infinite horizon problem is satisfied.

Theorem 10. *The belief state transition model for a 2-state Streaming Bandit arm with deterministic arrival time T_1 and departure time T_2 can be reduced to a belief state model for the standard restless bandit arm with $T_2 + (T_2 - T_1)^2$ states.*

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

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

$\langle b_w(t'), t-1\rangle$ to the state $\langle b_w(t''), t+1\rangle$ is same as the probability of changing from belief value $b_w(t')$ to $b_w(t'')$ in one time step on taking action w.

- For time step $t \geq T_2$, we create one sink state $\langle U, T_2\rangle$. This state represents unavailability of the arm subsequent to time step $T_2 - 1$. For any $t' \in \{0, 1, \ldots, T_2 - T_1\}$, the probability of transitioning from $\langle b_w(t'), T_2\rangle$ to $\langle U, T_2\rangle$ is 1.

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

$$\begin{aligned} & T_1 - 1 + 2(1 + \ldots + (T_2 - T_1)) + 1 \qquad (3.4)\\ = \;& T_1 + (T_2 - T_1)(T_2 - T_1 + 1)\\ = \;& T_2 + (T_2 - T_1)^2 \end{aligned}$$

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

Lemma 1. *If a forward (or reverse) threshold policy π is optimal for a subsidy m for the belief states MDP of the infinite horizon problem, then π is also optimal for the augmented belief state MDP.*

Proof. First, we define the value function for the modified belief states.

$$V_m^p(\langle b, t\rangle) = \begin{cases} b + m + \beta V_m(\langle bP_{11}^p + (1-b)P_{01}^p, t+1\rangle) & \text{if } b \neq U \\ b + m + V_m(\langle b', t+1\rangle) & \text{otherwise} \end{cases}$$

$$V_m^a(\langle b, t\rangle) = \begin{cases} b + \beta(V_m(\langle bP_{11}^a, t+1\rangle) + (1-b)V_m(\langle P_{01}^a, t+1\rangle)) & \\ & \text{if } b \neq U \\ b + V_m(\langle b', t+1\rangle) & \text{otherwise} \end{cases}$$

where b' is the next belief state.

The minimum value of m_U that makes the passive action as valuable as active action at the states $\langle U, t\rangle$ for

$T_1 \leq t < T_2$, can be obtained by equating

$$V^{p}_{m_U}(\langle U, t\rangle) = V^{a}_{m_U}(\langle U, t\rangle) \tag{3.5}$$

$$\Rightarrow U + m_U + V_{m_U}(\langle b', t+1\rangle) = U + V_{m_U}(\langle b', t+1\rangle) \tag{3.6}$$

$$\Rightarrow m_U = 0. \tag{3.7}$$

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

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

□

Theorem 11. *A Streaming Bandits instance is indexable when there exists an optimal policy, for each arm and every value of $m \in \mathbb{R}$, that is forward (or reverse) threshold optimal policy.*

Proof. Using Theorem 1 and Lemma 1, it is straightforward to see that an optimal threshold policy for infinite horizon problem can be translated to a threshold policy for Streaming bandits instance. Moreover, using the fact that the existence of threshold policies for each subsidy m and each arm $i \in N$ is sufficient for indexability to hold (Theorem 1 of [101]), we show that the Streaming bandit problem is also indexable. □

3.4.2 Index decay for finite horizons

In this section we describe a phenomenon called *index decay* which is observed considering short horizon. Here, the Whittle index values are low when the residual lifetime of an arm is 0 or 1. We formalize this observation in Theorem 21. We use this phenomenon as an anchor to develop our algorithm (detailed in 3.4.3). We proceed by stating one fact and proving one useful Lemma, building up towards the Theorem.

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

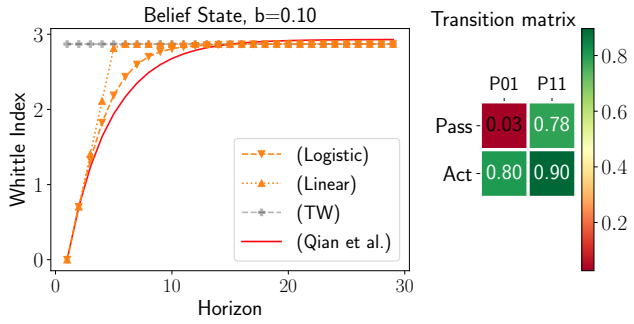


Figure 3.1: Whittle Indices for a belief state as computed by different algorithms. Both our algorithms capture index decay providing good estimates.

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

Theorem 12 (Index Decay). *Let $V^p_{m,T}(b)$ and $V^a_{m,T}(b)$ be the T-step passive and active value functions for a belief state b with passive subsidy m. Let m_T be the value of subsidy m, that satisfies the equation $V^p_{m,T}(b) = V^a_{m,T}(b)$ (i.e. m_T is the Whittle Index for a residual life time of T). Assuming indexability holds, we show that: $\forall T > 1$: $m_T > m_1 > m_0 = 0$.*

Proof. We provide our argument for a more general reward criterion than the total reward introduced in Section 3.3. Consider a discounted reward criterion with discount factor $\beta \in [0,1]$ (where $\beta = 1$ corresponds to total reward). m_0 is simply the m that satisfies: $V^p_{m,0}(b) = V^a_{m,0}(b)$ i.e., $b + m = b$, thus $m_0 = 0$. Similarly, m_1 can be solved by equating $V^p_{m_1,1}(b)$ and $V^a_{m_1,1}(b)$ and obtained as: $m_1 = \beta\Delta b = \beta\Big(\big(b\,P^a_{11} + (1-b)\,P^a_{01}\big) - \big(b\,P^p_{11} + (1-b)\,P^p_{01}\big)\Big)$

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

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

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

$$V^p_{m_1,T}(b_0) = b_0 + m_1 + \beta\rho^p(b_0, 1) + \sum_{t=2}^{T} \beta^t \rho^p(b_0, t) \tag{3.8}$$

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

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

$$< b_0 + \beta\rho^a(b_0, 1) + \sum_{t=2}^{T} \beta^t \rho^a(b_0, t) + \Big(\sum_{t=1}^{T} \beta^t m_1 . 1_{\{\pi^p(t)=passive\}}\Big) \tag{3.9}$$

(by Lemma 2)

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

□

3.4.3 Proposed algorithms

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

Recall that we establish in Theorem 21 that the Whittle Index for arms with a zero residual lifetime, is always

zero. Similarly, indices for arms with residual lifetime of 1 are simply the myopic indices, computed as:

$$\Delta b = \left(b\, P_{11}^{a} + (1-b)\, P_{01}^{a}\right) - \left(b\, P_{11}^{p} + (1-b)\, P_{01}^{p}\right).$$

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

$$\hat{W}(b, \Delta b, \overline{W}) = \min\{b\, \Delta b, \overline{W}\}. \tag{3.10}$$

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

$$\hat{W}(b, \Delta b, \overline{W}) = \frac{C_1}{1 + e^{-C_2 b}} + C_3. \tag{3.11}$$

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

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

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

Algorithm 3: Interpolation Algorithm Template

1: Pre-compute $\overline{W}(b, P^i)\ \forall b \in \mathcal{B}_i,\ \forall\, i \in [N]$, with transition matrix P^i and set of belief states $\mathcal{B}_i$.
2: **Input:** $\bar{b}_{N\times 1} \in [0,1]^N,\ \bar{h}_{N\times 1} \in [L]^N$, containing the belief values and remaining lifetimes for the N arms.
3: Initialize $\hat{W}_{N\times 1}$ to store estimated Whittle Indices.
4: **for** each arm i in N **do**
5: Let $b := \bar{b}_i, h := \bar{h}_i$ and let P be i's transition matrix.
6: Compute the myopic index Δb as:
$\Delta b = \left(b\, P^a_{11} + (1-b)\, P^a_{01}\right) - \left(b\, P^p_{11} + (1-b)\, P^p_{01}\right)$.
7: Set $\hat{W}_i(h, \Delta b, \overline{W})$ according to one of the interpolation functions (3.10) or (3.11).
8: **end for**
9: Pull the k arms with the largest values of $\hat{W}$.

We compare the index values computed by our interpolation algorithms with the exact solution by [132]. Figure 3.1 shows an illustrative example, plotting the index values as a function of the residual lifetime and shows that the interpolated values agree well with the exact values.

Infinite horizon index: For transition matrices that satisfy the conditions for forward threshold policies to be optimal, Mate et al. [101] present an algorithm that computes $\overline{W}$ cheaply. The cornerstone of their technique is to leverage forward threshold optimality to map the passive and active actions to two different forward threshold policies, and find the value of subsidy m that makes the expected reward of the policies equal. We extend this reasoning to reverse threshold optimal arms.

3.4.4 Complexity analysis

For the complexity analysis of the algorithms, we denote by $\bar{X}$ the expected number of arms arriving each time step and $\bar{L}$ their average expected lifetimes. The expected number of arms at any point in time is then $\mathcal{O}(\bar{X}\bar{L})$ [94]. Our algorithms (both versions) require a per-period cost of $\mathcal{O}(\bar{X} * |\mathcal{B}_i| = \bar{X} * 2\bar{L})$ for the Threshold Whittle pre-computations, plus $\mathcal{O}(\bar{X})$ computations for the myopic cost, plus $\mathcal{O}(\bar{X}\bar{L} * \bar{L})$ calculations (for $\bar{X}\bar{L}$ arms, each requiring up to $\bar{L}$ additions or multiplications) and $\mathcal{O}(\bar{X}\bar{L})$ for determining the top k indices. The overall per-period complexity of our algorithm is thus $\mathcal{O}(\bar{X}\bar{L}^2)$.

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

best-known complexity of solving a linear program with $\bar{L}$ variables[69] .

3.4.5 Reverse Threshold Arms

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

All the belief states that an arm can ever visit during its lifetime L can be enumerated and organized into two chains — each chain corresponding to one of the two possible observations ($\omega \in \{0, 1\}$) last observed for that arm. These chains are shown in Figure 3.2.[101] present an algorithm to compute the index for forward threshold arms with belief states belonging to the NIB process (i.e. whenever $b > b_{stationary} = \frac{P^p_{01}}{P^p_{01}+P^p_{10}}$). The algorithm relies on mapping the active and passive actions to two different forward threshold policies (with corresponding threshold states on the two chains indexed as X_0, X_1) and equating the policies' rewards to solve for the passive subsidy m, that makes the two actions equal.

We extend this reasoning to reverse threshold arms with belief chains belonging to the $\omega = 0$ chain of the SB (split-belief) process, as shown in Figure 3.2. The belief states belonging to the increasing chain ($\omega = 0$ chain) satisfy $b < b_{stationary} = \frac{P^p_{01}}{P^p_{01}+P^p_{10}}$. We identify two different reverse threshold policies that correspond to the active and passive actions, which can be used to set up similar indifference equations. For a given belief state on the increasing chain with index in the chain X, the corresponding reverse threshold policies can be indexed by $(X_0, X_1) = (1, X)$ and $(X_0, X_1) = (1, X + 1)$ and used to solve for the whittle index using the indifference equation outlined in Algorithm 1 of[101].

3.5 Experimental evaluation

We evaluate the performance and runtime of our proposed algorithms against several baselines, using both, real as well as synthetic data distributions. Logistic and Linear are our proposed algorithms. Our main baselines are: (1) a precise, but slow algorithm by Qian et al., which accounts for the residual lifetime by solving the expensive finite-horizon POMDP on each of the N arms and finds the k best arms to pull and (2) Threshold-

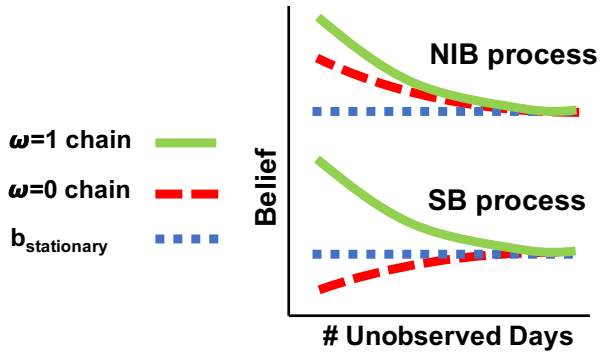
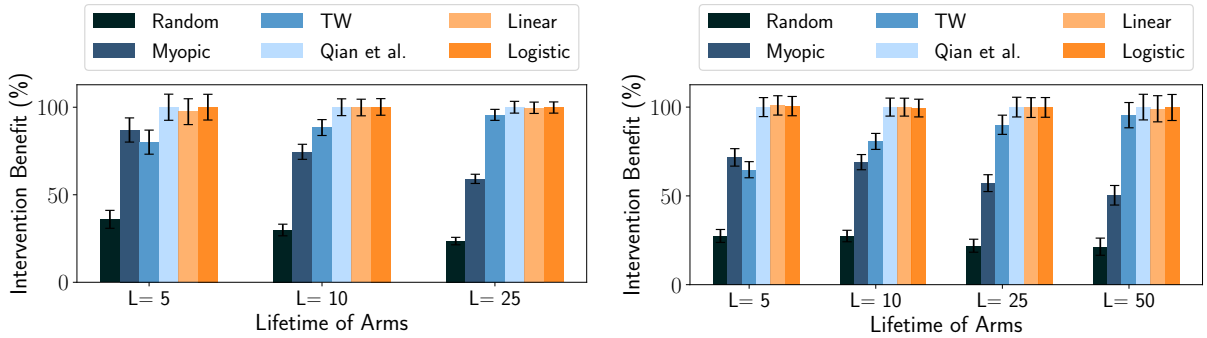


Figure 3.2: Belief values arranged in chains as presented in [101]. For every possible last observed state of the arm, ω, there is a corresponding chain of belief states.

Whittle [101] (marked as TW), a much faster algorithm, that is only designed to work for infinitely long residual time horizons. MYOPIC policy is a popularly used baseline [101,132,95] that plans interventions optimizing for the expected reward of the immediate next time step. RANDOM is a naive baseline that pulls k arms at random.

Performance is measured as the excess average intervention benefit over a 'do-nothing' policy, measuring the sum of rewards over all arms and all timesteps minus the reward of a policy that never pulls any arms. Intervention benefit is normalized to set [132] equal to 100% and can be obtained for an algorithm ALG as: $\frac{100\times(\overline{R}^{\text{ALG}}-\overline{R}^{\text{No intervention}})}{\overline{R}^{\text{Qian et al.}}-\overline{R}^{\text{No intervention}}}$ where $\overline{R}$ is the average reward. All simulation results are measured and averaged over 50 independent trials and error bars denote the standard errors.



(a) Arms arriving synchronously

(b) Arms arriving asynchronously as a stochastic process

Figure 3.3: (a) Performance of Threshold Whittle algorithm degrades when the lifetime of arms gets shorter, even when all arms start synchronously (b) The performance dwindles further if arms arrive asynchronously.

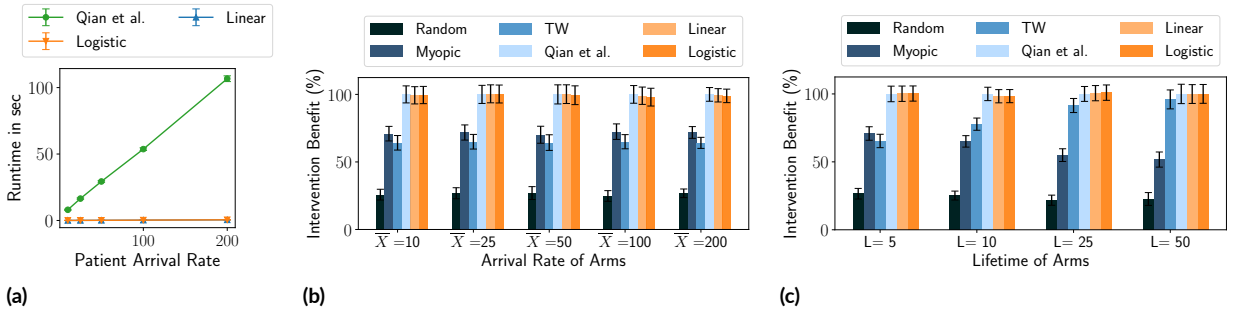


Figure 3.4: (a) Linear and Logistic interpolation algorithms are nearly 200× faster than Qian et al. (b) & (c) The interpolation algorithms achieve the speedup without sacrificing on performance, while other fast algorithms like Threshold Whittle deteriorate significantly for small residual horizons.

3.5.1 Real domain: Monitoring tuberculosis medication adherence

We first test on an anonymized real-world data set used by[79], consisting of daily adherence data of tuberculosis patients in Mumbai, India following a prescribed treatment regimen for six months. For our study, we only obtain the summary statistics capturing the transition probabilities of these patients moving between the adherent and non-adherent states as extracted from the dataset. We then follow the same data imputation steps adopted by[101] for arriving at the transition matrices, $P^a_{ss'}$ and $P^p_{ss'}$ for each patient. We sample transition matrices from this real-world patient distribution and run simulations over a simulation length much longer than the lifetimes of the patients in the simulation.

In Figure 3.3a, we first demonstrate the impact of a short horizon alone on the performance of various algorithms in a simple, non-streaming setting. In Figure 3.3b, we contrast this with a similar comparison for the short horizon setting combined with a stochastic incoming stream of patients.

In Figure 3.4, we again consider the finite horizon setting with a deterministic incoming stream of patients. In Figure 3.4a, we plot the runtimes of our algorithms and that of Qian et al., as a function of the daily arrival rate, $\bar{X}$ of the incoming stream. Figure 3.4b measures the intervention benefits of these algorithms for these values of $\bar{X}$. The lifetime of each arm, L is fixed to 5 and the number of resources, k is set to $10\% \times (\bar{X}L)$. Each simulation was run for a total length T such that $\bar{X}T = 5000$, which is the total number of arms involved in the simulation. Runtime is measured as the time required to simulate L days. The runtime of Qian et al. quickly far exceeds that of our algorithms. For the $\bar{X} = 200$ case, a single trial of Qian et al. takes 106.69 seconds to run on an average, while the proposed Linear and Logistic interpolation algorithms take 0.47 and 0.49 seconds respectively, while

attaining virtually identical intervention benefit. Other competing fast algorithms like Threshold Whittle, which assume an infinite residual horizon, suffer a severe degradation in performance for such short residual horizons. Our algorithms thus manage to achieve a dramatic speed up over existing algorithms, without sacrificing on performance.

In Figure 3.4c, we consider an S-RMAB setting, in which arms continuously arrive according to a deterministic schedule, and leave after staying on for a lifetime of L, which we vary on the x-axis. The details about the other parameters are deferred to the appendix. We also study the isolated effects of small lifetimes and asynchronous arrivals separately as well as performance in settings with stochastic arrivals, in the appendix. Across the board, we find that the performance of TW degrades as the lifetime becomes shorter and that this effect only exacerbates with asynchronous arrivals. The performance of our algorithms remains on par with Qian et al., in all of the above.

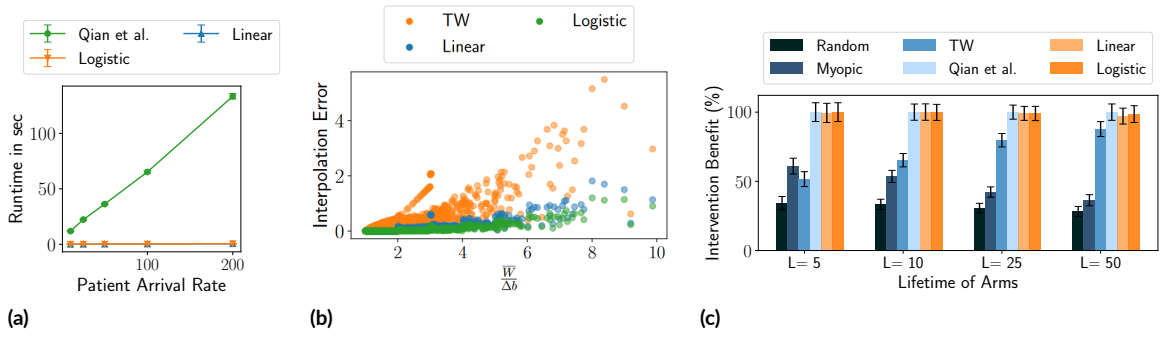


Figure 3.5: (a) The interpolation algorithms achieve a speedup of about $250\times$ over baselines.(b) The error between the actual and estimated indices is largest for TW and lower for our interpolation algorithms (c) The good performance is maintained even for reverse threshold optimal arms.

3.5.2 Real domain: ARMMAN for improving maternal healthcare

Considering an alternate real-world domain, we again only use summary statistics (transition probabilities) from an application domain consisting of intervention planning for improving maternal healthcare[21]. Individuals (arms) are labeled to be in one of three states at any time step, of which one is the good state. [102] cast the problem as an RMAB with 2-state MDP on each arm. We also focus on maximizing the number of individuals in the good state, merging the other two states from the data into a single bad state. The data set consists of three types

of transition matrices for different groups, only one of which satisfies the constraints mentioned in Section 3.3 and is used in our subsequent analysis, which is otherwise analogous to Section 3.5.1. Figure 3.5a establishes similar large runtime gains achieved by our algorithm as against other baselines, while maintaining similar performance figures in this domain. In the supplementary material we also present more details and analyses of the performance of our algorithms and baselines for this domain.

3.5.3 Synthetic domains

Finally, in this section, we test our algorithms on synthetic domains. We identify corner cases where our solutions do poorly and construct adversarial domains based on those. The ratio between the infinite horizon Whittle Index $\bar{W}$ and the myopic index Δb is an important driver of the approximation quality of our algorithms. The linear interpolation takes $\frac{\bar{W}}{\Delta b}$ steps to reach the finite horizon value, hence the higher this ratio is, higher the potential for approximation errors. In figure 3.5b we sum the approximation error over this interval $\varepsilon := \sum_{b=1}^{b=\frac{\bar{W}}{\Delta b}} (\| \hat{W}(b) - W_{Qian}(b) \|)$ and plot it for different ratios $\frac{\bar{W}}{\Delta b}$. As expected, the approximation error increases with $\frac{\bar{W}}{\Delta b}$. We construct an adversarial domain by simulating cohorts with varying proportions of such patients. The results in the supplementary material show the intervention benefit of our algorithms decreases but remains within one standard error of Qian et al.

In Figure 3.5c, we simulate a population consisting of reverse threshold optimal patients exclusively and show similar good performance even though the previous theoretical guarantees of Threshold Whittle apply to forward threshold optimal patients only. In the supplementary material, we test multiple synthetic domains by varying the proportion of forward threshold optimal patients. In addition, we perform several other robustness checks varying important problem parameters and find that the run time and strong performance of our algorithms remains consistent across the board.

3.6 Conclusion

We study *streaming bandits*, or S-RMAB, a class of bandits where heterogeneous arms arrive and leave asynchronously under possibly random streams. While efficient RMAB algorithms for computing Whittle Indices for infinite horizon settings exist, for the finite horizon settings however, these algorithms are either compara-

tively costly or not suitable for estimating the Whittle Indices accurately. To tackle this, we provide a new scalable approach that allows for efficient computation of the Whittle Index values for finite horizon restless bandits while also adapting to more general S-RMAB settings. Our approach leverages a phenomenon called *index decay* to compute the indices for each arm. Through an extensive set of experiments on real-world and synthetic data, we demonstrate that our approach provides good estimates of Whittle Indices, and yield over 200$\times$ runtime improvements without loss in performance.

4

Field Study in Deploying Restless Bandit Algorithms for Healthcare

4.1 Introduction

The wide-spread availability of cell phones has allowed non-profits to deliver targeted health information via voice or text messages to beneficiaries in underserved communities, often with significant demonstrated benefits to those communities [126,75]. We focus in particular on non-profits that target improving maternal and infant health in low-resource communities in the global south. These non-profits deliver ante- and post-natal care information via voice and text to prevent adverse health outcomes [70,12,56].

Unfortunately, such information delivery programs are often faced with a key shortcoming: a large fraction of beneficiaries who enroll may drop out or reduce engagement with the information program. Yet non-profits often have limited health-worker time available on a periodic (weekly) basis to help prevent engagement drops. More specifically, there is limited availability of health-worker time where they can place crucial service calls (phone calls) to a limited number of beneficiaries, to encourage beneficiaries' participation, address complaints and thus prevent engagement drops.

Optimizing limited health worker resources to prevent engagement drops requires that we prioritize beneficiaries who would benefit most from service calls on a periodic (e.g., weekly) basis. We model this resource optimization problem using Restless Multi-Armed Bandits (RMABs), with each beneficiary modeled as an RMAB arm. RMABs have been well studied for allocation of limited resources motivated by a myriad of application domains including preventive interventions for healthcare [101], planning anti-poaching patrols [133], machine repair and sensor maintenance [47] and communication systems [142]. However, RMABs have rarely seen real-world deployment, and to the best of our knowledge, have never been deployed in the context of large-scale public health applications.

This chapter presents first results of an RMAB system in real-world public health settings. Based on available health worker time, RMABs choose m out of N total beneficiaries on a periodic (e.g., weekly) basis for service calls, where the m are chosen to optimize prevention of engagement drops. The chapter presents two main contributions. First, previous work often assumes RMAB parameters as either known or easily learned over long periods of deployment. We show that both assumptions do not hold in our real-world contexts; instead, we present clustering of offline historical data as a novel approach to infer unknown RMAB parameters.

Our second contribution is a real-world evaluation showing the benefit of our RMAB system, conducted in partnership with ARMMAN*, an NGO in India focused on maternal and child care. ARMMAN conducts a large-scale health information program, with concrete evidence of health benefits, which has so far served over a million mothers. As part of this program, an automated voice message is delivered to an expecting or new mother (beneficiary) over her cell phone on a weekly basis throughout pregnancy and for a year post birth in a language and time slot of her preference.

Unfortunately, ARMMAN's information delivery program also suffers from engagement drops. Therefore,

*https://armman.org/

in collaboration with ARMMAN we conducted a service quality improvement study to maximize the effectiveness of their service calls to ensure beneficiaries do not drop off from the program or stop listening to weekly voice messages. More specifically, the current standard of care in ARMMAN's program is that any beneficiary may initiate a service call by placing a so-called "missed call". This beneficiary-initiated service call is intended to help address beneficiaries' complaints and requests, thus encouraging engagement. However, given the overall decreasing engagement numbers in the current setup, key questions for our study are to investigate an approach for effectively conducting additional ARMMAN-initiated service calls (these are limited in number) to reduce engagement drops. To that end, our service quality improvement study comprised of 23,003 real-world beneficiaries spanning 7 weeks. Beneficiaries were divided into 3 groups, each adding to the current standard of care. The first group exercised ARMMAN's current standard of care (CSOC) without additional ARMMAN-initiated calls. In the second, the RMAB group, ARMMAN staff added to the CSOC by initiating service calls to 225 beneficiaries on average per week chosen by RMAB. The third was the Round-Robin group, where the exact same number of beneficiaries as the RMAB group were called every week based on a systematic sequential basis.

Results from our study demonstrate that RMAB provides statistically significant improvement over CSOC and round-robin groups. This improvement is also practically significant — the RMAB group achieves a $\sim$ 30% reduction in engagement drops over the other groups. Moreover, the round-robin group does not achieve statistically significant improvement over the CSOC group, i.e., RMAB's optimization of service calls is crucial. To the best of our knowledge, this is the first large-scale empirical validation of use of RMABs in a public health context. Based on these results, the RMAB system is currently being transitioned to ARMMAN to optimize service calls to their ever-growing set of beneficiaries. Additionally, this methodology can be useful in assisting engagement in many other awareness or adherence programs, e.g., Thirumurthy & Lester [149], Chen et al. [31]. *Our RMAB code would be released upon acceptance*.

4.2 Related Work

Patient adherence monitoring in healthcare has been shown to be an important problem [99], and is closely related to the churn prediction problem, studied extensively in the context of industries like telecom [34], finance [170,139], etc. The healthcare domain has seen several studies on patient adherence for diseases like HIV [154], cardiac prob-

lems [144,33], Tuberculosis [82,127], etc. These studies use a combination of patient background information and past adherence data, and build machine learning models to predict future adherence to prescribed medication [†]. However, such models treat adherence monitoring as a single-shot problem and are unable to appropriately handle the sequential resource allocation problem at hand. Additionally, the pool of beneficiaries flagged as high risk can itself be large, and the model can't be used to prioritize calls on a periodic basis, as required in our settings.

Campaign optimization (via phone outreach) has also been studied previously. Most existing works [87,37] however, rely on the availability of a customer social network based on preferences, behavior or demographics, to help identify the set of key customers who will increase the reach of the campaign. In our domains of interest, there is no evidence of a social network among the beneficiaries, so such campaign optimization techniques are inapplicable. Furthermore, campaign optimization relies on single-shot interventions for optimization, whereas, our problem requires tracking progress of beneficiaries over multiple timesteps.

The Restless Multi-Armed Bandit (RMAB) framework has been popularly adopted to tackle such sequential resource allocation problems [164,71]. Computing the optimal solution for RMAB problems is shown to be PSPACE-hard. Whittle proposed an index-based heuristic [164], that can be solved in polynomial time and is now the dominant technique used for solving RMABs. It has been shown to be asymptotically optimal for the time average reward problem [160], and other families of RMABs arising from stochastic scheduling problems [47]. Several works as listed in Section 4.1, show applicability of RMABs in different domains but these unrealistically assume perfect knowledge of the RMAB parameters, and have not been tested in real-world contexts. Biswas et al. [22], Avrachenkov & Borkar [15], present a Whittle Index-based Q-learning approach for unknown RMAB parameters. However, their techniques either assume identical arms or rely on receiving thousands of samples from each arm, which is unrealistic in our setting, given limited overall stay of a beneficiary in an information program — a beneficiary may drop out or stop engaging with the program few weeks post enrolment unless a service call convinces them to do otherwise. Instead, we present a novel approach that applies clustering to the available historical data to infer model parameters.

Clustering in the context of Multi-Armed Bandit and Contextual Bandits has received significant attention in the past [43,90,172,88], but these settings do not consider restless bandit problems. [111] tackles a non-stationary

[†]Similarly, in our previous preliminary study (anonymous 2020) published in a non-archival setting, we used demographic and message features to build models for predicting beneficiaries likely to drop-off from ARMMAN's information program.

setup with stochastic rewards, while [16] infers model parameters from independent studies in absence of historic data. In contrast, we focus on learning RMAB parameters using clustered historic beneficiary data. [176,92] propose building predictive models per beneficiary in an online fashion, which is infeasible in our setup given the short stay of the beneficiaries.

4.3 Preliminaries

4.3.1 Background: Restless Multi-Armed Bandits

An RMAB instance consists of N independent 2-action Markov Decision Processes (MDP) [130], where each MDP is defined by the tuple $\{\mathcal{S}, \mathcal{A}, R, \mathcal{P}\}$. $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the set of possible actions, R is the reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ and $\mathcal{P}$ represents the transition function. We use $P^{a}_{s,s'}$ to denote the probability of transitioning from state s to state s' under the action a. The policy π, is a mapping $\pi : \mathcal{S} \to \mathcal{A}$ that selects the action to be taken at a given state. The total reward accrued can be measured using either the discounted or average reward criteria to sum up the immediate rewards accrued by the MDP at each time step. Our formulation is amenable to both, although we use the discounted reward criterion in our study.

The expected *discounted reward* starting from state s_0 is defined as $V^{\pi}_{\beta}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t R(s_t, \pi(s_t), s_{t+1} | \pi, s_0)\right]$ where the next state is drawn according to $s_{t+1} \sim P^{\pi(s_t)}_{s_t,s_{t+1}}$, $\beta \in [0, 1)$ is the discount factor and actions are selected according to the policy mapping π. The planner's goal is to maximize the total reward.

We model the engagement behavior of each beneficiary by an MDP corresponding to an arm of the RMAB. Pulling an arm corresponds to an active action, i.e., making a service call (denoted by $a = a$), while $a = p$ denotes the passive action of abstaining from a call. The state space $\mathcal{S}$ consists of binary valued states, s, that account for the recent engagement behavior of the beneficiary; $s \in [NE, E]$ (or equivalently, $s \in [0, 1]$) where E and NE denote the 'Engaging' and 'Not Engaging' states respectively. For example, in our domain, ARMMAN considers that if a beneficiary stays on the automated voice message for more than 30 seconds (average message length is 1 minute), then the beneficiary has engaged. If a beneficiary engages at least once with the automated voice messages sent during a week, they are assigned the engaging (E) state for that time step and non-engaging (NE) state otherwise. For each action $a \in \mathcal{A}$, the beneficiary states follow a Markov chain represented by the 2-state Gilbert-Elliot model [44] with transition parameters given by $P^{a}_{ss'}$, as shown in Figure 5.1. With slight abuse of notation, the

reward function $R(.)$ of n^{th} MDP is simply given by $R_n(s) = s$ for $s \in \{0, 1\}$.

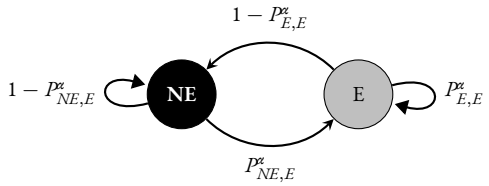


Figure 4.1: The beneficiary transitions from a current state s to a next state s' under action α, with probability $P^{\alpha}_{ss'}$.

We adopt the Whittle solution approach described previously for solving the RMAB. It hinges around the key idea of a "passive subsidy", which is a hypothetical reward offered to the planner, in addition to the original reward function for choosing the passive action. The Whittle Index is then defined as the infimum subsidy that makes the planner indifferent between the 'active' and the 'passive' actions, i.e.,:

$$W(s) = inf_{\lambda}\{\lambda : Q_{\lambda}(s, p) = Q_{\lambda}(s, a)\} \tag{4.1}$$

4.3.2 Data Collected by ARMMAN

Beneficiaries enroll into ARMMAN's information program with the help of health workers, who collect the beneficiary's demographic data such as age, education level, income bracket, phone owner in the family, gestation age, number of children, preferred language and preferred slots for the automated voice messages during enrolment. These features are referred to as Beneficiary Registration Features in rest of the chapter. Beneficiaries provided both written and digital consent for receiving automated voice messages and service calls. ARMMAN also stores listenership information regarding the automated voice messages together with the registration data in an anonymized fashion.

4.4 Problem Statement

We assume the planner has access to an offline historical data set of beneficiaries, $\mathcal{D}_{train}$. Each beneficiary data point $\mathcal{D}_{train}[i]$ consists of a tuple, $\langle f, \mathcal{E} \rangle$, where f is beneficiary i's feature vector of static features, and $\mathcal{E}$ is an episode storing the trajectory of (s, α, s') pairs for that beneficiary, where s denotes the start state, α denotes the action taken (passive v/s active), and s denotes the next state that the beneficiary lands in after executing α in state

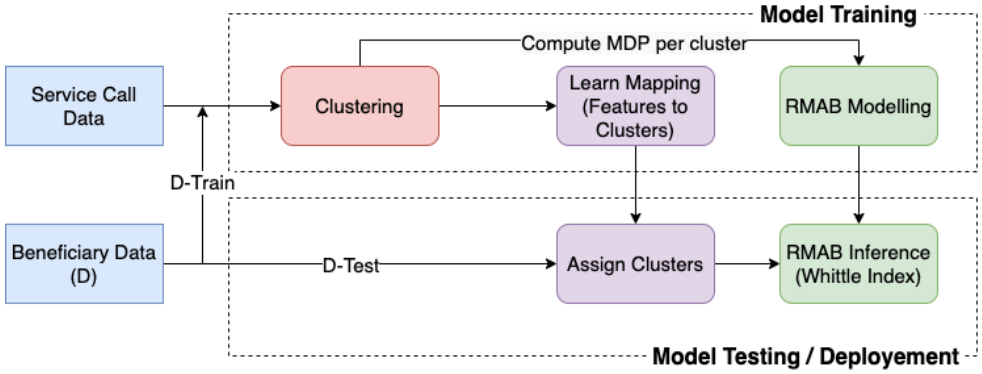


Figure 4.2: RMAB Training and Testing pipelines proposed

s. We assume that these (s, a, s') samples are drawn according to fixed, latent transition matrices $P^a_{ss'}[i]$ and $P^p_{ss'}[i]$ (corresponding to the active and passive actions respectively), unknown to the planner, and potentially unique to each beneficiary.

Given D_{train}, we now consider a new beneficiary cohort $\mathcal{D}_{test}$, consisting of N beneficiaries, marked $\{1, 2, \ldots, N\}$, that the planner must plan service calls for. The MDP transition parameters corresponding to beneficiaries in $\mathcal{D}_{test}$ are unknown to the planner, but assumed to be drawn at random from a distribution similar to the joint distribution of features and transition parameters of beneficiaries in the historical data distribution. We assume the planner has access to the feature vector f for each beneficiary in $\mathcal{D}_{test}$.

We now define the service call planning problem as follows. The planner has upto m resources available per round, which the planner may spend towards delivering service calls to beneficiaries. Beneficiaries are represented by N arms of the RMAB, of which the planner may pull upto m arms (i.e., m service calls) at each time step. We consider a round or timestep of one week which allows planning based on the most recent engagement patterns of the beneficiaries.

4.5 Methodology

Figure 4.2 shows our overall solution methodology. We use clustering techniques that exploit historical data D_{train} to estimate an offline RMAB problem instance relying solely on the beneficiaries' static features and state transition data. This enables overcoming the challenge of limited samples (time-steps) per beneficiary. Based on this estimation, we use the Whittle Index approach to prioritize service calls.

4.5.1 Clustering Methods

We use historical data $\mathcal{D}_{train}$ to learn the impact of service calls on transition probabilities. While there is limited service call data (active transition samples) for any single beneficiary, clustering on the beneficiaries allows us to combine their data to infer transition probabilities for the entire group. Clustering offers the added advantage of reducing computational cost for resource limited NGOs; since all beneficiaries within a cluster share identical transition probability values we can compute their Whittle index all at once. We present four such clustering techniques below:

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

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

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

4. Passive Transition-Probability based Clustering (PPF):

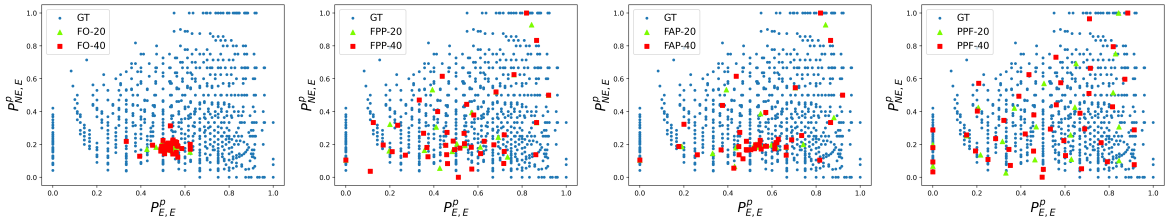
The key motivation here is to group together beneficiaries with similar transition behaviors, irrespective of their features. To this end, we use k-means clustering on passive transition probabilities (to avoid issues with missing

active data) of beneficiaries in D_{train} and identify cluster centers. We then learn a map φ from the feature vector f to the cluster assignment of the beneficiaries that can be used to infer the cluster assignments of new beneficiaries at test-time solely from f. We use a random forest model as φ.

The rule-based clustering on features involved in *FPP* and *FAP* methods can be thought of as using one specific, hand-tuned mapping function φ. In contrast, the *PPF* method *learns* such a map φ from data, eliminating the need to manually define accurate and reliable feature buckets.

4.5.2 Evaluation of Clustering Methods

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



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

Figure 4.3: Comparison of passive transition probabilities obtained from different clustering methods with cluster sizes $k = \{20, 40\}$ with the ground truth transition probabilities. Blue dots represent the true passive transition probabilities for every beneficiary while red or green dots represent estimated cluster centres.

1. Representation: Cluster centers that are representative of the underlying data distribution better resemble the ground truth transition probabilities. This is of prime importance to the planner, who must rely on these values to plan actions. Fig 4.3 plots the ground truth transition probabilities and the resulting cluster centers determined using the proposed methods. Visual inspection reveals that the *PPF* method represents the ground truth well, as is corroborated by the quantitative metrics of Table 4.1 that compares the RMSE error across different clustering methods.

2. Balanced cluster sizes: A low imbalance across cluster sizes is desirable to preclude the possibility of arriving at few, gigantic clusters which will assign identical whittle indices to a large groups of beneficiaries. Working with smaller clusters also aggravates the missing data problem in estimation of active transition probabilities.

Considering the variance in cluster sizes and RMSE error for the different clustering methods with $k = \{20, 40\}$ as shown in Table 4.1, *PPF* outperforms the other clustering methods and was chosen for the pilot study.

Table 4.1: Average RMSE and cluster size variance over all beneficiaries for different methods. Total Beneficiaries = 4238, $\mu_{20} = 211.9$, $\mu_{40} = 105.95$ (μ = average beneficiaries per cluster)

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

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

Finally, we adopt the Whittle solution approach for RMABs to plan actions and pre-compute all of the possible $2 * k$ index values that beneficiaries can take (corresponding to combinations of k possible clusters and 2 states). The indices can then be looked up at all future time steps in constant time, making this an optimal solution for large scale deployment with limited compute resources.

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

4.6 Experimental Study

In this section, we discuss a real-world quality improvement study. We also simulate the expected outcome in other synthetically constructed situations and demonstrate good performance of our approach across the board.

4.6.1 Service Quality Improvement Study

Setup

This cohort of beneficiaries registered in the program between Feb 16, 2021 and March 15, 2021 as D_{test} and started receiving automated voice messages few days post enrolment as per their gestational age. Additionally, as per the current standard of care, any of these beneficiaries could initiate a service call by placing a "missed call". The 23003 beneficiaries are randomly distributed across 3 groups, each group adding to the CSOC as follows:

- **Current-Standard-of-Care (CSOC) Group**: The beneficiaries in this group follow the original standard of care, where there are no ARMMAN initiated service calls. The listenership behavior of beneficiaries in this group is used as a benchmark for the RR and RMAB groups.
- **RMAB group**: In this group, beneficiaries are selected for ARMMAN-initiated service call per week via the Whittle Index policy described in Section 4.3. Even though all beneficiaries within a cluster are modeled by identical MDP parameters, their states may evolve independently, and so the Whittle indices are tracked for each beneficiary separately, leading to an RMAB with 7668 arms.
- **Round Robin (RR) group**: By default, NGOs including ARMMAN often conduct service calls using some systematic set order – the idea here is to have an easily executable policy, that services enough of a cross-section of beneficiaries and can be scaled up or down per week based on available resources. To recreate this setting, we generate service calls to beneficiaries based on the ascending order of their date of enrollment for this RR group, as recommended by ARMMAN. If this method succeeds compared to CSOC, then a simple manual strategy is enough; RMAB style optimization may not be needed.

Table 4.2 shows the absolute number of beneficiaries in states E or NE, where the state is computed using one week of engagement data between April 19 - April 26, 2021.

Beneficiaries across all three groups receive the same automated voice messages regarding pregnancy and post-birth care throughout the program, and no health related information is withheld from any beneficiary. The study only aims to evaluate the effectiveness of ARMMAN-initiated outbound service calls with respect to im-

Table 4.2: Beneficiary distribution in the three groups and their start states during week 0 of the study.

Group	Engaging (E)	Non-Engaging (NE)	Total
RMAB	3571	4097	7668
RR	3647	4021	7668
CSOC	3661	4006	7667

proving engagement with the program across the three groups. No interviews or research data or feedback was collected from the beneficiaries.

The study started on April 26, 2021, with m beneficiaries selected from the RMAB and RR group each ($m \ll N$) per week for ARMMAN-initiated service calls. ARMMAN staff performing service calls were blind to the experimental groups that the beneficiaries belonged to. Recall, the goal of the service calls is to encourage the beneficiaries to engage with the health information message program in the future. For this study, number of service calls m was on average 225 per week for each of RMAB and RR groups to reflect real-world constraints on service calls. The study was scheduled for a total of 7 weeks, during which 20% of the RMAB (and RR) group had received a service call, which is closer to the percentage of population that may be reached in service calls by ARMMAN. [‡]

Results

We present our key results from the study in Figure 4.4. The results are computed at the end of 7 weeks from the start of the quality improvement study on April 26, 2021.

Figure 4.4 measures the impact of service calls by the RMAB and RR policies in comparison to the CSOC Group. Beneficiaries' engagement with the program typically starts to dwindle with time. In Figure 4.4, we measure the impact of a service call policy as the cumulative drop in engagement prevented compared to the CSOC Group. We consider drop in engagement instead of the raw engagement numbers themselves, because of the slight difference in the numbers of beneficiaries in engaging (E) state at the start of the study. The drop in en-

[‡]Each beneficiary group also received very similar beneficiary-initiated calls, but these were less than 10% of the ARMMAN-initiated calls in RMAB or RR groups over 7 weeks.

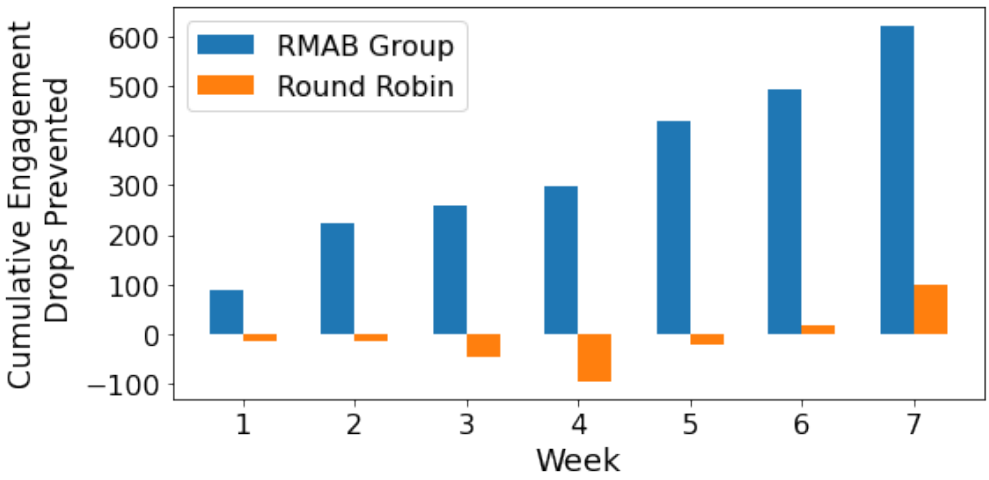


Figure 4.4: Cumulative number of weekly engagement drops prevented (in comparison to the CSOC group) by RMAB far exceed those prevented by RR.

gagement under a policy π at time t can be measured as the change in engagement:

$$\Delta^{\pi}_{current}(t) := \sum_{n \in N} (R_n(s_0) - R_n(s_t)) \tag{4.2}$$

where $R_n(s_t)$ represents the reward for n^{th} beneficiary in state s_t at time step t and cumulative drop in engagement is:

$$\Delta^{\pi}_{cumulative}(t) := \sum_{n \in N} \sum_{\zeta=0}^{\zeta=t} (R_n(s_0) - R_n(s_\zeta)) \tag{4.3}$$

The cumulative drop in engagement prevented by a policy π, in comparison to the CSOC Group is thus simply:

$$\Delta^{\pi}_{cumulative}(t) - \Delta^{CSOC}_{cumulative}(t) \tag{4.4}$$

and is plotted on the y-axis of Figure 4.4.

Figure 4.4 shows that the RMAB policy prevents a total 622 instances of a drop in automated health message engagement, at the end of 7 weeks, as compared to CSOC. RR group, on the other hand, only prevents 101 engagement drops by the end of week 7. Given that there are a total of 1944 engagement drops in the CSOC group, we show in the first row of Table 4.3, that the RMAB group has 32.0% and 28.3% less cumulative engagement drops as compared to the CSOC and RR groups respectively by the end of the study.

Table 4.3: Statistical significance for service call policy impact at week 7 is tested using a linear regression model. We use: $^{*}p < 0.05$; $^{\dagger}p < 0.1$

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

To investigate the benefit from use of RMAB policy over policies in the RR and CSOC groups, we use regression analysis[6] [§]. Specifically, we fit a linear regression model to predict number of cumulative engagement drops at week 7 while controlling for treatment assignment and covariates specified by beneficiary registration features. The model is given by:

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

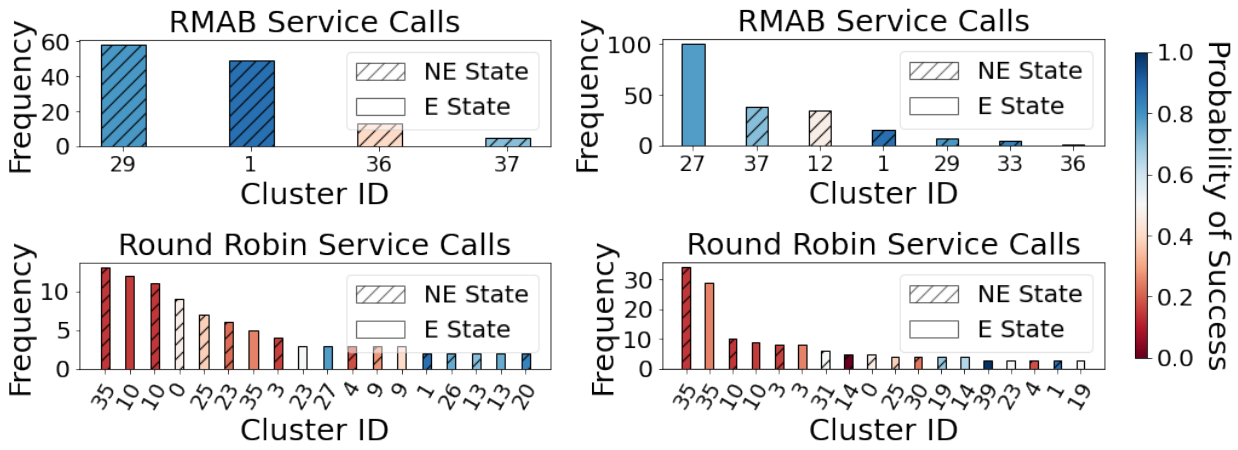
where for the i_{th} beneficiary, Y_i is the outcome variable defined as number of cumulative engagement drops at week 7, k is the constant term, β is the treatment effect, T_i is the treatment indicator variable, x_i is a vector of length J representing the i_{th} beneficiary's registration features, γ_j represents the impact of the j_{th} feature on the outcome variable and ε_i is the error term. For evaluating the effect of RMAB service calls as compared to CSOC group, we fit the regression model only for the subset of beneficiaries assigned to either of these two groups. T_i is set to 1 for beneficiaries belonging to the RMAB group and 0 for those in CSOC group. We repeat the same experiment to compare RR vs CSOC group and RMAB vs RR group.

The results are summarized in Table 4.3. We find that RMAB has a statistically significant treatment effect in reducing cumulative engagement drop (negative $\beta, p < 0.05$) as compared to CSOC group. However, the treatment effect is not statistically significant when comparing RR with CSOC group ($p = 0.740$). Additionally, comparing RMAB group with RR, we find β, the RMAB treatment effect, to be significant ($p < 0.1$). This shows that RMAB policy has a statistically significant effect on reducing cumulative engagement drop as com-

[§]See Appendix D.1 for erratum

pared to both the RR policy and CSOC. RR fails to achieve statistical significance against CSOC. Together these results illustrate the importance of RMAB's optimization of service calls, and that without such optimization, service calls may not yield any benefits.

RMAB Strategies



(a) Week 1 Service Calls

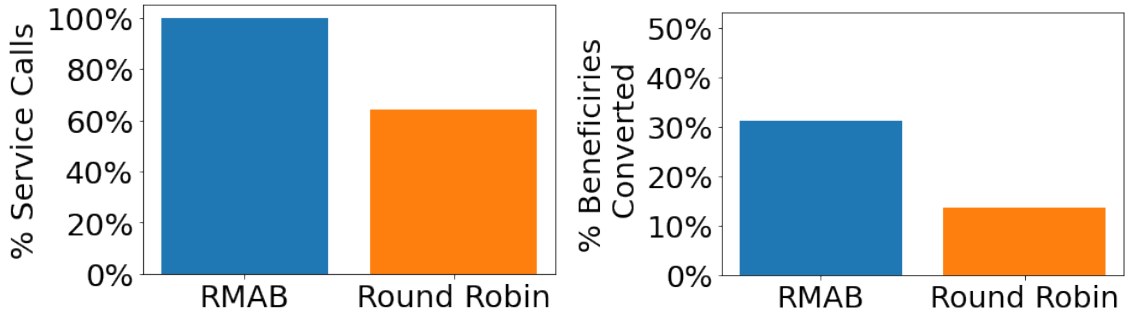
(b) Week 2 Service Calls

Figure 4.5: Distributions of clusters picked for service calls by RMAB and RR are significantly different. RMAB is very strategic in picking only a few clusters with a promising probability of success, RR displays no such selection.

We analyse RMAB's strategic selection of beneficiaries in comparison to RR using Figure 4.5, where we group beneficiaries according to their whittle indices, equivalently their $\langle$`cluster, state`$\rangle$. Figure 4.5 plots the frequency distribution of beneficiaries (shown via corresponding clusters) who were selected by RMAB and RR in the first two weeks. For example, the top plot in Figure 4.5a shows that RMAB selected 60 beneficiaries from cluster 29 (NE state).

First, we observe that RMAB was clearly more selective, choosing beneficiaries from just four (Figure 4.5a) or seven (Figure 4.5b) clusters, rather than RR that chose from 20. Further, we assign each cluster a hue based on their probability of transitioning to engaging state from their current state given a service call. Figure 4.5 reveals that RMAB consistently prioritizes clusters with high probability of success (blue hues) while RR deploys no such selection; its distribution emulates the overall distribution of beneficiaries across clusters (mixed blue and red hues).

Furthermore, Figure 4.6a further highlights the situation in week 1, where RMAB spent 100% of its service calls on beneficiaries in the non-engaging state while RR spent the same on only 64%. Figure 4.6b shows that RMAB converts 31.2% of the beneficiaries shown in Figure 4.6a from non-engaging to engaging state by week 7, while RR does so for only 13.7%. This further illustrates the need for optimizing service calls for them to be effective, as done by RMAB.



(a) **(b)** Week 2 Service Calls

Figure 4.6: (a) % of week 1 service calls on non-engaging beneficiaries (b) % of non-engaging beneficiaries of week 1 receiving service calls that converted to engaging by week 7

4.6.2 Synthetic Results

We run additional simulations to test other service call policies beyond those included in the quality improvement study and confirm the superior performance of RMAB. Specifically, we compare to the following baselines: (1) Random is a naive baseline that selects m arms at random. (2) Myopic is a greedy algorithm that pulls arms optimizing for the reward in the immediate next time step. Whittle is our algorithm. We compute a normalized reward of an algorithm ALG as: $\frac{100\times(\overline{R}^{\text{ALG}}-\overline{R}^{\text{CSOC}})}{\overline{R}^{\text{WHITTLE}}-\overline{R}^{\text{CSOC}}}$ where $\overline{R}$ is the total discounted reward. Simulation results are averaged over 30 independent trials and run over 40 weeks.

Figure 4.7 presents simulation of an adversarial example [101] consisting of $x\%$ of non-recoverable and $100-x\%$ of self-correcting beneficiaries for different values of x. Self-correcting beneficiaries tend to miss automated voice messages sporadically, but revert to engaging ways without needing a service call. Non-recoverable beneficiaries are those who may drop out for good, if they stop engaging. We find that in such situations, MYOPIC proves brittle, as it performs even worse than RANDOM while WHITTLE performs well consistently. The actual

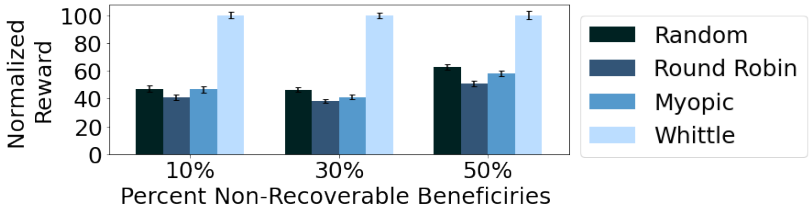


Figure 4.7: Performance of myopic can be arbitrarily bad and even worse than Random, unlike the Whittle policy.

quality improvement study cohort consists of 48.12% non-recoverable beneficiaries (defined by $P_{01}^{p} < 0.2$) and the remaining comprised of self-correcting and other types of beneficiaries.

4.7 Conclusions and Lessons Learned

The widespread use of cell-phones, particularly in the global south, has enabled non-profits to launch massive programs delivering key health messages to a broad population of beneficiaries in a cost-effective manner. We present an RMAB based system to assist these non-profits in optimizing their limited service resources. To the best of our knowledge, ours is the first study to demonstrate the effectiveness of such RMAB-based resource optimization in real-world public health contexts. These encouraging results have initiated the transition of our RMAB software to ARMMAN for real-world deployment. We hope this work paves the way for use of RMABs in many other health service applications.

Some key lessons learned from this research, which complement some of the lessons outlined in [166,41,151] include the following. First, social-impact driven engagement and design iterations with the NGOs on the ground is crucial to understanding the right AI model for use and appropriate research challenges. As discussed in footnote 1, our initial effort used a one-shot prediction model, and only after some design iterations we arrived at the current RMAB model. Next, given the missing parameters in RMAB, we found that the assumptions made in literature for learning such paramters did not apply in our domain, exposing new research challenges in RMABs. In short, *domain partnerships with NGOs to achieve real social impact automatically revealed requirements for use of novel application of an AI model (RMAB) and new research problems in this model*.

Second, *data and compute limitations of non-profits are a real world constraint, and must be seen as genuine research challenges in AI for social impact, rather than limitations.* In our domain, one key technical contribution in our RMAB system is deploying clustering methods on offline historical data to infer unknown RMAB param-

eters. Data is limited as not enough samples are available for any given beneficiary, who may stay in the program for a limited time. Non-profit partners also cannot bear the burden of massive compute requirements. Our clustering approach allows efficient offline mapping to Whittle indices, addressing both data and compute limits, enabling scale-up to service 10s if not 100s of thousands of beneficiaries. Third, *in deploying AI systems for social impact, there are many technical challenges that may not need innovative solutions, but they are critical to deploying solutions at scale.* Indeed, deploying any system in the real world is challenging, but even more so in domains where NGOs may be interacting with low-resource communities. We hope this work serves as a useful example of deploying an AI based system for social impact in partnership with non-profits in the real world and will pave the way for more such solutions with real world impact.

Finally, there are also some important topics for future work in improving the RMAB system, which include handling fairness [104], changing the current RMAB model with two actions to incorporate multiple actions [80], and improving the RMAB model from interactions with beneficiaries [22].

5

Non-Stationary and Restless Bandits for Improved Intervention Planning

5.1 Introduction

In many public health contexts, monitoring patient adherence or beneficiary engagement with prescribed health programs has been shown to be an important problem[99]. The healthcare domain has seen several studies on patient adherence for diseases like HIV[154], cardiac problems[144,33], tuberculosis[82,127], etc. In low-resource settings, such monitoring resources may be severely limited, thereby making it critical to utilize these resources optimally.

As a motivating example, we consider a popular maternal healthcare task in which limited health workers

strive to monitor and improve engagement of enrolled beneficiaries with the information program. Non-profits such as[70,12,56] deliver ante- and post-natal care information via voice and text to enrolled beneficiaries, aiming to prevent adverse health outcomes. Unfortunately, such information programs often see dwindling rates of engagement among enrollees, with a large fraction even dropping out completely leading to negative health outcomes. Such non-profits have very limited health worker resources at their disposal to place crucial service calls to beneficiaries, to encourage participation, address complaints and prevent dropouts. Viewed algorithmically, these health workers must choose k out of the total N beneficiaries (where $k \ll N$) on a periodic basis (eg. weekly) for service calls, such that the chosen k beneficiaries are likely to benefit the most from the service calls.

Several works in public health have modeled similar challenges as Restless Multi-Armed Bandit (RMAB) problems, in contexts such as monitoring tuberculosis medication adherence[100], screening for hepatocellular carcinoma[86], spreading health awareness[18], delivering health services through mobile health clinics[123], besides existing work on improving maternal health outcomes[22].

One key assumption about the RMAB setting however, is that the bandit arms in an RMAB must follow fixed, but possibly distinct Markov Decision Processes (MDPs) with fixed transition parameters. In context of the maternal healthcare application (used as a running example hereafter), we find from real-world data collected over a 3-month long study presented in[102], involving 23,003 real-world beneficiaries that the Markov model may not fit the data well. In fact, we find that the state transition parameters of beneficiaries, assumed to remain fixed in the MDP model, show transient patterns.

In this chapter, in a bid to improve resource allocation for the engagement monitoring problem, our contributions are as follows: (1) We cast the planning challenge as a Non-stationary Restless Multi-Armed Bandit (RMAB-NS) that admits time-varying transition parameters, $\mathcal{P}(t)$ instead of a fixed point $\mathcal{P}$ and present a Whittle index based solution technique. (2) We prove 'indexability' for the RMAB-NS problem. Indexability is a technical condition that guarantees existence and asymptotic optimality of the Whittle index solution. Our proof hinges on showing a reduction from the RMAB-NS problem to the standard RMAB framework via a state-space expansion trick. (3) We propose a practical algorithm based on an index interpolation idea, useful particularly in long-horizon planning problems when the transition function $\mathcal{P}(t)$ changes linearly. Our algorithm provides a $30\times$ speed-up for a planning horizon $L = 100$, with a marginal sacrifice on performance. (4) From a practical standpoint, the time-dependent transition parameter model exacerbates the challenge of predicting these values

for real-world agents, especially in data-scare settings. We solve this challenge via a technique that identifies a few, distinct behavior patterns among agents using clustering, and pools data by bucketing agents according to these behavior patterns.(5) Finally, we evaluate using both synthetic as well as real data from a maternal healthcare application, and show improved performance over existing baselines.

5.2 Preliminaries

5.2.1 Stationary Restless Multi-Armed Bandits

A stationary RMAB instance consists of N independent 2-action Markov Decision Processes (MDP)[130], where each MDP is defined by the tuple $\{\mathcal{S}, \mathcal{A}, R, \mathcal{P}\}$. $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the set of possible actions, R is the reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ and $\mathcal{P}$ represents the transition function. L denotes the total length of horizon over which the MDP continues. In stationary RMABs, $\mathcal{P}$ is assumed to be constant (stationary in time) throughout this time horizon. We use $P^{\alpha}_{s,s'}$ to denote the probability of transitioning from state s to state s' under the action α. The policy π, is a mapping $\pi : \mathcal{S} \to \mathcal{A}$ that selects the action to be taken at a given state. The total reward accrued can be measured using either the discounted or average reward criteria to sum up the immediate rewards accrued by the MDP at each time step. Our formulation is amenable to both, although we use the discounted reward criterion in our study.

The expected *discounted reward* starting from state s_0 is defined as $V^{\pi}_{\beta}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t R(s_t, \pi(s_t), s_{t+1} | \pi, s_0)\right]$ where the next state is drawn according to $s_{t+1} \sim P^{\pi(s_t)}_{s_t,s_{t+1}}$, $\beta \in [0, 1)$ is the discount factor and actions are selected according to the policy mapping π. The planner's goal is to maximize the total reward.

In context of the engagement monitoring problem, the stationary RMAB formulation models the engagement behavior of each beneficiary as an MDP corresponding to an arm of the RMAB. Pulling an arm corresponds to an active action, i.e., making a service call (denoted by $\alpha = a$ or $\alpha = 1$), while $\alpha = p$ or $\alpha = 0$ denotes the passive action of abstaining from a call. The state space $\mathcal{S}$ consists of binary valued states, s, that account for the recent engagement behavior of the beneficiary; $s \in [NE, E]$ (or equivalently, $s \in [0, 1]$) where E and NE denote the 'Engaging' and 'Not Engaging' states respectively. For each action $\alpha \in \mathcal{A}$, the beneficiary states follow a Markov chain represented by the 2-state Gilbert-Elliot model[44] with transition parameters given by $P^{\alpha}_{ss'}$, as shown in Figure 5.1. With slight abuse of notation, the reward function $R(.)$ of the MDP is simply given by $R(s) = s$

for $s \in \{0, 1\}$.

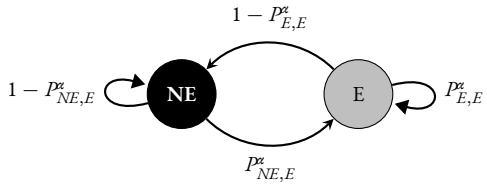


Figure 5.1: The beneficiary transitions from a current state s to a next state s' under action α, with probability $P^{\alpha}_{s,s'}$.

5.2.2 Whittle Index Solution Approach

Computing the optimal policy for an RMAB is PSPACE-hard in general, even when the transition functions are stationary and perfectly known. [125]. Whittle in [164], proposed an index-based heuristic, known today as the Whittle Index, that can be solved in polynomial time. Today, the Whittle solution approach is the predominant technique used for solving the RMAB. It hinges around the key idea of a "passive subsidy", which is a hypothetical reward offered to the planner, in addition to the original reward function for choosing the passive action. The Whittle Index is then defined as the infimum subsidy that makes the planner indifferent between the 'active' and the 'passive' actions, i.e.,:

$$W(s) = inf_{\lambda}\{\lambda : Q_{\lambda}(s, 0) = Q_{\lambda}(s, 1)\} \tag{5.1}$$

However, passive subsidy is well-defined only if a technical condition called 'indexability' is satisfied. Further, indexability also guarantees the asymptotic optimalilty of the Whittle Index technique. The Whittle approach has been shown to be asymptotically optimal for the time average reward problem [160] and other families of RMABs arising from stochastic scheduling problems [47].

5.2.3 Related Work

Healthcare Intervention Planning The healthcare domain has seen several studies on patient adherence for diseases like HIV [154], cardiac problems [144,33], tuberculosis [82,127], etc. These studies propose models that operate by predicting beneficiaries at a high-risk of defaulting on the prescribed programs. However, oftentimes, the pool of beneficiaries marked as high-risk itself can be large, and thus prioritization of beneficiaries for intervention becomes a challenge.

Optimization of limited intervention resources has been widely studied in the past, for applications such as campaign optimization via phone outreach [87,37]. Some studies also consider application to public health [123], but assume that access to some social network of beneficiaries is available. When such a network is unknown or when optimization over multiple timesteps is involved, RMABs have been shown to be useful in assisting engagement monitoring in many other adherence or awareness programs such as [149,31].

Restless Multi-Armed Bandits (RMABs) This framework has been popularly adopted to tackle sequential resource allocation problems in a myriad of application domains such as anti-poaching patrol planning [132], multi-channel communication systems [95], sensor monitoring tasks [45], UAV routing [85] etc. Even in the public health domain, several works [104,86,18,123] show applicability of RMABs for the engagement monitoring and intervention resource planning problem. However, most of these works assume agents to follow the Markov property. In addition, most of these works also unrealistically assume perfect knowledge of the RMAB parameters. [22,15,115], present Whittle Index-based Q-learning or Deep RL approaches for handling unknown RMAB parameters. However, their techniques either assume identical arms or rely on receiving large number of samples from each arm. This may be an unrealistic requirement in setting such as the maternal healthcare program, where the overall stay may be limited — a beneficiary may drop out or stop engaging with the program few weeks post enrolment unless an intervention convinces them to do otherwise. [102] reports results from a real-world experiment with 23003 beneficiaries demonstrating the utility of RMABs for optimizing use of limited service calls for maximizing beneficiary engagement in the program.

Clustering in the context of Multi-Armed Bandit and Contextual Bandits has received significant attention previously [43,90,172,88], but these studies do not consider the restless bandit setup. [111] tackles a non-stationary setup with stochastic rewards, while [16] infers model parameters from independent studies in absence of historic data.

Non-Stationary Bandits Several other works consider non-stationary bandits [175,136,17] in which the rewards from bandit arms are allowed to change with time, but are not dependent on arms chosen or satisfy other simplifying assumptions and thus do not capture the complexities of the RMAB setting.

5.3 Engagement Monitoring Problem

We formalize the engagement monitoring and intervention planning problem as follows. The setup consists of a cohort of N agents in a health program, each with 2-states ('Engaging (E)' denoted by $s = 1$ and 'Not Engaging (NE)' denoted as $s = 0$) participating for a duration of L timesteps. At a each timestep, there are k intervention resources available to encourage agents to engage with the program. Depending on whether an agent received such an intervention or not, and depending on the time-step in the program, each agent may transition from a state s_t at time t to a state s_{t+1} at timestep $t+1$, with some probability $P^{a}_{s_t s_{t+1}}(t)$. The states and state transition matrix may be unique for each agent $i \in [N]$ and may be formally represented as $s_t[i]$ and $P^{a}_{s_t s_{t+1}}(t)[i]$, but we drop i from the notation (as above) when the agents index i is irrelevant. Note that this setup is different from a stationary RMAB in that the transition function P is not stationary, but is a function of time t. In this problem, the planner represents the health worker responsible for managing the engagement outcomes for the set of agents being catered to.

Data We assume that the planner has access to a bunch of historical data $\mathcal{D}$ from previous agents enrolled in the program. $\mathcal{D}$ consists of two parts: For the first part, each agent in $\mathcal{D}$ has an associated static feature vector $\mathbf{f}$. In the context of maternal healthcare, these may be demographic features such as age, income, location, etc. The second part consists of state and action trajectories over the L timesteps recording the engagement status of agents each timestep. In our problem, both the state and action are assumed to be binary variables. In the maternal healthcare example, the state represents the engagement status of enrolled mothers and the action variable denotes whether or not the mother was screened for receiving service calls that timestep.

Problem Definition The planner's goal is to maximize the total engagement across the agent cohort over the duration of the program. Formally, the planner wishes to maximize the total reward (either discounted or average):

$$\bar{R} := \sum_{t \in [L]} \sum_{i \in [N]} R_t(i). \tag{5.2}$$

In doing so, the planner has access to a limited budget k of intervention resources (such as manual service calls to mothers) to nudge agents to the engaging state. The goal of the planner is to decide which k agents to select

for delivery of intervention (out of the N agents in total) at each time step t, to maximize the overall engagement. Note that the state transition matrix $P^{a}_{s_t s_{t+1}}(t)[i]$, governing the agent behavior is unknown to the planner for all agents.

5.4 Methodology: Planning in RMAB-NS

5.4.1 Whittle Index Approach

We adopt the Whittle Index technique as our solution approach. Typical approaches entail adopting value iteration to estimate the value functions $Q_\lambda(s,0)$ and $Q_\lambda(s,1)$. This method however fails, given the time-varying transition matrices. We circumvent this issue in two stages. First, for small residual horizon values, we modify the Bellman equation by appropriately supplying the applicable $\mathcal{P}(t)$ at time-step t, to compute the finite-horizon value functions, $Q^t_\lambda(s,1)$instead. These can be computed as shown below:

$$Q^t_\lambda(s,a) = R(s,a) + \lambda.1_{\{a==0\}} + \beta.\big(\sum_{s'} \mathcal{P}(t).V^{t+1}_\lambda(s')\big)$$

where $\mathcal{P}(t) = P^{a}_{s,s'}(t)$. Finally, we modify Equation 5.1 and apply a binary search based algorithm of [132] to search for the index λ satisfying: $Q^t_\lambda(s,0) = Q^t_\lambda(s,1)$. This method provides apparatus to compute the index for any size of planning horizon L. However, the second stage of our solution consists of leveraging the structure of $\mathcal{P}(t)$ to compute the index for larger horizons inexpensively. We present this technique in Section 5.4.3.

5.4.2 Conditions for Indexability in RMAB-NS

In this subsection, we show results on indexability from the RMAB-NS setup. Indexability is crucial to guaranteeing existence of the Whittle index in the first place, and also to prove its optimality.

Theorem 13. *The RMAB-NS problem instance with time-varying transition parameters $\mathcal{P}(t)$ is indexable under the same sufficient conditions of indexability for the standard RMAB problem.*

Proof Sketch. We prove the theorem by showing a reduction of the RMAB-NS problem to a standard RMAB setup, using a state space expansion technique that incorporates the time-step t within the state definition. We

also add a dummy sink state that the MDP transitions to almost surely at the end of the horizon length L. This expands the state space to size $\|\mathcal{S}\|L + 1$. We then show that the passive and active Q-values in the original RMAB-NS problem match the corresponding Q-values in the reduced problem (differing by upto constant). Thus if the reduced problem with augmented state space satisfies the indexability condition in difference of the Q-values, it also implies indexability for the RMAB-NS problem. Existing results on conditions for indexability of observable MDPs [5] thus also extend to our setup. □

5.4.3 Fast Algorithm

In this section we propose a fast algorithm for computing the intervention policy when the transition parameters change as a linear function of residual horizon. Our algorithm is particularly useful if the planning horizon is large. The central idea powering our algorithm is based on the following proposition:

For an agent with transition parameters $\mathcal{P}(t)$ changing linearly with time t, the Whittle index $W_\tau(s)$ can also be approximated as a linear function of the residual horizon, τ for large values of τ.

As shown in Figure 5.2, the Whittle index trend can be decomposed and approximated as a two-piece function. The first piece represents the index for small horizon values ($\approx 0 \leq x \leq 5$ in the figure). Previous work [100] establishes the 'index decay' phenomenon which shows that the index value decays as the residual horizon approaches zero. [100] also proposes a technique in which the Whittle index in the index decay phase can be approximated as a linear or logistic function.

The second piece captures the linear trend for large horizon values, stemming from the non-stationary transition parameters ($\approx x > 5$ in figure 5.2). Previous work [100] fails to account for the impact of non-stationary parameters, stopping at only addressing index decay. Whereas previous work regards the index value to be constant for large horizons, our contribution is to leverage the observed linear index trend and propose to approximate the Whittle index using a one-switch piece-wise linear function. This technique allows sampling the Whittle index only at a few values of residual horizon τ and interpolate to other values, without having to compute the exact index for all τ. We describe the algorithmic details in Algorithm 4.

Algorithm 4: Fast Algorithm

Input: Transition function $\mathcal{P}(t)$ changing linearly with size $S \times A \times T$ for every time step in horizon
Compute Whittle Indices for four values of residual horizon:
$W_0(s) := WI(horizon = 0) = 0$(known to be 0)
$W_1(s) := WI(horizon = 1) = P_{s,1}^{a=1} - P_{s,1}^{a=0}$
$W_{L-1}(s) := WI(horizon = L-1)$
$W_{L-2}(s) := WI(horizon = L-2)$
Compute t^*, the point of intersection of the two-piece linear curves as:
$t^* = L - \frac{W_1(s) - W_{L-1}(s)}{W_{L-2}(s) - W_{L-1}(s)} - 1$
initialize $W_t(s)$ as an array of size L.
Set $W_t(s) = t.W_1(s) \; \forall \, t$ in $0 \leq t \leq t^*$.
Set $W_t(s) = W_{L-1} + (W_{L-2} - W_{L-1})(L-1-t) \; \forall \, t$ in $t^* \leq t \leq L$.
return $W_t(s)$

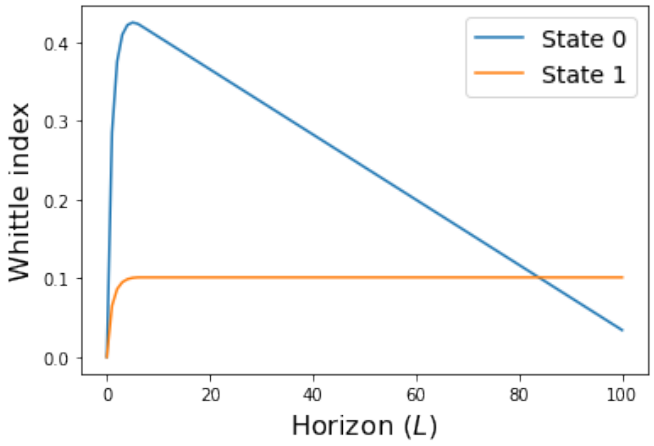


Figure 5.2: Whittle index (on y-axis) computed for the two possible states (blue and orange lines) shows an approximately piece-wise linear trend as a function of residual horizon (on x-axis), when the transition functions change linearly with time.

5.5 Inferring Transition Parameters

While the richer RMAB-NS model comes with more expressive power and allows for better optimized intervention planning, it also significantly amplifies the challenge of learning the appropriate model parameters, especially from scarce data. We counter this issue by proposing a novel technique of grouping agents with similar transition parameters into clusters to aid parameter estimation.

5.5.1 Discovering Behavior Patterns

We hypothesize that the transition behavior of agents being catered to, belongs to one of a finite collection of unique behavior types. These behavior types may be unique to the population of agents and type of application considered, and may be potentially different for different application domains. We aim to identify these unique behavior types through the available historical data $\mathcal{D}$. We leverage unsupervised clustering techniques, which provide a natural solution to this task of identifying agents with similar behavior patterns, and allow grouping them together.

Clustering Methodology

Clustering for trajectories has been extensively studied in the existing literature [19]. Our method operates by extracting relevant feature representations from the trajectories by boiling them down to a vector of informative attributes, which can then be supplied as an input to standard clustering algorithms. We use k-means clustering in our setup.

To unearth behavior patterns among agents, we consider their observed state and action trajectories $\{s_t\}$ and $\{a_t\}$. The limited active action resources mean that the passive action transition samples are more abundantly available than the active samples. We utilize these passive samples to compute point estimates of the average passive transition probabilities for the observed time horizon as, $\hat{P}^0_{ss'} := \frac{\eta_{\{s_t=s\}\to\{s_{t+1}=s'\}|a=0}}{\eta_{\{s_t=s\}|a=0}}$, where η denotes the number of agents in the group satisfying the condition on state transition specified in the subscript, under an action a. This point estimate, computed separately for each agent in the training set (historical data, $\mathcal{D}$) encodes the representative vector. We perform clustering using available techniques such as k-means, over these estimated $\hat{P}^0_{ss'}$ values, arriving at C distinct clusters of transition values. Each cluster represents a group of beneficiaries with

similar behavior patterns. These clusters are discovered organically from data without needing manual segregation or specification by hand of possible behavior patterns.

All agents within each cluster are assumed to share the same transition probability values. This allows pooling of samples from several agents together, helping circumvent the challenge of limited data for parameter estimation, particularly for active transitions. We estimate time-dependent transition parameters, $\hat{P}^{a}_{ss'}(t)$ for each cluster $c \in [C]$ as: $\hat{P}^{a}_{ss'}(t) := \frac{\eta_{\{s_t=s\}\rightarrow\{s_{t+1}=s'\}|a}}{\eta_{\{s_t=s\}|a}}$ where η denotes the number of agents in the group satisfying the condition on state transition specified in the subscript, under an action $a \in \{0, 1\}$.

Choosing Clustering Hyperparameters

The number of clusters, C is a tunable hyperparameter. For large values of C, some clusters could be spread too thin, leading to insufficient samples to accurately estimate transition parameter values. On the other hand, making C too small can conflate distinct innate behavior patterns into a single cluster, giving rise to a more coarse-grained model. We generate an elbow plot measuring the clustering error to inform the choice of C as shown in Figure 5.3(a). In our experiments, for the maternal health program dataset, we arrive at $C = 40$ as the optimal choice for empirical evaluation.

5.5.2 Inferring Behavior Patterns

We use these estimated parameters to construct a look-up table, $\wp$, in which, for each cluster $c \in [C]$, $\wp(c)$ stores the sequence of non-stationary transition probabilities $\hat{P}^{a}_{ss'}(t)$ computed $\forall t \in [L]$ using the above method. We employ a predictive model φ to map each beneficiary to their cluster assignment c, using the demographic feature vector $\mathfrak{f}$ as: $c = \varphi(\mathfrak{f})$. We implement φ using a Random Forest model, but our method can admit any other implementation for φ.

5.6 Evaluation Testbed

5.6.1 Limitation of existing approaches

Restless Bandit Models typically model each arm of the bandit as an MDP with fixed transition parameters, $P^{a}_{ss'}$. Simulations based on this model operate by computing a point estimate for the transition parameters $\hat{P}^{a}_{ss'}$ from

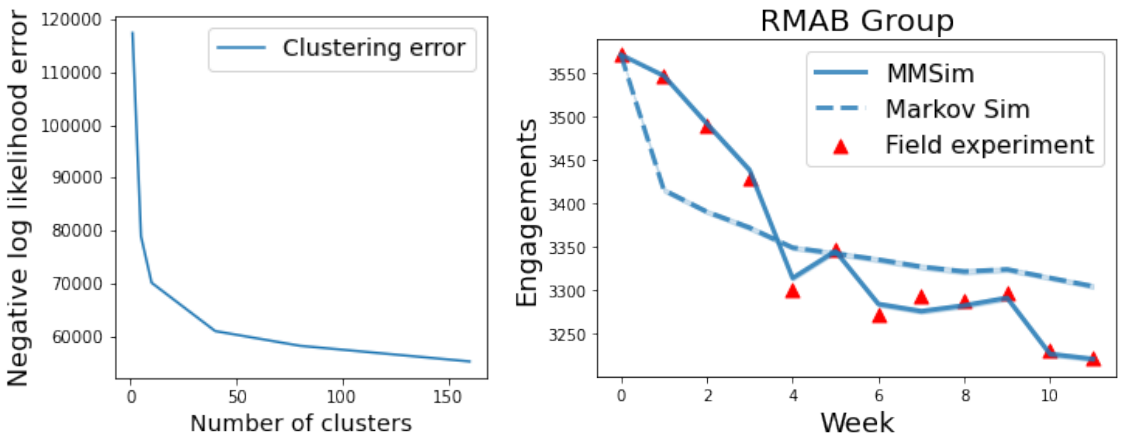


Figure 5.3: (a) Elbow plot measuring the clustering error informs the choice of ideal number of clusters. (b) The best-fit Markov simulator (dashed line) can only manage to crudely capture actual behavior, even when trained on the actual observations. The richer MMSim simulator on the other hand (solid line), is more expressive and is better suited for simulating an RMAB-NS environment.

available historical data to simulate transitions under new policies. However, while the markov assumption makes it analytically tractable, in effect, it restricts the possible trajectories spanned by the model to a specific subclass of all possible trajectories. For some classes of agents in the real-world, the Markov model can prove to be a crude approximation limited by the expressive power of the markov model.

On the other hand, simulations via fine-grained agent based models are impracticable to run in this scenario, given limited data on individual beneficiaries to inform tailored behavior models. Consequently, these simulated RMAB evaluations are only as veracious as the Markov assumption and may not dependably extend to real-world settings.

Figure 5.3(b) shows evidence of this phenomenon. We consider the engagement trajectories of beneficiaries participating in the maternal healthcare study presented in [102]. We extract the best fitting stationary transition parameters $\hat{P}^{a}_{ss'}$ for these beneficiaries from the observed transitions according to the Markov model and use that to simulate effect of the same actions as in the original study. In Figure 5.3, we plot the timestep on the x-axis and measure the engagements (either simulated or actual) on the y-axis. The dashed lines in Figure 5.3 show, via visual inspection, that the simulation output is far from the actual numbers (red triangles). The nature of curve is significantly different, as the output of the Markov simulator tends to saturate to its stationary state quickly, whereas the engagement values observed in practice continue to decline steadily.

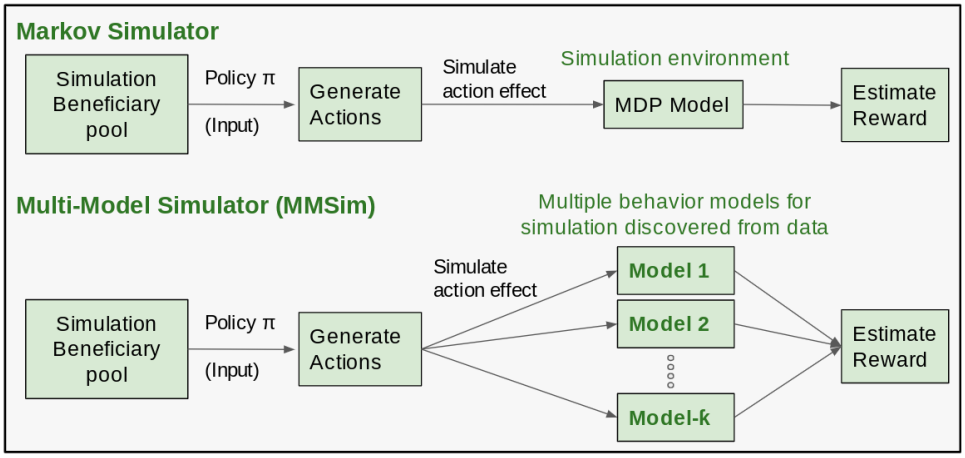


Figure 5.4: Overview of pipeline. The proposed Multi-environment Simulator involves discovering suitable environments from data that represent beneficiary behavior well. MDP environment, Oscillatory environment and others are examples of potential behavior models that could be discovered from actual data.

5.6.2 Overview of Simulator

To overcome this issue of limited evaluation accuracy in simulation, we propose a new simulation technique — the Multi-Model Simulator (MMSim)— that we show yields more accurate predictions. The key idea powering MMSim is to forgo the Markov assumption and build richer behavior models learnt from data. To avoid the challenges that ABMs encounter, MMSim pools data from agents with similar behavior patterns. Figure 5.4 gives a broad overview of MMSim. Note that the simulation environment operates independent of the policy generation module. The simulation environment treats the policy generation step as a black box, accepting the policy π as an input. It then simulates the effect of this policy π depending on the model it implements (for e.g. MDP model). In contrast to the Markov simulator, the coin tosses for each agent in MMSim are implemented via bespoke behavior models, designed to capture the unique transition behavior patterns, characteristic of their cluster.

5.7 Empirical Evaluation

In this section we evaluate our proposed RMAB-NS approach and solution technique on real as well as synthetic datasets. We build up the evaluation in five stages: (1) Evidence of non-stationary behavior patterns identified from real data (2) Evaluation of simulation testbed (3) Evaluation of planning algorithm (4) Evaluation of plan-

ning in combination with inferring real-world transition parameters (5) Synthetic data evaluation of fast algorithm.

Real Data: Service Call Allocation for Maternal Health For our empirical analysis, we obtain data collected in a real-world experiment reported in [102]. The experiment was carried out in partnership with an Indian non-profit, 'ARMMAN', and involved 23003 new or expecting mothers. The experiment spanned 13 weeks between May—July 2021, during which period all beneficiaries were provided with weekly automated voice calls delivering crucial health-related information. The weekly engagement status of these mothers with the voice calls was tracked, and on the basis of which, manual service calls were delivered to a small fraction of mothers each week, to encourage them to engage with the automated calls.

The data consists of two parts. For the first part, each beneficiary has an associated set of static demographic features, such as age, income, education level, location, etc. These features are recorded when beneficiaries enroll into ARMMAN's maternal health information program with the help of health workers. They also collect information such as phone owner in the family, gestation age, number of children, preferred language and preferred slots for the automated voice messages during enrolment. These features constitute $\mathfrak{f}$, the static feature vector available for each agent in the RMAB-NS model. Beneficiaries provided both written and digital consent for receiving automated voice messages and service calls.

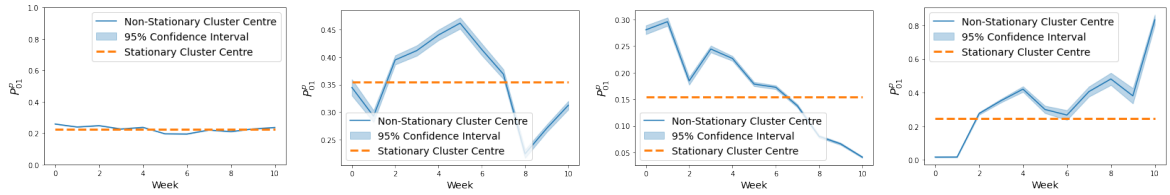
The second part consists of trajectories of engagement status of beneficiaries over the 13 weeks of the experiment, along with binary action data indicating whether or not the beneficiary was screened for receiving service calls that week. ARMMAN also stores this listenership information regarding the automated voice messages together with the registration data in an anonymized fashion.

ARMMAN considers that if a beneficiary stays on the automated voice message for more than 30 seconds (average message length is 1 minute), then the beneficiary has engaged. If a beneficiary engages at least once with the automated voice messages sent during a week, they are assigned the engaging (E or $s = 1$) state for that time step and non-engaging (NE or $s = 0$) state otherwise.

5.7.1 Real Data: Behavior Patterns Unearthed

We test out the proposed methodology on the real-world data generated from the maternal health experiment and find that it churns out a finite number of unique behavior patterns exhibited by beneficiaries. We generate the look-up table $\wp$ using k-means clustering on data pooled together from all participant beneficiaries in the experiment. Figures 5.5c, 5.5b and 5.5a, illustrate some examples of patterns identified in $\wp$. In each of these figures, we plot the transition probability $P^{\hat{p}}_{01}(t)$ on the y-axis, as a function of time, t on the x-axis. We also compute 95% confidence intervals for an estimated transition probability value $\hat{p}$, using standard results for bernoulli random variables as $C.I.(\hat{p}) = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

For instance, Figure 5.5a shows a group of beneficiaries displaying a flat (time-invariant) transition probability curve, indicating that their behavior could be approximated well using a Markov model. Figure 5.5c on the other hand shows beneficiaries with transition probabilities that decay with time and Figure 5.5b represents beneficiaries with a similar such unique pattern.



(a) Markov behavior **(b)** Unique pattern **(c)** Downward pattern **(d)** Upward Behavior

Figure 5.5: Distinct behavior patterns among beneficiaries are unearthed from data. (a): Some beneficiaries show decaying transition probabilities (b) Some beneficiaries display unique behavior characteristics. (c) Some beneficiaries show fixed probabilities, indicating a Markov model would be a good fit.

5.7.2 Real Data Evaluation of Simulation Testbed

In this section, we empirically evaluate our proposed simulation technique pitting it against the existing simulation technique for RMABs. We again test our approach on the real-world data from the maternal health application domain.

Predictive accuracy of simulator

The goal of this subsection is to evaluate the accuracy of our proposed simulation technique in predicting the engagement trajectories of new beneficiary cohorts, using the behavior patterns learnt from observing the trajectories of beneficiaries in the training data. We establish via two experiments, that our proposed simulator does a better job at simulating the engagement outcome than the existing Markov simulator. We test the following methods: (1) **Markov Sim** denotes the traditional Markov simulator that assumed an MDP model for each agent in simulation. (2) **MMSim** is our proposed simulator that works with temporally changing transition probabilities. (3) **Actual** denotes the actual engagement numbers observed in the real-world experiment mentioned previously.

Apportioning Improved Prediction Quality to Simulator

In this experiment, the goal is to compare and evaluate the simulation outputs of both simulators, given perfect information about the transition pattern clusters of all beneficiaries. The key idea behind this setup is to take the performance of the predictive model supplying the cluster predictions out of the equation. This allows us to apportion the contribution of the simulation technique, predicting the engagement behavior for a known class label. We divide the available data randomly, into a 80-20 train-test split as follows:

Training Data: We assume both simulators have access to the training data $\mathcal{D}_{train}$, consisting of N_{train} beneficiaries, where for each beneficiary, there is a trajectory of states $\{s_t\}\ \forall\ t\ \in\ [L]$ and a sequence of actions chosen $\{a_t\}\ \forall\ t \in [(L-1)]$ recorded. These state trajectories would later be used to perform clustering and assign cluster labels to each beneficiary.

Test Data: The test data $\mathcal{D}_{test}$, similarly consist of N_{test} beneficiaries each with their own state and action sequences, $\{s_t\}$ and $\{a_t\}$ respectively. Each simulator is fed with only the action sequences $\{a_t\}$ as an input, while the true state sequences $\{s_t\}$ are kept hidden. To sidestep the predictive model performance in predicting the cluster assignments of the test set beneficiaries, in this experiment, we also supply the true class labels $\bar{c}\ \in\ [C]^{N_{test}}$ of the test beneficiaries, to the simulator in both simulation methods.

Results: Figure 5.6a compares the total engagements predicted by both simulators each week, against the backdrop of actual numbers observed in the real-world experiment. Figure shows the output for a single simulation

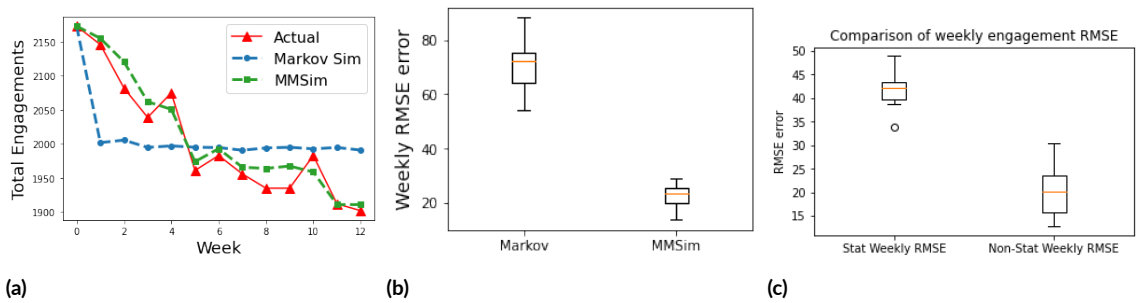


Figure 5.6: (a) MMSim matches the real numbers more closely as compared to the traditional MDP-based simulator, despite even knowing the behavior categories of beneficiaries. (b)MMSim shows much lower Root Mean Square Error (RMSE) in the number of weekly engagements. (c) MMSim still achieves lower RMSE than the Markov simulator, even after including the predictive model in the evaluation.

instance. The closer the simulated output is to the actual output, the better. We repeat the experiment over 30 random instances of the simulation (random seeds) and formalize the error measurement over these as explained below.

In Figure 5.6b, we measure and compare the RMSE error in weekly engagements. This is measured as:

$$\varepsilon_{RMSE} := \sqrt{\frac{\sum_{t=1}^{t=L} \left(\sum_{i \in N_{test}} \left(\hat{s}_t(i) - s_t(i) \right) \right)^2}{L}}$$

We argue that this error metric is more useful at reporting the true picture than an error metric comparing simply the total engagements, because it is possible for the latter to be very low by predicting the correct total engagements, despite having a large error in weekly engagement predictions. In Figure 5.6b, we find that our simulator produces significantly lower ε_{RMSE} than the Markov simulator.

Integrating Cluster Prediction in Evaluation

In this setup, we aim to test the simulator performance while also including the cluster prediction model (φ) in the pipeline. We retain the exact train-test split and training data design as the previous experiment. From the test set, we now withhold information about the true class labels, $\bar{c}$.

We use a Random Forest model to implement φ, i.e. to infer the cluster assignments for new beneficiaries with unknown behavior patterns, by utilizing their known demographic feature information. This problem fits within the standard supervised learning framework and can admit any available machine learning model for φ.

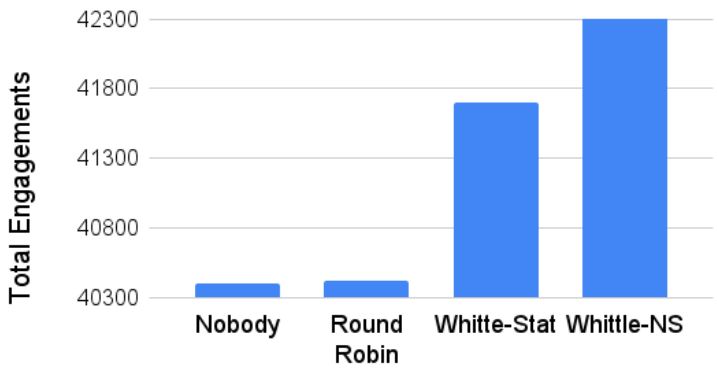


Figure 5.7: Planning Evaluation: Evaluated on planning performance alone, given full knowledge of the transition function, Whittle-NS outperforms other baselines.

Figure 5.6c shows the predictive error measured for 30 random instances of the simulation. Through the lower error statistics, we conclude that our new simulator performs better than the Markov simulator.

5.7.3 Evaluation of Planning Algorithm

In this section we evaluate our solution against other existing baselines when the non-stationary transition functions are perfectly known. We use the MMSim testbed to simulate test policies using transition parameters drawn from the maternal health program data. We consider the following baselines: (1) **Nobody** places no service call interventions. (2) **Round Robin** selects beneficiaries for service calls in a cyclic fashion in a set order. (3) **Whittle-Stat** is the Whittle index solution that models the beneficiary behavior as a stationary MDP. (4) **Whittle-NS** is our proposed solution using the RMAB-NS model.

We run the evaluation for a setup identical to the original experiment used to collect the data [102]. Specifically, our simulation consists of $N = 23003$ beneficiaries with an intervention budget of $k = 450$ service calls per week for a period of $L = 13$ weeks. We measure the performance of an algorithm in terms of the total engagements summed up across all beneficiaries over the L timesteps. All results are averaged over 30 independent runs of the simulation.

Figure 5.7 shows that our algorithm outperforms the other competitor algorithms. The intervention benefit achieved by Whittle-NS in improving engagement over the control ('Nobody') group is 55% higher than the intervention benefit of the best stationary baseline.

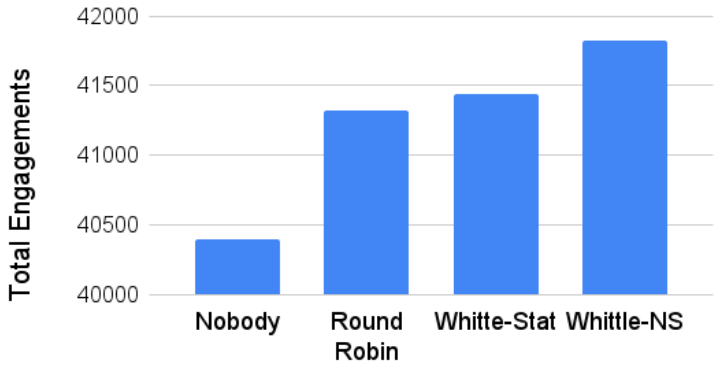


Figure 5.8: Full Pipeline Evaluation: RMAB-NS solution outperforms stationary baselines, notwithstanding the added challenge of inferring more complex model parameters.

5.7.4 Real Data: Evaluation of Full Pipeline

Having seen promising results from the planning component, we next test our model with the full pipeline involving the added challenge of inferring the unknown non-stationary transition parameters.

To enlarge the performance difference between policies for a clearer comparison, we ramp up the budget level used in this experiment, simulating $k = 1350$ service calls per week.

Figure 5.8 plots the weekly engagement numbers on the y-axis resulting from different intervention policies with the week numbers on x-axis. We see that by end of week-12, our proposed Whittle-NS policy improves over existing methods and leads to 367 additional engagements compared to the state-of-the-art Whittle index policy for the stationary RMAB model.

We show that Whittle-NS outperforms stationary baselines despite the added challenge of learning richer transition parameters. The intuition is that owing to a limited number of unique behavior patterns among beneficiaries, both models — the stationary RMAB or the non-stationary RMAB-NS— perform similarly in terms of predicting behavior patterns. However, because the RMAB-NS model encodes much richer information within each behavior pattern, it is able to leverage the same for planning better actions.

5.7.5 Synthetic Domains: Evaluation of Fast Interpolation Algorithm

In this subsection, we test out our fast interpolation algorithm on synthetically generated datasets in which the transition parameters of agents vary linearly with time. We sample a linear function $\mathcal{P}(t)$ at random, and measure

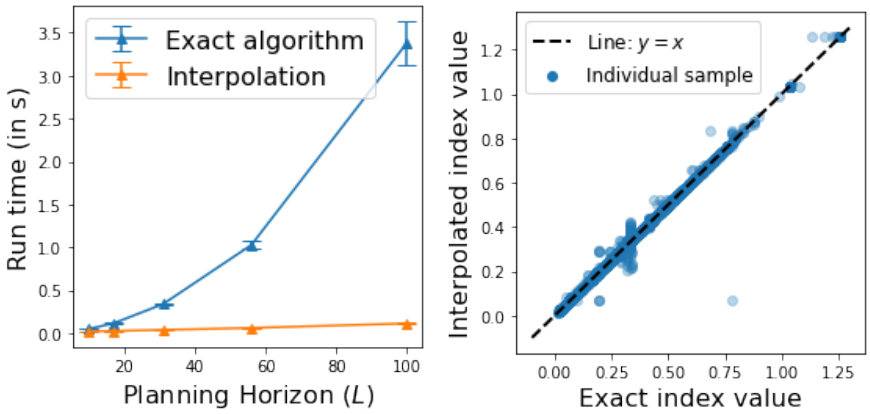


Figure 5.9: (a) Runtime comparison figure. Our linear interpolation algorithm brings a $30\times$ speedup for $L = 100$. (b) Performance evaluation figure: the speed-up comes with a marginal error in estimation of Whittle index .

the exact and interpolated whittle indices for horizon values of upto $L = 50$. We repeat the same process for 30 independently sampled transition matrices.

Figure 5.9(a) compares the run-time involved for planning using the exact Whittle-NS algorithm and for the interpolation algorithm. We see that interpolation dramatically speeds up index computation achieving a speedup of $30\times$ for $L = 100$. Figure 5.9(b) shows a scatter plot with the exact and interpolated values on the x-axis and y-axis respectively. We see that the interpolation algorithm unlocks this speedup while estimating index values nearly as well as the exact Whittle-NS values.

5.8 Conclusion

In this work, we focus on Restless Bandits with non-stationary transition probabilities. We show from real-world data that such a model can be useful as real-world agents may not conform to the MDP assumption of the standard RMAB framework. We propose a Whittle index based solution technique for time varying transition parameters and derive conditions on performance guarantees of the algorithm. We also present an approximate, but faster interpolation algorithm that achieves a $30\times$ speedup for a planning horizon of 100. Additionally, we also present a technique to infer the richer set of transition parameters in the real world, and show via evaluations on both real-wold as well as synthetic data that our proposed RMAB-NS approaches outperform the stationary baselines across the board.

6

Deployed "SAHELI" for Increasing Impact of Mobile Health Programs

6.1 Introduction

Mobile health (mHealth) programs, that leverage the widespread use of cellphones, are a crucial resource for bridging information inequities for underserved and marginalized communities in the global south [153,50], especially in areas such as public health and social services where access to authoritative information is unevenly distributed. Many non-governmental organizations (NGOs) periodically send automated voice messages to improve health outcomes of beneficiaries. However, in spite of high adoption, adherence is a key challenge in public

health information programs [10,65,40,110]. NGOs often employ live service calls made by health workers to boost engagement via encouragement or through logistic changes requested by beneficiaries. However, given the comparatively large number of potential beneficiaries, it is important to maximally utilize the limited availability of health workers, and thus it is crucial to identify the best recipients for such service calls.

Scale of ARMMAN	
States Covered in India	19
Partner NGOs	40
Partner Hospitals	97
Health Workers Trained	235K
Beneficiaries	27.2M

Figure 6.1: A beneficiary receiving preventive health information

While AI models can help health workers in optimizing their service calls, deploying these models in the context of mHealth programs for underserved communities presents unique challenges. First, available data is sparse and skewed (because data is necessarily limited from small numbers of service calls). Second, NGOs are constrained by a very limited compute budget. Third, responsible deployment of the AI models is particularly important in such settings.

In this chapter, we show how we address these research challenges in our deployed AI model – a deployed Restless Multi-Armed Bandits (RMAB) model for public health – together with our NGO partner ARMMAN[9] to help improve the quality of service of their mHealth program focusing on maternal and child care. India suffers from high maternal and neonatal mortality rates [107,169], and ARMMAN[9] runs one of the largest mHealth programs in this domain in India. Our system, Saheli (System for Allocating Healthcare-resources Efficiently given Limited Interventions), is the result of deep partnership of an interdisciplinary team of researchers. Saheli (meaning 'female friend' in Hindi) is designed to assist, rather than substitute, health workers in their normal workflow. The key contributions of deployed Saheli are:

- Saheli includes the first deployed application of RMABs for public health, and it is continuously in use by our partner NGO ARMMAN.
- A key novelty of the deployment is that it both predicts RMAB model parameters and computes optimal policies; in contrast with most past research that has focused on computing optimal policies. To that

end, we provide an improved and robust machine learning prediction framework by performing model selection and evaluation of real-world RMAB systems.

- We deployed Saheli on cloud infrastructure with an emphasis on frugality throughout the end-to-end pipeline given the resource constraints of the NGO partner.
- We present Responsible AI practices to address ethical considerations for deploying an AI system for impact in underserved communities, particularly in this non-western context.

Saheli has been developed as a platform, with the ability to be scaled to more NGOs in more domains. Our source code and data dictionary are available on Github*.

6.2 Related Work

While several works in the healthcare domain have studied patient adherence for diseases like HIV [154], cardiac problems [144,33], and tuberculosis [82,127], these largely focus on building machine learning classifiers to predict future adherence to prescribed medication. With such models, the pool of beneficiaries flagged as 'high-risk' can itself be very large. Furthermore, the one-shot predictions of these models fail to capture the sequential decision making aspect of the problem. Other approaches that consider sequential decision making challenges, such as Pollack et al. [128], Liao et al. [92], Brisimi et al. [25] adopt reinforcement learning techniques to build personalized health monitors that can send timely notifications or activity suggestions to users. However, these models assume notifications can be sent at will, and as such, do not address the challenge of limited service call resources.

Alternatively, RMABs have seen significant theoretical investigation, motivated by resource allocation challenges, such as in anti-poaching patrols [132], multi-channel communication [95], sensor monitoring and machine maintenance tasks [45]. While they provide important contributions, none of these works have seen a real-world deployment, and most have not been field tested.

Key reasons for the lack of RMAB deployment are their significant computational and data requirements. For example, just the optimization problem of computing the optimal allocation π, while assuming the transition parameters $\mathcal{P}$ are available is already known to be PSPACE-hard [125]. Furthermore, in the real world, these transition

*https://github.com/armman-projects/SAHELI

parameters are not just unknown but also hard to infer for real beneficiaries enrolling with ARMMAN and other similar health programs, as they come with no historical transition data. Despite such difficulties, our work is the first to deploy RMABs in tackling a real-world maternal healthcare task via frugal design choices discussed below.

6.3 Problem Introduction

ARMMAN is a non-governmental nonprofit organization based in India, focused on improving maternal and child health outcomes among underserved and underprivileged communities[9]. Their flagship program, 'mMitra', is a mHealth service that aims to leverage the extensive cellphone penetration in India to send out critical preventive health information to expectant or new mothers via automated voice messages. A large fraction ($\sim$ 90%) of mothers in the mMitra program are below the World Bank international poverty line[168]. Despite the acute economic disadvantages faced by these mothers, such automated voice messages prove to be a feasible mode of information dissemination at scale, thanks to the wide accessibility of low-cost phones.

After enrollment into the mMitra mHealth program, beneficiaries receive 1-2 minute voice messages with health information according to beneficiary's gestational age or age of the infant. Unfortunately, despite the proven effectiveness of this information program in improving maternal health outcomes, ARMMAN often sees dwindling engagement rates among beneficiaries, including frequent dropouts. Around 22% of beneficiaries dropout of the program after just 3 months. To counter this issue, ARMMAN leverages health workers that place live service calls (phone calls) to a limited number beneficiaries on a weekly basis to encourage beneficiaries' participation, address requests/ complaints, and attempt to prevent engagement drops. This raises the key question of deciding which beneficiaries to pick for live service call in order to improve engagement rates among the beneficiaries.

6.4 Restless Multi-Armed Bandits (RMAB)

The Restless Multi-Armed Bandits (RMABs) model was first introduced by Whittle[163] to address limited resource allocation problems, but has not received much attention in terms of real-world deployments. An RMAB consists of a set of N arms, where each arm is associated with a two-action MDP[131]. An MDP $\{\mathcal{S}, \mathcal{A}, r, P\}$ consists of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$, and a transition

function P, where $P^{\alpha}_{s,s'}$ is the probability of transitioning from state s to s' when action α is chosen. The reward function in our set up is given as $r(s, \alpha, s') = s'$. An MDP policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ maps to the choice of action to take at each state. The long-term discounted reward for a policy π, starting from state $s_0 = s$ is defined as $R^{\pi}_{\gamma}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}) | s_0 = s\right]$ where $s_{t+1} \sim P^{\pi(s_t)}_{s_t, s_{t+1}}$ and $\gamma \in [0, 1)$ is the discount factor. The total reward in the RMAB is defined as the sum of the total rewards accrued by individual arms of the RMAB.

In the setup we consider, each arm of the RMAB models a beneficiary enrolled with ARMMAN, who can be in one of two states $\mathcal{S} = \{0, 1\}$ (corresponding to 'Not Engaging (NE)' and 'Engaging (E)' respectively). Engagement in our setup was defined in consultation with the subject matter experts at ARMMAN: we define a beneficiary as engaged when she listens to at least one call in a week for more than 30 seconds. The action space for each arm consists of two actions, $\mathcal{A} = \{0, 1\}$, where $1(0)$, typically called the active (passive) action, refers to selecting (not selecting) the beneficiary for the live service call. Beneficiaries may transition from say their E state to NE state (or other transitions) from one week to the next week based on their transition probabilities defined on passive or active actions. The planner's goal is to select actions on arms (deliver live service calls) so as to maximize the total reward, i.e. number of beneficiaries in the engaged state, accrued by the RMAB. However, the budget constraint demands that the planner can choose no more than k arms ($k \ll N$) for the active action at any given timestep, i.e., no more than k live service calls per week.

The dominant technique for solving RMABs uses the Whittle Index heuristic [163], which is shown to have asymptotic optimality under some conditions [161], and to provide excellent performance in practice [132]. Whittle indexes are formulated using the idea of passive subsidy, and informally rank arms so as to choose the top k, based on how attractive it is for a planner to activate each arm. For computing Whittle index, we use binary search algorithm from Qian et al. [132]

Previous Study:

Our previous study conducted in April 2021 [103] is the first to present real-world service quality improvement using RMABs in the context of mMitra program. This study tested an RMAB-based policy against two baselines of interest, and showed RMAB outperforming its competitors. The study spanned 7 weeks and included 23,003 real-world beneficiaries who were distributed in three groups corresponding to the RMAB policy, round robin

(RR) and current standard of care (CSOC). Whereas RR corresponds to a non-AI heuristic for systematically calling beneficiaries, CSOC did not call any individuals. The results from this pilot study are shown in Table 6.1.

Improvements	**RMAB over CSOC**	**RMAB over RR**	**RR over CSOC**
% reduction in total beneficiary engagement drops	32.0%	28.3%	5.2%
p-value	0.044	0.098	0.740

Table 6.1: RMABs demonstrate statistically significant superior performance when compared against other non-AI approaches, namely current standard of care (CSOC) and round robin (RR), as showed by Mate et al. [103].

The pilot results demonstrated that the RMAB method cuts $\sim$ 30% of the beneficiary engagement drops experienced by the other groups. Furthermore, whereas RMAB achieves statistically significant improvement against CSOC ($p < 0.05$) and RR ($p < 0.1$), RR fails to achieve any statistically significant improvement over CSOC. This key result forms the basis of relying on RMAB-based strategy over other non-AI strategies as a basis of Saheli. In this chapter, we describe the journey from this initial study to the final deployment. Whereas we use the same overall RMAB learning and optimization approach, we made multiple changes to provide significant enhancements that reduce data anomalies and improve computational performance of this RMAB-based strategy. Additionally, our deployed cloud application now automates the data exchange process with the NGO's systems while requiring minimal compute resources to be feasibly handled by the NGO. We now describe the end-to-end Saheli system.

6.5 Deploying Saheli

We now introduce Saheli and its architecture. We begin by discussing the different components, and follow that up with the description of the AI pipeline. We then discuss the frugal design choices – both in modeling and infrastructure – that were required to finalize the deployment.

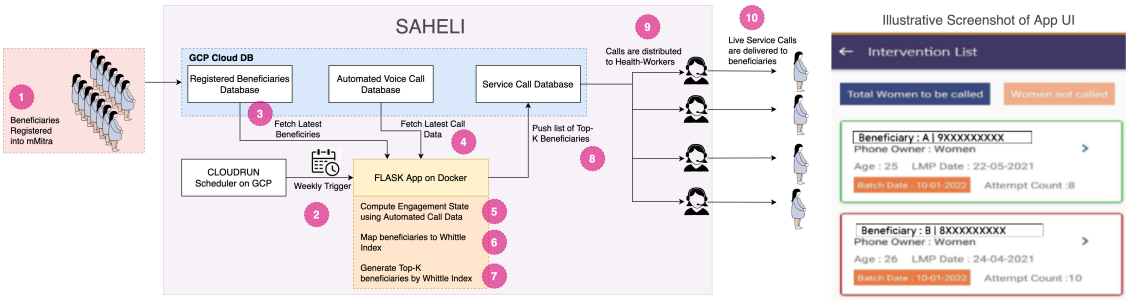


Figure 6.2: Pipeline of Deployed System. Beneficiary information on app UI is available only to the health worker in charge.

6.5.1 System Architecture

We first describe all the interactions within Saheli's ecosystem (refer Figure 6.2). The health workers in the field periodically register beneficiaries through door-to-door visits or at the hospitals (step 1). The socio-demographic data such as age, language, income range, as well as the information on gestational age is then entered into the database maintained by ARMMAN (step 3). Automated voice messages tailored to the beneficiaries' gestation age are sent with the help of a telecommunication provider (step 4). The meta-data of the outcome such as duration of the call, failure reason etc, is also pushed to ARMMAN's database . As beneficiaries' engagement with the voice messages diminishes over time, live service calls are made by ARMMAN to encourage beneficiaries to engage with the program (step 10). However due to limited resources on the NGO's side, only a limited number of live service calls can be made each week. The AI pipeline predicts which beneficiaries would benefit most from receiving a service call in any given week. This list of beneficiaries is then generated at the start of each week and distributed across health workers in an automated fashion as shown on Figure 6.2 in steps through 2-9.

The AI pipeline (described in the next section) for a dynamically growing population is deployed on infrastructure hosted on Google Cloud Platform (GCP). The AI pipeline is wrapped as an application using Flask, which is containerized using Docker. The docker image is created to contain the requisite code scripts for the AI pipeline with apt environment requirements. Our default GCP container settings are to use 6 vCPUs and 16GiB memory. A weekly scheduler job on GCP triggers the Flask application, which then generates the list of beneficiaries.

Step 8 in Figure 6.2 shows the generation of the list of beneficiaries that should be intervened in the given week using the AI pipeline. This list is ingested in ARMMAN's cloud databases, which serve as the back-end of a

client mobile application (screenshot provided in Figure 6.2) used by the health workers. This client application randomly distributes the list of scheduled service calls among health workers based on their weekly availability. An illustrative screenshot (not real beneficiary) is also shown in Figure 6.2. The health worker sees a list of beneficiaries that he/she can call, along with certain features like number of call attempts. They can also click on a particular beneficiary and see more information about the beneficiary and past calls with them (not shown). The calls are made through the week with a maximum of 3 call attempts to the same beneficiary. All the beneficiaries in the generated list receive the aforesaid service calls. The model is currently providing services to beneficiaries enrolling at an average rate of $20K$ beneficiaries per month with a budget of 1000 calls per week.

Saheli streamlines the entire deployment workflow in a singular pipeline, and automates its orchestration and execution, making this process computationally efficient, cost-effective, and easy to debug. As more beneficiaries get enrolled periodically, the beneficiary cohort in the application can now be updated automatically.

Health workers can then make the calls (step 10 in Figure 6.2) to these beneficiaries motivating them to listen to the voice messages and address any logistic issues (e.g. time slots, language of communication, and others) that might be affecting their engagement. As we show later in the chapter, motivating the beneficiaries is key to driving adherence. However, it bears repeating that given the limited availability of the health workers, they can only make a limited number of calls. In our AI pipeline we focused on identifying the right set of beneficiaries to call, and not on automating the contents of the service call. *This is a key design choice in Saheli*: we thus complement the human-to-human engagement between the health worker and the beneficiary, and together they contribute towards aiding a particular beneficiary and driving higher engagement with the mHealth program. This model of working together with the health workers embodies ARMMAN's core 'tech plus touch' philosophy[9] and is essential to our successful outcomes.

6.5.2 Pipeline Description

This section describes the modules in the AI pipeline for both the offline model training and the online model execution. The offline model creation begins with the processing of the training data (i.e. historic data from past mHealth studies), clustering of processed data, and the RMAB modeling per cluster. The transition probabilities and the Whittle indexes are then learned per cluster. Additionally, a mapping from socio-demographic features

of a beneficiary to a cluster is also learned offline. This mapping is used to treat a new beneficiary during model execution – transition probabilities and Whittle index values for the new beneficiary are given by the corresponding values of the beneficiary's mapped cluster. These individual modules are now described. For data privacy reasons, the data pipeline only uses anonymized data and no personally identifiable information (PII) is made available to the AI models.

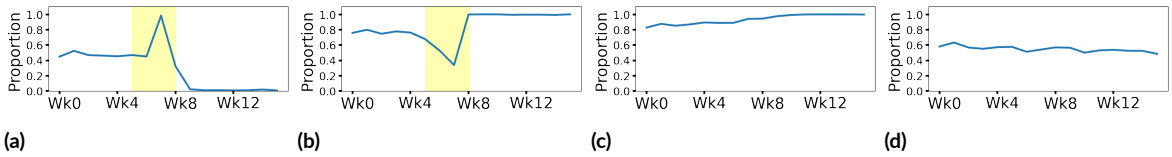


Figure 6.3: Figures (a) and (b) show anomalous engagement behavior while figures (c) and (d) are genuine behaviors. The y-axis shows the proportion of cluster-population in engaging state.

Data Processing:

We train the model on a dataset obtained from historic data collected by ARMMAN, consisting of demographic features and listenership patterns. However, during the pre-deployment trials, we observed some anomalous engagement behaviors – the engagement behavior for some beneficiaries was extremely spiky and unexpected. Figures 6.3(a) and (b), shows two such anomalous groups with a clear peak and dip contrasted with groups having genuine engagement behavior. Upon investigation we found that this spiky behavior resulted from unanticipated real-world events like network outages.

We detect and exclude such anomalies from Saheli's data training pipeline. We first group beneficiaries based on their passive transition probabilities. For grouped beneficiaries, we then obtain a running mean of their engagement over time where the mean is calculated over a window of 3 weeks. We filter out all groups with more than 20% change in running mean engagement within a week. Figures 6.3(c) and (d) show two groups that don't exhibit anomalous behavior and are maintained in the data pipeline.

Additionally, further discussions with ARMMAN pointed out long-term engagement issues in some beneficiaries, such as the registration of a wrong or out-of-service phone number, or the beneficiary not being pregnant. Live service calls in these cases are not productive. Thus, as a pre-processing step, we do not consider beneficiaries who have not listened to any automated voice calls in the past 6 weeks.

Clustering:

We face a data scarcity and skew challenge in our domain. Specifically, our training dataset comprises of beneficiaries from our own past studies where intervention data is available for only a limited set of these beneficiaries. Thus, to define the parameters of the RMAB model, we cluster beneficiaries as an effective way of addressing data scarcity. We cluster the beneficiaries per their transition behaviors for passive actions using *k-means* clustering. We obtain transition probabilities for each of these clusters by aggregating their transitions as a whole.

However, the optimal number of clusters is a design choice not readily addressed by *k-means*. We experimented with the number of clusters ranging from 1 to 100, and looked at the *distortion* metric. Distortion is the sum of squared distances of each point from its corresponding centroid, where smaller distortion implies better clustering. We plot the distortion values for multiple number of clusters and find 20 to be the ideal choice using elbow-method. The results are shown in Figure 6.4a where the x-axis is the number of clusters and the y-axis is the distortion value. This has the added advantage of offering computational frugality.

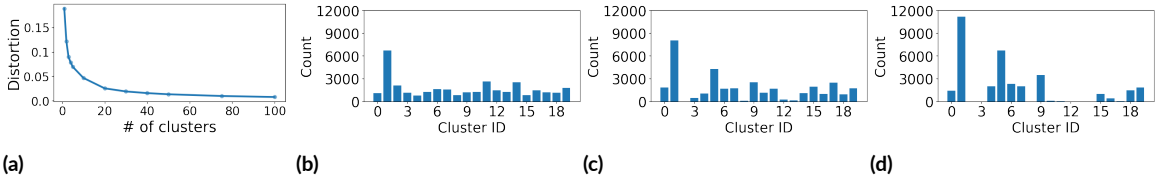


Figure 6.4: Figure (a) shows elbow plot with distortion for varying number of clusters. Figures (b), (c), and (d) show the distribution of predicted clusters using the Feature Only (FO), Feature and Warm-up (FW), and Warm-up Only (WO) mapping functions.

Mapping Features to Clusters:

When a new beneficiary enrolls into the system, the system only knows about their demographic data. We therefore need to learn a mapping of a beneficiary's socio-demographic features to clusters, to enable inferring transition probabilities and Whittle indexes for newly enrolled beneficiaries (step 6 in Figure 6.2). We experimented with different mapping functions to identify the best one: Features Only (FO) mapping - beneficiaries' socio-demographic features only; Warm-up Only (WO) mapping - transition probabilities computed from warm-up period (first 6 weeks post enrollment); and lastly Feature and Warm-up (FW) mapping - using a combination of the above two.

We compute Mean Absolute Error between predicted and ground truth passive transition probabilities as a performance metric and found them as [0.40, 0.37, 0.38] for FO, FW, and WO strategies respectively. In addition to MAE, we plot the distribution of beneficiaries predicted in different clusters (refer Figures 4.3(b), (c) and (d)). Having a sparse cluster distribution is undesirable since large clusters lowers the granularity of Whittle index planning. As an extreme example, if all beneficiaries are mapped to a single cluster, they would all have the same transition probability and thus the same Whittle indexes. Since the cluster size is now much larger than the number of arms to be pulled, the beneficiaries within that cluster would be chosen randomly for receiving service calls, which would degrade the performance.

Thus, to ensure equitable cluster distribution, we computed Entropy and Gini index values for the predicted distribution of number of beneficiaries per cluster. Entropy values came out to be [2.81, 2.56, 2.04] for FO, FW, and WO respectively, and Gini indexes were [0.29, 0.48, 0.57]. Given the error similarities for the three strategies, and higher entropy / lower Gini index implies more equitable clusters, we chose FO as our strategy.

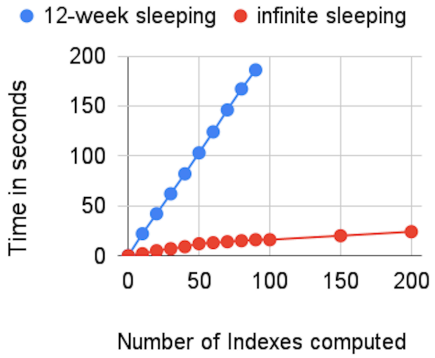


Figure 6.5: Index computation is significantly faster with the infinite sleeping approximation.

RMAB Modeling and Whittle Index Computation:

These transition probabilities per cluster are used to compute Whittle indexes for all beneficiaries, similar to Mate et al. [103], i.e., computing $2 \times k$ unique indexes where k is the number of clusters. There are two Whittle indexes per cluster as beneficiaries may be in the engaging or non-engaging states. Whittle index indicates the benefit of performing an active action on a beneficiary: higher Whittle indexes are chosen to receive service calls (step 7 in Figure 6.2). By mapping beneficiaries to clusters, the Whittle indexes can be pre-computed per cluster at the

beginning of the deployment, thus providing a frugal solution ideal for large scale deployment with minimal resources.

Frequency of Repeated Live Service Calls:

We initially enforced a frequency restriction that required ensuring no beneficiary be called more than once in η+1 weeks (we set $\eta = 3$). Algorithmically, we implement this by appending η sets of dummy 'sleeping states' to the state space that we force the beneficiaries to transition through each time they are called. This augmentation yields a state space of size $2\eta+2$ and a transition matrix of size $(2\eta+2)\times(2\eta+2)$. However, our pilot tests reveal that repeat calls made within just $\eta = 3$ weeks are less effective. For instance, we observed that 30% of 'Non-engaging' beneficiaries converted to 'Engaging' due to the first service call; however this number drops to 20% for repeat calls made just three weeks later. To address this, along with the subject matter experts at ARMMAN, we increased the sleeping period, η, to 12 weeks.

6.5.3 Frugality of System Design

Successful deployments of AI systems like Saheli in social good settings requires conscious focus on frugality across the system design. This is to reduce both the direct costs (e.g. number of calls) and indirect costs (e.g. computational requirements) on our NGO partners. Here are some design choices in Saheli that have led to frugality in its operations:

1. Clustering of beneficiaries allows us to compute transition probabilities and Whittle indexes at a cluster level as opposed at the beneficiary level. Since we use 20 clusters for thousands of beneficiaries, it provides a significant scale-up in performance, while simultaneously reducing data demands for learning RMAB model parameters.

2. As described above, we updated the 'sleeping states' parameter η to 12. However, this increases the Whittle index computation time sharply, owing to a bulky transition matrix of size 26×26. With frugality in mind, we use the insight that a sleeping constraint with large η can be approximated as a permanent sleeping constraint, akin to setting η to $+\infty$, for the purposes of index computation. This is because in index computation, the contribution of reward terms appearing after η timesteps is discounted by a factor of γ^{η} ($\gamma < 1$), which precipitously

diminishes to zero. This simplification compresses the transition matrix to 4×4, and unlocks a $25\times$ speedup in index computation, as shown in Figure 6.5.

3. Lastly, multiple frugal design choices were made in the orchestration of cloud infrastructure. Specifically, we run our services on-demand using a task scheduler on default container settings of 6 vCPUs and 16GiB memory.

6.6 Application Use and Payoff

We now discuss the impact of Saheli on both the beneficiaries as well as the AI community in more detail. Saheli is deployed and in continuous use at ARMMAN. It has already reached 50K beneficiaries, and is on track to reach one million beneficiaries by the end of 2023.

6.6.1 Engagement Results

In order to evaluate the impact of live service calls through Saheli, we track the engagement behavior of a cohort of 5000 beneficiaries for 12 weeks, registered between February 2022 to April 2022. We further filter 2538 beneficiaries with engagement between 10% to 90% as these would benefit the most from live service calls. Additionally, we create a holdout set of beneficiaries registered in the same time period but are not given any live service calls (we obtained ethical approvals before our studies; see section Responsible AI practices for further discussion). We make sure that both the Saheli and holdout groups have equal number of beneficiaries, equal number engaging beneficiaries at the start of experiment, and similar socio-demographic features.

Figure 6.6(a) shows how many engagements did not occur in the holdout group that occurred in the Saheli group, aggregated cumulatively across months. It demonstrates that the Saheli group received significant benefit: the Saheli group has an additional 300 engaging beneficiaries over the holdout group cumulatively at the end of three months.

We also measured the benefit for the Saheli group over the holdout group in terms of time spent listening to mMitra voice calls. More time spent implies more content exposure for our beneficiary population, as well as better adherence with the mHealth program. In particular, by the end of month 3, the Saheli group had listened to 60, 000 seconds *more of content than the holdout group* (Figure 6.6(b)). Overall, at the end of three

months, SAHELI prevented **drop in engagements by** 30.5% with an **additional content exposure of** 46.4% in comparison to the holdout group. This analysis demonstrates SAHELI's success in achieving our core objectives of improving information dissemination.

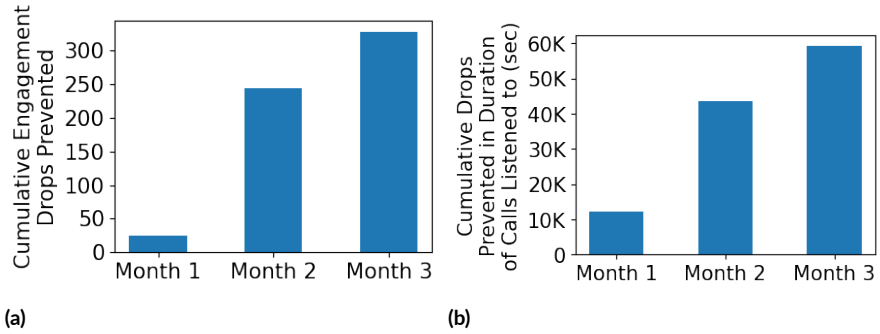


Figure 6.6: (a) Prevention in drop in engagement (cumulative) (b) Increased time spent listening to calls (cumulative)

6.6.2 IMPACT OF LIVE SERVICE CALLS

We performed a qualitative study to understand human-AI collaboration due to the AI system. We conducted a total of 24 interviews, 2 focus group discussions, and approximately 90 hours of observation. We found that healthcare workers engaged positively with targeted predictions through the AI system that integrated into their day-day workflows seamlessly. It helped them improve the engagement of beneficiaries, provided an opportunity to support them in their care journeys and understand their needs.

We also investigated the reasons for why live service calls helped improve engagement with ARMMAN's mMitra mHealth program from the perspective of the beneficiary. Specifically, we conducted a follow-up study with a sample of beneficiaries who were given live service calls one year ago. We could successfully reach out to 306 beneficiaries, out of which 134 recalled the details of the service call from a year ago. Table 6.2 shows the responses to our follow-up study by these 134 beneficiaries. Particularly, 50.75% beneficiaries engaged more with mMitra calls after getting more information about the program. The service calls also helped improve listenership by making logistical updates such as updating delivery date (9.7%), changing time slot of receiving the call (8.21%) or updating the phone number (2.99%).

Did the call help you to listen to the mMitra calls more regularly?	# of Beneficiaries	% of Beneficiaries
Yes, after getting more information about mMitra, I am listening to the calls more regularly	68 (in 134)	50.75%
Not really	30	22.39%
Yes, after updating my delivery date, I was able to get the right information	13	9.7%
Yes, after changing time slot, I am able to listen to the calls more regularly	11	8.21%
Have not asked my wife	4	2.99%
Yes, after changing the number, I am able to listen to the calls more regularly	4	2.99%
Any other	4	2.99%

Table 6.2: Follow-up study responses

6.6.3 Fairness of the RMAB model

Model fairness in non-western contexts has not received much attention in the literature [138]. Responsible AI principles of the Government of India's NITI AAYOG [120] for example, requires non-discrimination based on sensitive markers like caste and religion. These sensitive data are specifically not collected by ARMMAN for mMitra, thereby, making it inaccessible to Saheli's AI models. We worked with public health and field experts to evaluate other indicators such as education, and income levels that signify markers of socio-economic marginalization. ARMMAN's goals for SAHELI are to favor beneficiaries of lower income and lower education levels for service calls. We conducted a post-hoc analysis of the deployment to evaluate if SAHELI indeed met such preferences.

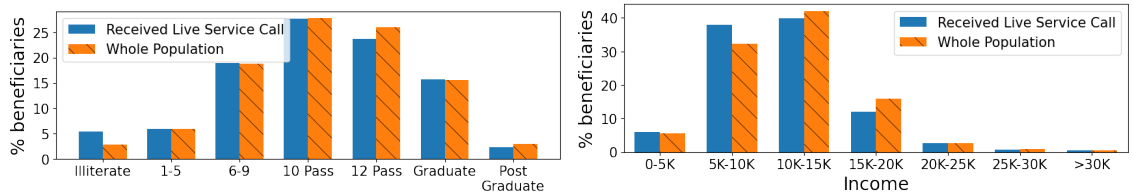


Figure 6.7: Distribution of (a) education (highest education received) and (b) income (monthly family income in Indian Rupees) across cohort that received service call and the whole population.

Figure 6.7(a) shows the distribution of beneficiaries aggregated across Saheli's enrollments split into different education levels in India. We compare those who were chosen for live service calls by Saheli versus the enrolled

population. The x-axis portrays the education levels; for instance grade 1-5 represents primary school, grade 6-9 middle school, 10th pass junior high, and 12th pass represents senior high school. The y-axis is the % of beneficiaries per education category. For instance, Saheli calls 5.5% of beneficiaries who had no formal education (illiterate), whereas this group was 2.8% of the overall enrolled population.

We did a similar analysis split by income as depicted in Figure 6.7(b). The x-axis contains buckets of average monthly income of the beneficiary household in Indian Rupees, and the y-axis denotes the % of beneficiaries in that income category. As an example, the category '5K-10K' contains around 30% of the beneficiaries in the population, and almost 40% of the beneficiaries who received a service call.

Both these figures show that Saheli favors the beneficiaries in the 'illiterate' education category and in the '5K-10K' income category. This distribution is in line with ARMMAN's goals – Saheli favors beneficiaries of lower income and lower education levels for service calls.

6.6.4 Enabling New Research

From identifying the right problem to solve, to creating an AI solution, testing it in pilot, iterating on learnings and finally, establishing an end-to-end integrated system, we made our journey to this deployment. With this, we provide other AI researchers an important case study to take an AI model from the lab out on the field. In our pursuit of deployment of Saheli, we uncover several research challenges, e.g., we overcame the challenges of data scarcity and frugal design. This hopefully inspires additional research in robust and computationally efficient approaches for RMABs and other AI applications for mHealth.

6.7 Responsible AI Practices

We recognize the responsibility associated with deploying real-world AI systems that impacts underserved communities. In our approach, we have iteratively designed, developed and deployed the system in constant coordination with an interdisciplinary team comprised of ARMMAN's field staff, social work researchers, public health researchers and ethical experts. Along with seeking ethical approvals through review boards at Google and ARMMAN, we have taken additional steps to constantly monitor and mitigate the risks associated with Saheli by abiding with AI principles at Google [48] as well as key policy making bodies in India such as the NITI

AAYOG [120]. Our success draws attention to the practices around responsible AI including ethics, fairness and accountability in the non-western context [138] where Saheli is deployed. We now discuss three of the core Responsible AI principles that impacted the design of Saheli.

Socially beneficial: The intent of this work is to bring the power of AI in service to some of the most marginalized communities in the global south. The challenges faced by our team were limited resources in every dimension – limited data on the beneficiaries, limited compute available to the NGO, and limited health workers to make the outreach calls. Thus, we had to develop new algorithms that were not data hungry, and were bounded in their computational requirements. To that affect, Saheli is the first large-scale deployment of RMABs for public health.

Avoid reinforcing unfair bias: As discussed in the previous section, we have undertaken extensive analysis to study model's fair treatment of beneficiaries.

Incorporate privacy design principles: We take significant measures to ensure participant consent is understood and recorded in a language of the community's choice at each stage of the program. Data stewardship resides in the hands of the NGO, and only the NGO is allowed to share data. This dataset will never be used by Google for any commercial purposes. In this dataset, sensitive features such as caste and religion are never collected and stored. Saheli's data pipeline only uses anonymized data and no personally identifiable information (PII) is made available to the AI models. Lastly, domain experts at ARMMAN have been deeply involved in the development and testing of Saheli and have provided continuous input and oversight in data interpretation, data consumption and model design.

6.8 Maintenance

Since SAHELI has been automated end-to-end, there has not been any manual intervention in the run of the system. We have been reviewing the system regularly in collaboration with ARMMAN. Though no updates have been required since deployment, the modular composition of Saheli enables us to make updates to the AI model without affecting other components.

6.9 Lessons Learned

Over the course of one year of our experiments moving from Pilot study to Deployment, we learned several lessons along the way. Most importantly, we learned that even a successful pilot study can't be translated as-is in to a full-scale deployment, and that several considerations are critical for wide-scale adoption of AI tools and scaling up of impact.

Selecting the right problem: There are multitude of problems that require to be solved to address the needs of the underserved communities. In our interactions with ARMMAN, *we realized that we could create the most impact with our techniques by improving the selection of the right beneficiaries for manual intervention*, as opposed to automating the communication with the beneficiary. Our choice of problem is consistent with the 'tech plus touch' philosophy of ARMMAN [9], and ensures that we complement the human expertise of the health worker. This way, each chosen beneficiary continued to have a one-on-one interaction with a health worker, while simultaneously improving the overall engagement with the mHealth program.

Immersion into the real-world problem: We learned that immersing in the working of a NGO and public health infrastructure is critical in understanding the context of the problem. The authors went on multiple field visits to understand the stakeholders involved in the mMitra's workflow. The health workers interact with the beneficiaries across multiple mHealth programs, and thus can speak to the needs and behaviors of the beneficiaries. For instance, upon interacting with these health workers, we understood how telecom outages lead to more anomalous and incomplete data than we had anticipated. We also understood the decreased value in utility of calling the same beneficiary again shortly after a previous call. *These field visits forced us to re-evaluate our assumptions, and led to better data processing and modeling choices*, as discussed in the earlier sections. For instance, after these discussions, we incorporated a new anomaly detection mechanism in our data pipeline, and impacted our choice of horizon (η) in our RMAB model.

Fairness of AI models: AI algorithms and datasets can reflect, reinforce, or reduce unfair biases. It is imperative on AI designers to seek to avoid unfair impacts on people, particularly on underserved and marginalized communities. As discussed in the section on Responsible AI practices, we worked with public health and field experts to demonstrate fairness of our approach. As we mentioned before, 94% of our potential beneficiary population are below WHO's poverty index. *Studying multiple socio-demographic attributes was essential to evaluate*

fairness of our approach. We worked closely with ethics experts, the ARMMAN's ethics team, and Google's ethics teams and extensively validated the fairness of our models.

End-to-end integration testing: In addition to the lessons learned on problem selection and model development, we also ran into several issues in our end-to-end integrated pipeline. On one occasion, we saw poor results because the data schema had evolved in the data storage pipeline at ARMMAN. *Testing of our application required our NGO partner to be equally involved in the validation of Saheli's outputs – as domain experts, they are better equipped to identify counter-intuitive behaviors.* Our experiences uncovering issues in the end-to-end pipeline led to improved communication practices, better documentation and tighter test goals. Social good applications like Saheli has real-world consequences for beneficiaries in underserved communities, and it is critical that there be a real partnership for testing and integration.

6.10 Conclusion

In this chapter, we presented Saheli, the first ever deployment of restless multi armed bandits in the public health domain for allocation of limited resources. Saheli is built on an improved and robust framework that both predicts RMAB parameters and computes optimal policies for it, in contrast with most past research that has only focused on computing optimal policies. It has been built with careful design choices inspired by close interactions with all stakeholders. It incorporates numerous lessons learned by embedding ourselves in the real-world domain. Saheli has been deployed on cloud infrastructure with an emphasis on frugality, and has reached out to 50k beneficiaries so far and aims to reach 1 million by 2023. Furthermore, in this chapter, we also discuss the importance of responsible AI practices in deploying AI systems at scale, especially in the social domain. This work serves as an important case study for AI researchers and NGO communities alike to take ML models from the lab and deploy them in the field.

7

Improved Evaluation of Algorithmic Resource Allocation Policies

7.1 Introduction

We consider a subclass of randomized controlled trials (RCTs) wherein the goal of the trial is to evaluate the efficacy of an algorithmic resource allocation policy. Such policies recommend an allocation (action), commonly utilizing tools such as reinforcement learning[42,148], variations of the multi-armed bandit framework[66], network optimization[166,157], etc. As machine learning becomes increasingly widely applied in socially critical settings, such policies have been used to allocate limited resources in a variety of domains, including campaign optimiza-

tion [87,37], improving maternal healthcare [21], screening for hepatocellular carcinoma [86], monitoring tuberculosis patients [104], etc. Such a policy may rank all individuals within a group, and offer a scarce resource, such as a home visit by a health worker, to the highest-ranked individuals within the group.

RCTs evaluating these resource allocation policies consist of M experimental arms labeled $\{1, \ldots, M\}$, with each arm consisting of N unique, randomly assigned individuals. The policy in each experimental arm prescribes an allocation of resources while respecting some resource constraints. This process may be either single-shot (allocations made once) or sequential (new allocation decisions made adaptively over a series of rounds). Each policy, dictating its own resource allocation strategy, is evaluated at the end of the trial by analyzing the outcomes data from its corresponding experimental arm. The group-level decision making of resource allocation policies creates new challenges for their experimental analysis. In a standard RCT, the aim is to evaluate the treatment effect – there is no resource constraint, so all participants receive treatment – for the average participant [7]. As the outcomes of all participants are independent, analysts can simply compare outcomes in the treatment versus control groups. However, trials of resource allocation policies aim to evaluate the *group* outcome of a set of individuals to whom the policy is applied (some of whom receive the resource, and some who do not). Even when the number of individuals in the trial is large by the standards of a normal RCT, randomized trials of allocation policies can suffer from high variance, leading to noisy and potentially erroneous estimates of the treatment effect.

This high variance stems from two sources. First, the outcomes of individuals within a given arm are correlated because allocation policies typically consider all individuals *jointly* in making an allocation decision (e.g., to respect budget constraints on the total number of individuals who may receive an intervention). This interdependence implies that we only see one independent sample of a policy's performance per RCT, instead of many (in contrast with one sample per participant in a standard RCT). Second, for many individuals the allocation decisions made by different policies being compared may coincide. For instance, no matter the policy applied, many individuals may consistently get screened in, or many may get screened out of receiving a resource. Because these individuals are oblivious to the policy employed and receive identical allocations under different test policies, their outcomes are never truly impacted by the policy employed. Yet their presence generates additional variance in the average outcome of the group due to stochasticity in their outcomes, independent of the policy being evaluated. *To our knowledge, no prior work has proposed strategies to mitigate these sources of variance in RCTs of resource allocation algorithms, despite the increasing use of ML-based policies in socially critical domains.*

An intuitive first attempt at variance reduction would be to share information across arms. For example, we could average out fluctuations in the outcomes of individuals who do not receive an allocation by pooling together such individuals across all of the test policies. Unfortunately, such naive estimators are biased because the distribution of *which* individuals receive a resource (or not) is different across the trial arms – indeed, assessing the comparative efficacy of these distributions is exactly the point of the trial. A more sophisticated strategy would be to exploit overlap between the policies via inverse propensity weighting estimators which reweight participants in one arm to match the actions a different policy would have taken [84,14]. However, such estimators are themselves subject to notoriously high variance, particularly when the overlap between policies is poor [177]. Moreover, as we discuss in Section 7.4.1, propensity scores are impossible to calculate for sequential intervention settings, where unseen states prevent us from evaluating individual-level propensities for a different policy.

Our key contribution fixing these issues is a novel estimator for the treatment effect of resource allocation policies which exploits overlap between arms in a principled manner, is guaranteed to reduce variance, and is applicable to either single-stage or sequential settings. The main idea is to find a subset of individuals with the following property: if we ran a hypothetical trial where the arm assignments of these individuals were swapped, the allocations made for all individuals remain unchanged. Our estimator identifies such sets of individuals with this property and averages the outcomes of all corresponding hypothetical trials (which are observable because no allocation decisions were altered from the original trial). We make the following contributions: (1) we propose this novel estimator and prove that it produces an unbiased estimate of the treatment effect with a guaranteed reduction in variance compared to the standard estimator, implying that it has strictly smaller average error; (2) we show how this estimator can be implemented in a computationally efficient manner for a class of policies encompassing those most commonly used in real-world allocation decisions; (3) we conduct experiments on three domains leveraging synthetic, semi-synthetic, and real-world data. We show the application of our techniques to real-world case study data illustrating its usefulness in solically critical domains. Across the board, our estimator substantially cuts error (by up to $\sim 70\%$) compared to available estimates.

7.2 Problem Formulation

RCT Setup: We consider a set of N individuals. Each has a feature vector $\mathbf{x} := [\mathbf{x}_o, \mathbf{x}_u]$, where $\mathbf{x}_o \in \mathcal{R}^o$ denotes the individual's observable features (such as demographics) while $\mathbf{x}_u \in \mathcal{R}^u$ denotes unobservable characteristics which nevertheless influence outcomes. A resource allocation policy π jointly considers all N individuals, their joint matrix of observable features $\mathbf{X}_o \in \mathcal{R}^{N \times o}$, and prescribes an allocation (action) vector $\mathbf{a} := \pi(\mathbf{X}_o) \in \mathcal{A}^N$ where $\mathcal{A}$ denotes the space of possible actions for each individual. The allocations $\mathbf{a}$ must respect resource constraints specific to the domain. On example is a budget constraint $||\mathbf{a}|| \leq B$ capping the total cost of allocated actions. Each individual then stochastically yields an outcome state $s \in \mathcal{S}$, according to an unknown function $P^*(\mathbf{x}, a, s)$ denoting the probability of the individual receiving outcome s. We use $r(s, a)$ to denote the reward accrued by the policymaker from this outcome. We remark that this notation is equivalent to the standard potential outcomes framework [137] where $r(s, a)$ is the realization of the individual's potential outcome given action a and P governs the joint distribution between $\mathbf{x}$ and the potential outcomes. We adopt the present notation for easy generalization to the sequential setting.

The goal of an RCT is to compare M such resource allocation policies using M randomly constructed experimental arms $C_1, \ldots C_M$ where C_i denotes the set of individuals assigned to the i^{th} arm. We use $\mathfrak{C} := \{C_1, \ldots C_M\}$ to denote this particular assignment of individuals to experimental arms, chosen uniformly at random from $\mathcal{C}$, which denotes the universal set of all possible assignments. At the end of an RCT, the analyst can compare the performance of the M allocation policies by comparing the sum total of rewards accrued by participants within each arm, given as $\text{Eval}(\pi_m) := \sum_{i \in C_m} r\big(s(i), a(i)\big)$.

Sequential RCTs: We also consider RCTs involving a multi-stage resource allocation setup. Such RCTs run for a total of T rounds (the single-stage setting above corresponds to $T = 1$), where allocations $\mathbf{a}_t \in \mathcal{A}^N$ must be made at each timestep $t \in [T]$ subject to potentially time-dependent resource constraints. At each time t, each individual has a state s belonging to state space $\mathcal{S}$. We use $\mathbf{s}_t$ to denote the N individuals' states at timestep t. The policy π may consider the entire history of previous states and actions to generate a new action allocation as $\mathbf{a}_t := \pi(\mathbf{X}_o, \mathbf{s}_{0:t-1}, \mathbf{a}_{1:t-1})$. Similarly, the individual transitions to a new state s_{t+1} according to the probability function $P^*(\mathbf{x}, s_{0:t}, a_{1:t+1}, s_{t+1})$. The state and action trajectories of all N participants are recorded for each of the M experimental arms as matrices: $S_1, \ldots, S_M \in \mathcal{S}^{N \times T+1}$ and $A_1, \ldots, A_M \in \mathcal{A}^{N \times T}$, which facilitates similar

computation of $\text{Eval}(\pi_m) \coloneqq \sum_{i \in C_m} r(S[i], A[i])$ as defined for the single-shot setting.

Index-based policy: We define a specific class of policies – called 'index-based policies' – for which a computationally efficient estimator can be derived. Index-based policies rank individuals according to some scoring rule that determines a prioritization among individuals for allocation of a resource. Such policies encompass most relevant resource allocation policies commonly employed in practice due to their transparency and ease of implementation [156]. Note that this class includes seemingly unlikely candidates that allocate resources to individuals cyclically in a set order or benchmarks such as 'control' groups (details in Appendix E.2). Formally, index-based policies are defined for a binary action space $\mathcal{A} \coloneqq \{0, 1\}$ with a budget constraint allowing at most B individuals to receive the action $a = 1$. The policy computes a time-dependent index $\Upsilon(\mathbf{x}_o, s_{0:t-1}, a_{1:t-1})$ for each individual at each timestep t, based solely on the individual's observable features $\mathbf{x}_o$, trajectory of states s and history of actions a received. The policy π allocates action $a = 1$ to the top B individuals with the largest values of index Υ. At any given time t, we assume the value of Υ is unique for each individual, i.e., the policy induces a total ordering. The key feature of an index-based policy is that Υ is computed independently for each individual based only on their own features and history.

Problem Statement: $\text{Eval}(\pi_m)$ provides a single random sample with which to estimate the *expected* performance of allocation policy π_m, combining the randomness in the assignments $\mathfrak{C} \in \mathcal{C}$ and the randomness in outcomes within each $\mathfrak{C}$, engendered by stochasticity in state transitions. Let $\text{Eval}^*(\pi_m)$ be the expected value of the performance of π_m:

$$\text{Eval}^*(\pi_m) \coloneqq \mathbb{E}_{S_m \sim P^*} \mathbb{E}_{\mathfrak{C} \sim \mathcal{C}} \Big[\text{Eval}(\pi_m) \Big] \tag{7.1}$$

We assume data available from only a single run of an RCT. Our goal is to build an estimator that estimates $\text{Eval}^*(\pi_m)$ accurately from just the single RCT instance. We remark that our results make no distributional assumptions about the set of individuals in the trial, e.g., we do not require them to be IID from some distribution. Rather, we treat the observed individuals as fixed (nonrandom) and propose techniques that use only the randomness in the assignment process of the RCT. The only assumption required is that individuals' state transitions are independent, akin to the standard stable unit treatment values assumption (SUTVA) in causal inference [57]. This is aligned with the growing emphasis on design-based inference in causal inference (c.f. [62,112,1]),

which allows us to formulate methods which require minimal modeling assumptions.

7.3 Related Work

Off-Policy Evaluation: The most closely related previous work is the off-policy evaluation (OPE) literature, where the goal is to use samples collected under some baseline policy to inform the evaluation of a new policy [146,68,159]. OPE makes frequent use of inverse propensity weighting estimators [96,89]; we discuss strategies for constructing such estimators as well as their disadvantages in Section 4.1. To date, the OPE literature has largely focused on individual-level decisions, as opposed to the group-level resource allocation we study. One exception is slate-level OPE, where the policy recommends a ranked list of items to a user [147]. Slate OPE is most similar to our single-step case, while we develop methods that extend to the multi-step setting. Additionally, slate OPE is motivated by unobservable individual-level rewards, while we assume that individual rewards are observable and the challenge for the single-step setting is that policies may be deterministic (preventing us from using their methods).

Individual treatment rules: Recent work in statistics has studied experimental design and analysis for individual treatment rules (ITR), which are similar to our class of index-based policies [64,13]. The crucial difference is that ITRs make decisions independently for each individual, while in our setting the policy considers the group of individuals jointly, which is required for exact enforcement of constraints such as budgets. Our techniques are motivated by the need to reduce variance when policies can only be evaluated at group level.

Cluster-randomization and interference: Our setting is related to a family of RCTs known as cluster-randomized trials (see [54] for an overview). In such trials, treatment is assigned at a group level (e.g., assignment of classrooms within a school instead of students), just as groups of individuals are assigned to policies in our setting. However, cluster-randomized trials are motivated by the potential for spillover effects, where the outcomes of one unit can influence others. By contrast, in our setting the outcomes of individuals are still independent conditioned on the actions of the policy. Accordingly, there is no need to account for potential correlations as in the interference literature; instead, we leverage this structured independence to develop lower-variance estimators.

7.4 Methodology

Our goal is to leverage the overlap in decisions of multiple resource allocation policies to improve our estimate of the reward from deploying each. We start by developing an estimator using inverse propensity weighting – a popular approach typically adopted for such a task – and show how it can be naturally applied to the single-stage setting. However, this natural estimator suffers from two challenges. First, inverse propensity estimators can suffer from notoriously high variance, a phenomenon that we empirically confirm in Section 7.6. Second, the approach breaks down entirely in the multi-stage setting, where (as detailed below) computation of propensities is impossible due to missing data. We resolve these challenges by developing a more stable "assignment permutation" estimator, which applies to both settings and is guaranteed to reduce estimation error.

7.4.1 Propensity Scores Approach

In typical off-policy evaluation settings, inverse propensity weighting (IPW) methods reweight samples according to the probability that observed actions would be taken by a given policy. These methods are not immediately applicable to our problem because we do not assume that policies are randomized – indeed, explicit randomization is rare in policies deployed by real-world governments, health systems, etc. When policies are deterministic, the probability that they would yield an alternate action is precisely zero, leaving standard propensity estimators undefined.

We show how to circumvent this issue in the single-step setting by leveraging an alternate source of randomness: the assignment in the trial itself. Calculated over the randomness in the assignment, each individual has some probability of being assigned a given action, denoted as $Pr_{\mathfrak{C},\pi}[a(i) = a]$ (intuitively, whether they receive a resource depends on who else the policy is comparing them to). Formally, exchanging the order of expectations allows us to write $\text{Eval}^*(\pi)$ as:

$$\sum_{i=1}^{N} Pr[i \in C_m] \sum_{a \in \mathcal{A}} Pr_{\mathfrak{C},\pi}[a(i) = a | i \in C_m] \mathbb{E}_s[r(s,a)]$$

Since the assignment $\mathfrak{C}$ is random, $Pr[i \in C_m] = \frac{1}{M}$. Moreoever, in the inner term, conditioning on $i \in C_m$ leaves the other members of C_m distributed uniformly at random. Accordingly, we can estimate $Pr_{\mathfrak{C},\pi}[a(i) =$

$a|i \in C_m]$ for any policy π by drawing repeated samples of the assignment $\mathfrak{C}$ and running π to reveal whether π would have assigned $a(i) = a$ given the group C_m containing individual i. Let $\hat{p}(i, a|\pi)$ denote the fraction of these samples in which $a(i) = a$. A standard IPW estimator for Eval(π) is given by

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{i\in C_m}\frac{\hat{p}(i,a(i)|\pi)}{\hat{p}(i,a(i)|\pi_m)}r(s(i),a(i)). \tag{7.2}$$

In a sequential setup ($T > 1$), propensity score methods become entirely inapplicable for two reasons. First, standard multi-time step IPW estimators require randomness in the policy, while we assume that policies may be deterministic. We cannot use the alternate approach described above (leveraging randomness in assignments) over multiple time steps, because the marginal probability that individual i receives action $a(i)$ on future steps depends on the state of all other individuals, and we do not have samples of such future states under counterfactual assignments. Second, even if we limited to randomized policies, standard off-policy methods calculate the probability of taking exactly the observed sequence of actions in the observed states. In our case, this requires computing the probability of π selecting the *vector* of actions assigned to each individual, i.e, we have a N-dimensional action space within each time step. Multi-step IPW estimators are already known to suffer from variance which explodes exponentially in T, often rendering them impractical[89]. In our case, their variance would (in the worst case) scale exponentially in N as well.

7.4.2 Main Contribution: Assignment Permutation

We present a novel approach that counters both challenges to compute a stable, accurate estimator. The key idea behind our estimator is to identify hypothetical trials with counterfactual experimental group assignments, whose reward outcomes can be exactly determined using the given outcomes from the original trial. We leverage the fact that although the state transitions depend only the received allocations, regardless of what policy π chooses those allocations.

As a warm-up, consider a single-shot trial $\mathcal{T}$ in which two individuals i and j are assigned to policies π_i and π_j, that make identical resource allocations a to both individuals, yielding outcomes s_i and s_j respectively. Now consider a hypothetical trial $\mathcal{T}^\dagger$, run exactly identical to $\mathcal{T}$ except that the assignments of i and j are switched. If in $\mathcal{T}^\dagger$, both i and j receive the same allocation a as in $\mathcal{T}$, allocations to other individuals would also remain unaf-

fected, and consequently, all individuals would see identical inputs in both $\mathcal{T}^\dagger$ and $\mathcal{T}$. Thus, the actual sample of outcomes $\mathbf{s}$ in $\mathcal{T}$ is a sample from the same distribution as that induced by $\mathcal{T}^\dagger$. Generalizing this idea, consider a sequential RCT $\mathcal{T}$, in which a subset of individuals' group assignments are permuted to construct a hypothetical trial $\mathcal{T}^\dagger$, which sees new allocations $\mathbf{a}_t^\dagger$ made at time t. If an individual experiences sub-trajectories of states $s_{0:t-1}$ and actions $a_{1:t-1}$ till time $t-1$ that are identical in both $\mathcal{T}^\dagger$ and $\mathcal{T}$, and if the new allocation $a_t^\dagger$ received is also identical to a_t, then original state sample s_t observed in $\mathcal{T}$, is also a valid sample in $\mathcal{T}^\dagger$, drawn from the same distribution, $P^*(\mathbf{x}, s_{0:t-1}, a_{1:t}, s_t)$. Furthermore, inductively, the entire original state trajectory $s_{0:T}$ of $\mathcal{T}$ can be treated as a valid sample for $\mathcal{T}^\dagger$ if $\forall\, t \in [T]$ the input sub-trajectory $s_{0:t-1}$, produces new allocations, $a_t^\dagger$ that are identical to a_t.

We exploit this concept to retrospectively check for all such possible reassignments, that would lead to the same sequence of output actions given the same input sub-sequence of the state-action trajectory at all times. The implication is that this allows us to uncover and aggregate outcomes from several such additional 'observable counterfactual assignments' (defined below) in estimating the performance of a given test policy. Algorithm 5 outlines the idea for a general M-arm setting. Later, in Algorithm 6 we present an efficient algorithm crafted for handling index-based policies.

Definition 7 (Observable Counterfactual Assignment)**.** *For an actual assignment $\mathfrak{C}$, we define $\mathfrak{C}^\dagger$ to be an observable counterfactual assignment if in a hypothetical trial with assignments $\mathfrak{C}^\dagger$, for each $t = 1 \ldots T$ the actions each policy would assign to each individual are identical to the original actions received, conditioned on the state and action histories ($\boldsymbol{s}_{0:t-1}, \boldsymbol{a}_{1:t-1}$) matching up until time $t-1$.*

Estimation via Assignment Permutation: Let $\mathcal{C}^\dagger(\mathfrak{C})$ be the set of all 'observable counterfactual assignments' engendered by a single actual experimental assignment $\mathfrak{C}$. Our proposed estimator averages the outcomes of all such observable counterfactuals:

$$\text{Eval}^\dagger(\pi_m) := \frac{\sum_{\mathfrak{C} \in \mathcal{C}^\dagger} \text{Eval}(\pi_m | \mathfrak{C})}{|\mathcal{C}^\dagger|} \tag{7.3}$$

In Theorem 14, we show that this is an unbiased estimator for the true expectation Eval*. The main technical step (Lemma 23) is to show that the that $\mathcal{C}^\dagger(\mathfrak{C})$ defines a partition over $\mathcal{C}$ (the set of all possible assignments)

where two assignments $\mathfrak{C}_1, \mathfrak{C}_2$ lie in the same part if $\mathcal{C}^\dagger(\mathfrak{C}_1) = \mathcal{C}^\dagger(\mathfrak{C}_2)$. Intuitively, this means that our estimator does not "overweight" any particular counterfactual assignment; it maintains the equal weight that each has in Eval*.

Algorithm 5: Estimation through Assignment Permutation

Input: States : $\{S_1^{N\times T+1}, \ldots, S_M^{N\times T+1}\}$, Actions : $\{A_1^{N\times T}, \ldots A_M^{N\times T}\}$, Assignment, $\mathfrak{C}$: $\{C_1, \ldots, C_M\}$

Output: Eval†

1: Compute $\mathcal{C}^\dagger$, the set of observable counterfactual assignments of $\mathfrak{C}$.

2: Compute $\text{Eval}^\dagger(\pi_m) := \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger} \text{Eval}(\pi_m|\mathfrak{C})}{|\mathcal{C}^\dagger|}$

3: **return** $\text{Eval}^\dagger(\pi_m)$

7.4.3 Theoretical Results

In this section, we prove theoretically that Eval$^\dagger(.)$ gives a more accurate estimate because it is unbiased and simultaneously reduces variance. Let $\dagger$ be a homogeneous relation on $\mathcal{C}$, defined as: $\dagger = \{(\mathfrak{C}_1, \mathfrak{C}_2) \in \mathcal{C}\times\mathcal{C} : \mathfrak{C}_2 \in \mathcal{C}^\dagger(\mathfrak{C}_1)\}$. Intuitively, $\dagger$ represents existence of a valid reshuffling to arrive at a counterfactual assignment $\mathfrak{C}_2$ from $\mathfrak{C}_1$.

Lemma 3. *The relation $\dagger$ is an equivalence relation and the family of sets defined by $\mathcal{C}^\dagger(\cdot)$ forms a partition over $\mathcal{C}$.*

All proofs may be found in the appendix. We leverage this property to prove unbiasedness:

Theorem 14. *$Eval^\dagger(\pi_m)$ is an unbiased estimate of the expected value of the performance, $Eval^*(\pi_m)\ \forall m \in [M]$, defined in equation 7.1. i.e.*

$$\mathbb{E}_{S_m\sim P^*}\mathbb{E}_{\mathfrak{C}\sim\mathcal{C}}[Eval^\dagger(\pi_m)] = Eval^*(\pi_m)\ \forall m \in [M]$$

Theorem 15. *The sample variance of our estimator, $Eval^\dagger(\pi)$ is smaller than the standard estimator, $Eval(\pi)$:*

$$(Eval(\pi)) - (Eval^\dagger(\pi)) =$$
$$\frac{1}{|\mathcal{C}|}\cdot\sum_{j\in[\gamma]}\left[\sum_{\mathfrak{C}\in\mathcal{P}_j} Eval^2(\pi|\mathfrak{C}) - \frac{\left(\sum_{\mathfrak{C}\in\mathcal{P}_j} Eval(\pi|\mathfrak{C})\right)^2}{|\mathcal{P}_j|}\right]$$

≥ 0, *where* $\{\mathcal{P}_1, \ldots, \mathcal{P}_\eta\}$ *is the partition of* $\mathcal{C}$ *induced by* †.

Proof Sketch. We compute the sample variance by first conditioning over the partition $\mathcal{P}_j$ (of the equivalence sets defined by †) that an instance of an assignment, $\mathfrak{C}$ belongs to and then accounting for the variance stemming from the candidate assignments $\mathfrak{C}$ within the partition. Finally, we use the Cauchy-Schwarz inequality to show that the right-hand-side expression in Theorem 15 is non-negative. □

The variance contraction expression of Theorem 15 reduces to zero if and only if $\text{Eval}(\pi|\mathfrak{C})$ is identical $\forall \mathfrak{C} \in \mathcal{P}_j,\ \forall j \in [\eta]$; if different assignments imply different rewards then our estimator exhibits a strict improvement in variance.

7.5 Efficient Swapping Algorithm

Identifying $\mathcal{C}^\dagger$ exhaustively in Algorithm 5 involves iterating through every possible assignment in $\mathcal{C}$ and running the policy to determine if the assignment belongs to $\mathcal{C}^\dagger$. However, the number of possible assignments $|\mathcal{C}|$ grows exponentially with N, making full enumeration infeasible. We show how this computational bottleneck can be circumvented for index-based policies with a modified estimator denoted $\text{Eval}^\dagger_\Upsilon(\cdot)$. This estimator implicitly averages over a subset of the possible permutations in $\mathcal{C}^\dagger$, trading off some variance reduction for computational efficiency.

For ease of exposition, here we consider RCTs with two experimental arms ($M = 2$) employing allocation policies π_0 and π_1 respectively. We aim to estimate $\text{Eval}^*(\pi_j)_{j=\{0,1\}}$. Intuitively, instead of working with the space of all possible assignments, we instead consider all individuals participating in the trial and to identify non-overlapping groups of individuals $\{\mathbf{G}_k\} \subset (C_0 \cup C_1)$ that satisfy certain desirable properties. Specifically, we intend to find sets of 'compatible' individuals such that any subset of individuals within each group $\{\mathbf{G}_k\}$ can be mutually swapped to arrive at either an unchanged assignment or a valid observable counterfactual assignment in $\mathcal{C}^\dagger(\mathfrak{C})$. Our intention is to compute an estimate by replacing the original reward of every individual $i \in \mathbf{G}_k$ by the average of rewards of all individuals in $\mathbf{G}_k$, for every such group $\mathbf{G}_k$ (justification in Theorem 16). We identify these groups by checking for two eligibility conditions pertaining to swappability of individuals.

The first eligibility condition for swapping two individuals i and j is that their original allocations $a(t)$ must be identical to each other $\forall t$, to continue to satisfy the resource constraints in both arms after the swap. To check

for this condition, we partition individuals into super-groups $\{\bar{\mathbf{G}}_1, \ldots, \bar{\mathbf{G}}_\kappa\}$, putting all individuals experiencing the same action vector $a(t) \in \mathcal{A}^T$, in the same super-group, where κ denotes the number of such super-groups. All individuals within each $\bar{\mathbf{G}}_k$ satisfy this first eligibility condition for being included in group $\mathbf{G}_k$. For convenience, we let $\varphi\colon C_0 \cup C_1 \to [\kappa]$ denote a many-to-one map identifying the super-group $\bar{\mathbf{G}}_{\varphi(i)}$ that an individual i belongs to.

The second eligibility condition for swapping an individual is that their new allocation $a^\dagger(t)$ under the new policy must be identical to the original $a(t)$, for the same sequence of input states as in the original trial. For each individual i, we use a binary-valued variable $\Lambda_i \in \{0, 1\}$ to indicate satisfaction of this second condition. We introduce and exploit the 'index-threshold' property here to verify this condition efficiently. We define an index threshold $\tau_j(t)$ as the smallest value among indices $\Upsilon(t)$ of individuals in C_j at time t, that get picked to receive the allocation $a = 1$ under policy π_j. To enable efficient computation, we only allow swaps within a group $\mathbf{G}_k$ that maintain the index thresholds $\tau_j(t)$ at the same values as the original. We implement this constraint by setting $\Lambda_i = 0$ for all individuals exactly at the index threshold. Furthermore, for other individuals i, Λ_i can be cheaply determined by just verifying if the index $\Upsilon_i^{\pi_j}(t)$ lies to the same side of threshold $\tau_j(t)\, \forall\, t \in [T]$ and for $j \in \{0, 1\}$. To summarize

$$\Lambda_i = \begin{cases} 1 & \text{if } \prod_{j=0}^{j=1} (\Upsilon_i^{\pi_j}(t) - \tau_j(t)) > 0\, \forall\, t \in [T] \\ 0 & \text{otherwise} \end{cases} \tag{7.4}$$

Intuitively, $\Lambda_i = 1$ means that $a^\dagger_{1:T}(i) = a_{1:T}(i)$ and indicates that individual i satisfies the second eligibility condition. Taking an intersection of both conditions, we form group $\mathbf{G}_k$ by including all individuals $i \in \bar{\mathbf{G}}_k$ that have $\Lambda_i = 1$. For each group $\mathbf{G}_k$, we compute the reward of a representative average individual as:

$$\tilde{r}_k := \frac{1}{|\mathbf{G}_k|} \sum_{i \in \mathbf{G}_k} r(S[i], A[i]) \tag{7.5}$$

In computation of the final estimate $\text{Eval}^\dagger_\Upsilon(\pi_j)$, we consider all individuals in the arm C_j, but replace the reward of swappable individuals among those (i.e. whose $\Lambda_i = 1$) by $\tilde{r}_{\varphi(i)}$. We leave the rewards of other individu-

als unchanged. Finally we compute $\text{Eval}^{\dagger}_{\Upsilon}(\pi_j)$ by summing up as:

$$\text{Eval}^{\dagger}_{\Upsilon}(\pi_j) = \sum_{i \in C_j} \Big(\Lambda_i \tilde{r}_{\varphi(i)} + (1 - \Lambda_i) r(S[i], A[i]) \Big) \tag{7.6}$$

Our theoretical analysis of this estimator establishes that it corresponds to an instance of the general permutation estimator which averages over a subset of the assignments in $\mathcal{C}^{\dagger}$ (instead of the entire set). The main idea is that we can sub-partition $\mathcal{C}^{\dagger}$ into sets with the same value of the index threshold (shown formally in Lemma 24). We denote the part of $\mathcal{C}^{\dagger}$ where the index thresholds are the same as in the actual trial as $\mathcal{C}^{\dagger}_{\Upsilon}$. Each permutation of individuals within groups $\{\mathbf{G}_k\}$ corresponds to an assignment in $\mathcal{C}^{\dagger}_{\Upsilon}(\mathfrak{C})$, and the final estimator averages over all such assignments:

Theorem 16. *$Eval^{\dagger}_{\Upsilon}(\cdot)$ computed as per Equation 7.6 computes the average of $Eval(|\mathfrak{C})$ over all assignments in $\mathcal{C}^{\dagger}_{\Upsilon}$. i.e. $Eval^{\dagger}_{\Upsilon}(\pi) = \frac{\sum_{\mathfrak{C} \in \mathcal{C}^{\dagger}_{\Upsilon}} Eval(\pi|\mathfrak{C})}{|\mathcal{C}^{\dagger}_{\Upsilon}|}$*

From this, $\text{Eval}^{\dagger}_{\Upsilon}(\pi_j)$ is easily shown to inherit the desirable properties of the general permutation estimator, e.g., Corollary 8 proves that it is also an unbiased estimator of $\text{Eval}^{*}(\pi_j)$. The tradeoff is a slight sacrifice in variance contraction as it yields smaller partitions $\mathcal{P}_j$ of $\mathcal{C}$ (as defined in Theorem 15), since we discard assignments with a different threshold. However, working with $\mathcal{C}^{\dagger}_{\Upsilon}$ enables a computationally efficient algorithm for index policies, avoiding the exponential runtime of the general estimator.

Algorithm 6: Reshuffling between index based policies

Input: $\mathfrak{C} \coloneqq \{C_0, C_1\}$, States: $S_0, S_1 \in \mathcal{S}^{N \times T+1}$, Actions: $A_0, A_1 \in \{0,1\}^{N \times T}$, Indexes: $\Upsilon^{\pi_0}, \Upsilon^{\pi_1} \in \mathbb{R}^{2N \times T}$

Output: Estimates: $\text{Eval}^{\dagger}_{\Upsilon}$

1: Group individuals according to action vectors into $\{\bar{\mathbf{G}}_k\}$ and determine $\varphi(i)\ \forall$ individuals $i \in C_0 \cup C_1$.
2: Determine $\forall t \in [T]$, index thresholds on both arms as: $\bar{\tau}_j(t) \coloneqq \min \left\{ \Upsilon_{\pi_j}(i,t) \,|A_j[i,t] = 1,\ i \in [N] \right\}$
3: Determine $\Lambda_i\ \forall$ individuals i according to Equation 7.4.
4: Compute group $\mathbf{G}_k$ as: $\{i \in \bar{\mathbf{G}}_k \,|\Lambda_i = 1\}, \forall k \in [\kappa]$
5: For each group $\mathbf{G}_k$, compute the average reward of a representative individual as per Equation 7.5.
6: Compute $\text{Eval}^{\dagger}_{\Upsilon}(\pi_j)$ per Equation 7.6

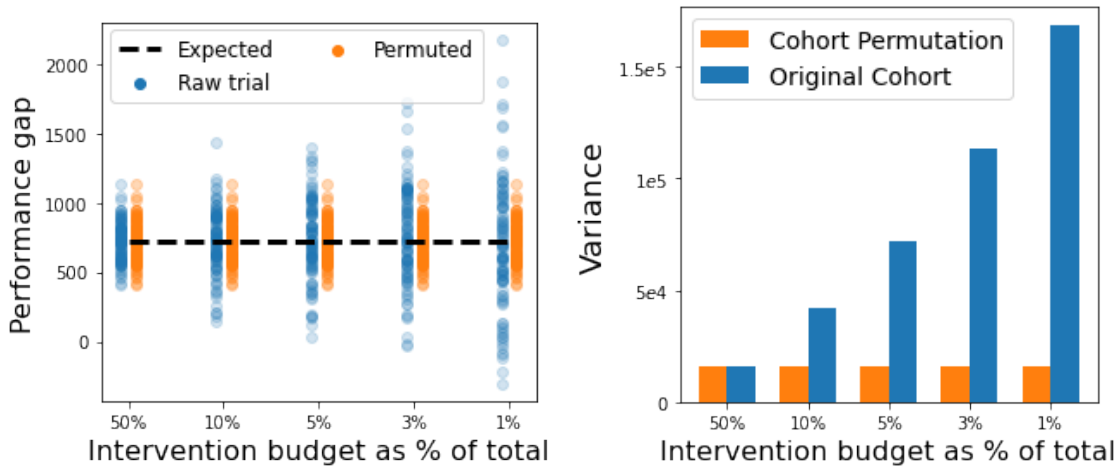
7.6 Empirical Evaluation

We test our proposed methodology empirically on several datasets: (1) synthetic example (2) semi-synthetic tuberculosis medication adherence monitoring data and (3) real-world field trial data from an intervention for a maternal healthcare. We consider a state space $\mathcal{S} = \{0, 1\}$, respectively representing an 'undesirable' and a 'desirable' state (of health, program engagement, etc.). The action space $\mathcal{A} = \{0, 1\}$, denotes 'no delivery' or 'delivery' of an intervention. We assume a budget constraint, limiting the total number of interventions per time step. The reward function is defined as $r(s_{0:T}, a_{1:T}) = \sum_t s_t$, translating to an objective of maximizing the total time spent by individuals in state $s = 1$.

7.6.1 Synthetic Dataset

This setup consists of three types of individuals characterized by their $P-$ matrices. P_1 and P_2 are designed such that it is always optimal to intervene on P_1 individuals consistently, whereas intervening on P_2 individuals is strictly sub-optimal (details in Appendix E.3.1). P_3 individuals are unaffected by interventions, with transition dynamics independent of the action received. We consider two test policies π_1 and π_2, which are designed such that policy π_j always chooses individuals of type P_j to intervene on, when available, making π_1 the optimal policy. We simulate 300 P_1 individuals, 300 P_2 individuals and $(300^*\eta\ P_3)$ individuals and set an intervention budget of 300 per timestep for $T = 20$ timesteps. We measure the performance lift of π_1 against π_2 as: $\Delta := \Delta(\pi_1, \pi_2) = \text{Eval}(\pi_1) - \text{Eval}(\pi_2)$.

In Figure 7.1a, we vary the budget on the x-axis. Each blue dot in the scatter plot shows one independent RCT instance and measures the raw difference in rewards Δ on the y-axis. Applying assignment permutation maps each blue dot to an orange dot. The black dashed line marks the expected value of the performance lift. Visually, both colors are centered on the black line, as both estimators are unbiased. The assignment-permuted estimates lie closer to the expected value than the raw estimates, indicating a smaller sample variance. Quantitatively, we measure the sample variance of Δ on the y-axis in Figure 7.1b. Variance reduces sharply upon applying assignment permutation – for instance, at a budget level of 3%, our approach cuts the variance by 7×, from 11.3×10^4 to 1.6×10^4. The intuition is that both π_1 and π_2 overlap in their decision to not intervene upon P_3 individuals. However, their final rewards are based partly on which P_3 individual gets (randomly) assigned to which group,



(a) Total engagements

(b) Sample Variance

Figure 7.1: Estimates and sample variance in synthetic domain.

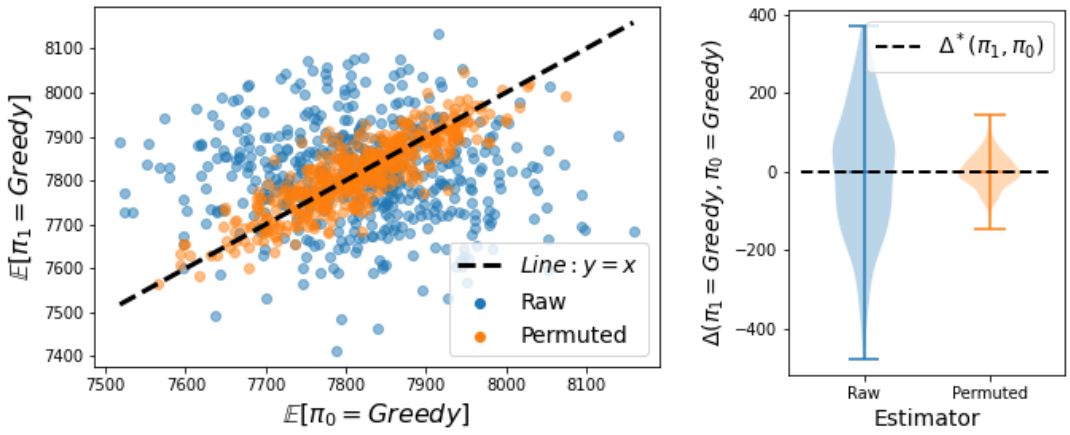
independent of the underlying policies. Assignment permutation counters this randomness by averaging over alternate assignments of the P_3 individuals.

7.6.2 Semi-synthetic evaluation with tuberculosis dataset

We use real tuberculosis medication adherence monitoring data, consisting of daily records of patients in Mumbai, India, obtained from[79] and simulate patient behavior by estimating the P matrix. More details can be found in the appendix.

We consider two policies: a "Whittle index" policy[163] that attempts to maximize long-run reward, and a greedy policy which optimizes an estimate of next-step reward. We simulate $N = 1000$ patients in each arm and vary the budget constraint. We consider both the multi-step setting ($T = 10$) and single step ($T = 1$). We compare three estimation methods: "Raw" is the naive average of outcomes in each arm, "Permuted" our proposed estimator, and "IPW" the inverse propensity estimator from Section 4.1 (available only for $T = 1$). In practice, we find it necessary to trim propensity scores for IPW[177] to the range $[0.01, 0.99]$, since extreme values lead to very large variance. This introduces a slight bias, visible in Figure 7.3b.

Table E.1 shows the sample variance of estimates returned by each method. For unbiased methods (Raw, Permuted) the variance is also their mean squared error; this holds approximately for IPW due to trimming. Ta-



(a) Total engagements

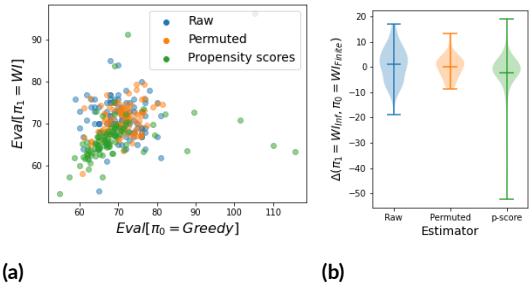
(b) Sample Variance

Figure 7.2: Multi-step setting. (a) The permuted estimates (orange) are closer to the true expectation (black line) than the raw estimates (blue) (b) Assignment permutation reduces variance.

ble E.1 includes an additional column labeled '$n-$value'. To benchmark the improvement produced by our method, this gives the minimum number of *independent* RCTs that would need to be run (and averaged over) to match the sample variance achieved by assignment permutation (computed by simulation).

For all comparisons and parameter settings, we find that our assignment permutation estimator produces a substantial improvement in variance. Indeed, achieving a comparably precise estimate using the naive raw estimator would require running anywhere from 2 to *13* independent RCTs. This underscores the importance of variance reduction – running RCTs is hugely costly and assigns many individuals to suboptimal policies; improved analysis allows us to draw comparably precise conclusions at dramatically lower cost.

Figure 7.2 illustrates this improvement in a single example where both trial arms are the Greedy policy and so the expected difference in rewards is exactly zero (with $B = 3\%$ and $T = 10$). Each dot in Figure 7.2a corresponds to a single instance of a trial, with the x- and y-axes giving total engagements in the two arms. The black dashed line ($y = x$) denotes the expectation, $\Delta = 0$. The blue dots, representing raw measurements, have a wider spread around the black line than the orange dots obtained via assignment permutation. Figure 7.2b, shows a violin plot of the sample difference in rewards between the two arms. Both violins are centered on the zero line, reflecting that the estimators are unbiased. The violin corresponding to assignment permutation is more compact, indicating lower sample variance.



(a) (b)

Figure 7.3: Illustration of results for single-step setting

Table 1: Sample variance in Measured Performance Lift

T	B	π_1 v π_0	RAW	PERMUTED	IPW	n-VAL
1	3%	π_{WI} v π_{GR}	49.09	4.94	**0.48**	9
1	10%	π_{WI} v π_{GR}	49.86	15.11	**6.66**	3
1	25%	π_{WI} v π_{GR}	49.45	**19.94**	78.12	2
10	3%	π_{WI} v π_{WI}	2381	**916**	NA	3
10	3%	π_{WI} v π_{GR}	2348	**728**	NA	4
10	3%	π_{GR} v π_{GR}	26356	**1860**	NA	13
10	10%	π_{GR} v π_{GR}	25983	**3808**	NA	7
10	25%	π_{GR} v π_{GR}	23619	**5477**	NA	5

In the single-step setting, the IPW estimator returns mixed results: it has the best variance for small values of the budget, but actually performs *worse* than the naive raw estimator for larger budgets. Essentially, the overlap between two policies becomes smaller as the budget increases because they agree only on the few highest-priority individuals. Low overlap translates into extreme propensity scores, inflating variance. Figures 7.3b and 7.3a show an illustration, where IPW often produces an improvement but is susceptible to large outliers. However, when overlap is high and we operate only in the single-step setting, IPW can be a valuable option.

7.6.3 Case study: Real-world Trial

Our method is directly applicable to real-world settings; we show this by considering an actual large-scale RCT reported in [103] evaluating a Restless Multi-Armed Bandit-based algorithm for resource allocation in a maternal and child healthcare. The data consists of 23,000 real-world beneficiaries, randomly split between three groups for the trial: RMAB algorithm, baseline algorithm and a control group, which sees no interventions. Real-world health workers delivered interventions recommended by the algorithms. We consider the performance lift of the RMAB algorithm in improving engagement with the program in comparison to the control group and apply our proposed permutation algorithm to the originally reported raw results. Figure 7.4 (left) plots the total engagement numbers on the y-axis as a function of time (in weeks) on the x-axis. Figure 7.4 (right) computes the lift provided by the RMAB algorithm, as defined in [103] on the y-axis. Our findings suggest that the performance lift of RMAB algorithm is larger than originally reported and by week 7, RMAB is estimated to prevent 815 en-

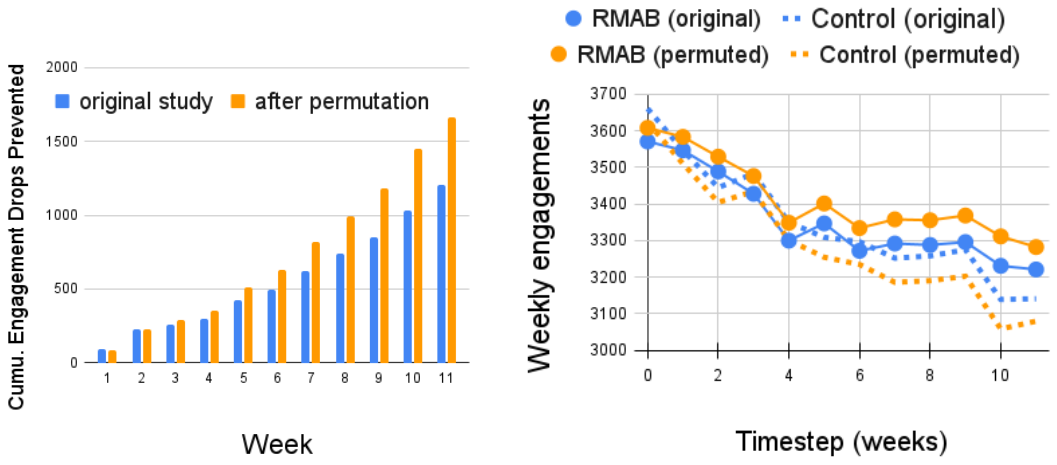


Figure 7.4: Impact of permutation estimator on real-world data.

gagement drops, vs the originally reported 622. Since this is real data, the true values are unknown. However, this case study provides evidence that the variance reduction provided by our estimator can be significant in practice.

7.7 Conclusion

We address the critical gap of mitigating the error in evaluation of resource allocation policies through RCTs. We propose a new estimator using a novel concept based on the idea of retrospective reassignment of participants to experimental arms. We prove that our estimator is unbiased while simultaneously reducing sample variance, and hence reduces error. Through empirical tests on multiple data sets – including a real-world dataset in a socially critical domain – we show that our approach cuts error by as much as 70% and from a single given RCT, can achieve benefits equivalent of running upto 13 *independent* RCTs in parallel.

8

Conclusion and Future Vision

To conclude, my thesis centers around designing and deploying innovative AI solutions aimed at improving public health outcomes, especially for the underserved and under-resourced communities. My thesis considers the data-to-realized positive social impact pipeline, consisting of optimization of available resources, deployment and measurement of impact. I show that achieving such impact may often require overcoming several fundamental research questions along this pipeline. In my thesis, I propose several ideas and algorithmic solutions to some of these problems in the context of two public health application domains: (1) Tuberculosis prevention and (2) improving maternal and child health.

Future Vision: My vision for future research is grounded in my drive and commitment to advancing AI techniques spanning the entire data-to-impact pipeline integrating tools from machine learning, decision-making and

(a) (b) (c)

Figure 8.1: (a) Field visit in Mumbai. (b) AI vs TB workshop. (C) Immersive discussions with ARMMAN staff

causal inference. Below I outline a few avenues of future research work that I deem important to address.

1. Limited Resource Allocation: Limited resources are ubiquitous, not just in public health settings, but in a myriad of other contexts and applications. Towards optimizing the use of these resources, it is important to push the frontiers of (sequential) planning to build robust, efficient algorithms for a wider class of models. I aim to generalize and expand on the MDP family of models and build scalable algorithms for non-markov or non-stationary resource allocation settings, building methods that are more useful in practical contexts.

2. Data Scarcity and Compute Efficiency: Scare, missing or incorrect data, present a difficult challenge in most applications of AI today. Decision-Focused Learning (DFL) is a paradigm that has shown promise in being able to squeeze out better quality of decisions despite limited data. However, current advances only scratch the surface. I intend to throw light on why DFL works or when it works best, improving our understanding of DFL through sharp theoretical insights. I am keen to combine ideas from surrogate models and DFL to discover clever ways to unlock computational speedup to make this technique scalable and accessible in real-world contexts.

3. Evaluation through deployment: I envision my recent work [106] on improved evaluation of resource allocation policies to open up an entire direction of research on AI in RCTs. Inference through RCTs for resource allocation methods has received little attention. I intend to address several unanswered ambitious questions pertaining to design of smart RCTs, with impactful consequences. For instance, the challenge of deriving confidence interval techniques for the evaluation of resource allocation policies through RCTs is still an important open problem to solve. I also intend to build new methodology using tools from causal inference to evaluate stochastic policies as well as methods to identify optimal experimental design, to reduce dependence on expensive large-scale RCTs.

A

Appendix to Chapter 1

A.1 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 2. *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: $\mathcal{B} \coloneqq \{b_{2T}, b_{2T-1}, \ldots, b_i, \ldots, b_1\}$.* 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 $b_{i+1} \geq b_{th} > b_i\ \forall i \in 1, \ldots, 2T$ 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 $\Pi = \{\pi_{2T+1}, \pi_{2T}, \ldots, \pi_1\}$ 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.† Thus the threshold policy π_i would follow: $b_i > b^*_{th}(\pi_i) \geq b_{i-1}\ \forall i \in 1, \ldots, 2T+1$, where $b_0 \coloneqq -\infty$ and $b_{2T+1} \coloneqq \infty$. 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:

$$\begin{aligned} &\forall m_1, m_2 \text{ such that } m_1 \leq m_2, \\ &\text{if } \pi^*(m_1) = \pi_i \text{ and } \pi^*(m_2) = \pi_j, \text{ then} \\ &\implies i \geq j \end{aligned} \tag{A.1}$$

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

Lemma 4. 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 $m^*_{2T+1} < m^*_{2T} < \ldots < m^*_1$.

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 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.

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

For the inductive case, assume the hypothesis, $m^*_{2T+1} < \ldots < m^*_{t+1} < m^*_i \; \forall i < t+1$. 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 $L = \{\pi_t, \pi_{t-1}, \ldots \pi_1\}$ as the set of "lower-side" polices and $U = \{\pi_{2T+1}, \pi_{2T}, \ldots \pi_{t+1}\}$ 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 $L = \{\pi_t, \pi_{t-1}, \ldots \pi_1\}$ (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 $\{\bar{m}_l\}_{l=1}^{\infty}$ 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 $\{\bar{m}_l\}_{l=1}^{\infty}$ 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 $U = \{\pi_{2T+1}, \pi_{2T}, \ldots \pi_{t+1}\}$. 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,

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

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

Let π_i be the optimal policy at some m_1.

$$\implies m_i^* \leq m_1$$

$$\implies m_j^* < m_i^* \leq m_1 \; \forall j > i \text{ using Lemma 4}$$

Let $V_\pi(m, b)$ be the discounted reward of policy π at arbitrary state b as defined in Eq. 1.2 of the main text. Then for any $V_{\pi_i}(m, b)$ and $V_{\pi_j}(m, b)$ such that $j > i$ we have:

$$V_{\pi_i}(m_j^*, b) < V_{\pi_j}(m_j^*, b) \; (\pi_j \text{ is optimal at } m_j^*) \tag{A.2}$$

$$V_{\pi_i}(m_i^*, b) \geq V_{\pi_j}(m_i^*, b) \; (\pi_i \text{ is optimal at } m_i^*) \tag{A.3}$$

$$m_j^* < m_i^* \text{ if } j > i \tag{A.4}$$

$$\implies \frac{dV_{\pi_i}}{dm} > \frac{dV_{\pi_j}}{dm} \forall j > i \tag{A.5}$$

Where Eq. A.2 is a strict inequality as implied by Lemma 4 and Eq. A.5 follows from Fact 2 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 $j \leq i$. Towards a contradiction, assume $j > i$. Then similar to the above

equations, we have the following:

$$V_{\pi_i}(m_j, b) \leq V_{\pi_j}(m_j, b) \,(\pi_j \text{ is optimal at } m_j) \tag{A.6}$$

$$V_{\pi_i}(m_i^*, b) \geq V_{\pi_j}(m_i^*, b) \,(\pi_i \text{ is optimal at } m_i^*) \tag{A.7}$$

$$m_i^* < m_j \tag{A.8}$$

$$\implies \frac{dV_{\pi_i}}{dm} \leq \frac{dV_{\pi_j}}{dm} \forall j > i \tag{A.9}$$

Where Eq. A.9 follows from Fact 2 and the value function's linear dependence on m (whether discounted or average reward criterion). which contradicts Eq. A.5. Therefore, our claim holds. From A.1, that implies indexability. □

A.2 Technical Condition for Forward Threshold Policies to be Optimal

We restate Eq. 1.2 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 $\tau(b) \coloneqq \tau_1(b)$ from Eq. 1.1. Simplified, $\tau(b)$ is simply a linear function of b given by the expression

$$\begin{aligned} \tau(b) &= bP^p_{1,1} + (1-b)P^p_{0,1} \\ &= (P^p_{1,1} - P^p_{0,1})b + P^p_{0,1} \end{aligned} \tag{A.10}$$

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

Fact 3. $\frac{d(\tau(b))}{db} = (P^p_{1,1} - P^p_{0,1}) \leq 1$.

Fact 4. $\forall b, b'$ s.t. $b \geq b'$, $\tau(b) \geq \tau(b')$.

Facts 5 and 6 follow from Eq A.10.

Lemma 5. $V_m(b_1) - V_m(b_2) \geq b_1 - b_2, \forall b_1, b_2$ 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 $V^1_m(b_1) - V^1_m(b_2) =$

$$m + b_1 - (m + b_2) = b_1 - b_2 \qquad \text{passive} \tag{A.11}$$

$$b_1 - b_2 = b_1 - b_2 \qquad \text{active} \tag{A.12}$$

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):

$$\begin{aligned} &= m + b_1 + \beta V^t_m(\tau(b_1)) - (m + b_2 + \beta V^t_m(\tau(b_2))) \\ &= b_1 - b_2 + \beta\Big(V^t_m(\tau(b_1)) - V^t_m(\tau(b_2))\Big) \\ &\geq b_1 - b_2 + \beta(\tau(b_1) - \tau(b_2)) \\ &\geq b_1 - b_2 \end{aligned} \tag{A.13}$$

Case 2 (both active):

$$\begin{aligned} &= b_1 - b_2 + \beta\Big((b_1 - b_2)V^t_m(P^a_{1,1}) + (b_2 - b_1)V^t_m(P^a_{0,1})\Big) \\ &= b_1 - b_2 + \beta\Big((b_1 - b_2)(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1}))\Big) \\ &= (b_1 - b_2)(1 + \beta(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1})) \\ &\geq (b_1 - b_2)(1 + \beta * 0) \\ &= (b_1 - b_2) \end{aligned} \tag{A.14}$$

□

Corollary 3. $V_m(b)$ 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 5 by setting $b_1 = b$ and $b_2 = b'$. □

Lemma 6. $V_m(b_1) - V_m(b_2) \leq \frac{b_1 - b_2}{1-\beta}, \forall b_1, b_2$ s.t. $b_1 > b_2$

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

$$m + b_1 - (m + b_2) = b_1 - b_2 \leq \frac{b_1 - b_2}{1-\beta} \qquad \text{both passive} \tag{A.15}$$

$$b_1 - b_2 = b_1 - b_2 \leq \frac{b_1 - b_2}{1-\beta} \qquad \text{both active} \tag{A.16}$$

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

Case 1 (both passive):

$$\begin{aligned}
&= \big(m + b_1 + \beta V^t_m(\tau(b_1))\big) - \big(m + b_2 + \beta V^t_m(\tau(b_2))\big) \\
&= (b_1 - b_2) + \beta\big(V^t_m(\tau(b_1)) - V^t_m(\tau(b_2))\big) \\
&\leq (b_1 - b_2) + \beta\left(\frac{\tau(b_1) - \tau(b_2)}{1-\beta}\right) \\
&\leq (b_1 - b_2) + \beta\left(\frac{(b_1 - b_2)}{1-\beta}\right) \text{ by Fact 6} \\
&= \frac{b_1 - b_2}{1-\beta}
\end{aligned} \tag{A.17}$$

Case 2 (both active):

$$\begin{aligned}
&= \Big(b_1 + \beta\big(b_1 V^t_m(P^a_{1,1}) + (1 - b_1) V^t_m(P^a_{0,1})\big)\Big) - \\
&\Big(b_2 + \beta\big(b_2 V^t_m(P^a_{1,1}) + (1 - b_2) V^t_m(P^a_{0,1})\big)\Big) \\
&= (b_1 - b_2) + \beta\Big(\big(b_1 - b_2\big)\big(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1})\big)\Big) \\
&\leq (b_1 - b_2) + \beta\left((b_1 - b_2).\frac{P^a_{1,1} - P^a_{0,1}}{1-\beta}\right) \\
&\leq (b_1 - b_2) + \beta\left(\frac{(b_1 - b_2)}{1-\beta}\right) \text{ by Fact 5} \\
&= \frac{b_1 - b_2}{1-\beta}
\end{aligned} \tag{A.18}$$

□

Lemma 7. $\frac{d(V_m(b))}{db} \geq 1 + \beta\alpha$

where, $\alpha = \min\{P^p_{1,1} - P^p_{0,1}, P^a_{1,1} - P^a_{0,1}\}$

Proof. Using Eq. 1.2, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} 1 + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))} \frac{d(\tau(b))}{db} & \text{passive} \\ 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) & \text{active} \end{cases} \tag{A.19}$$

Case 1 (passive):

$$= 1 + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))}(P^p_{1,1} - P^p_{0,1}) \tag{A.20}$$

$$= 1 + \beta \lim_{\delta \to 0} \frac{V_m(\tau(b) + \delta) - V_m(\tau(b))}{\tau(b) + \delta - \tau(b)}(P^p_{1,1} - P^p_{0,1}) \tag{A.21}$$

$$\geq 1 + \beta(P^p_{1,1} - P^p_{0,1}) \text{ by Lemma 5} \tag{A.22}$$

$$\geq 1 + \beta\alpha \tag{A.23}$$

Case 2 (active):

$$= 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \tag{A.24}$$

$$\geq 1 + \beta(P^a_{1,1} - P^a_{0,1}) \text{ by Lemma 5} \tag{A.25}$$

$$\geq 1 + \beta\alpha \tag{A.26}$$

□

Now we derive the technical condition for **Theorem 6**. 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 6 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:

$$(P^p_{1,1} - P^p_{0,1})(1 + \beta(P^a_{1,1} - P^a_{0,1}))(1 - \beta) \geq P^a_{1,1} - P^a_{0,1} \tag{A.27}$$

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.A.27 itself implies that $\alpha = P^a_{1,1} - P^a_{0,1}$. Thus, it becomes:

$$(P^p_{1,1} - P^p_{0,1})(1 + \beta\alpha)(1 - \beta) \geq P^a_{1,1} - P^a_{0,1} \tag{A.28}$$

$$\implies (P^p_{1,1} - P^p_{0,1})(1 + \beta\alpha) \geq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma 6} \tag{A.29}$$

$$\implies (P^p_{1,1} - P^p_{0,1})\frac{d(V_m(b))}{db} \geq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma. 7} \tag{A.30}$$

$$\implies 1 + \beta\frac{d(V_m(\tau(b)))}{d(\tau b)}\frac{d(\tau(b))}{db} \geq 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \text{ by Fact 5} \tag{A.31}$$

$$\implies \frac{d(V_m(b|a=0))}{d(b)} \geq \frac{d(V_m(b|a=1))}{d(b)} \tag{A.32}$$

$$\tag{A.33}$$

A.3 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 8. $\frac{d(V_m(b))}{db} \leq 1 + \frac{\beta\gamma}{1-\beta}$

where, $\gamma = \max\{P^p_{1,1} - P^p_{0,1}, P^a_{1,1} - P^a_{0,1}\}$

Proof. Using Equation A.2, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} 1 + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}\frac{d(\tau(b))}{db} & \text{passive} \\ 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) & \text{active} \end{cases} \tag{A.34}$$

Case 1 (passive):

$$= 1 + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))}(P^p_{1,1} - P^p_{0,1}) \tag{A.35}$$

$$= 1 + \beta \lim_{\delta \to 0} \frac{V_m(\tau(b) + \delta) - V_m(\tau(b))}{\tau(b) + \delta - \tau(b)}(P^p_{1,1} - P^p_{0,1}) \tag{A.36}$$

$$\leq 1 + \frac{\beta}{1-\beta}(P^p_{1,1} - P^p_{0,1}) \text{ by Lemma 6} \tag{A.37}$$

$$\leq 1 + \frac{\beta\gamma}{1-\beta} \tag{A.38}$$

Case 2 (active):

$$= 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \tag{A.39}$$

$$\leq 1 + \frac{\beta}{1-\beta}(P^a_{1,1} - P^a_{0,1}) \text{ by Lemma 6} \tag{A.40}$$

$$\leq 1 + \frac{\beta\gamma}{1-\beta} \tag{A.41}$$

□

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 7 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:

$$(P^p_{1,1} - P^p_{0,1})(1 + \frac{\beta(P^a_{1,1} - P^a_{0,1})}{1-\beta}) \leq P^a_{1,1} - P^a_{0,1} \tag{A.42}$$

Note that the Eq. A.42 itself implies that $\gamma = P^a_{1,1} - P^a_{0,1}$, thus giving:

$$(P^p_{1,1} - P^p_{0,1})(1 + \frac{\beta\gamma}{1-\beta}) \leq P^a_{1,1} - P^a_{0,1} \tag{A.43}$$

$$\implies (P^p_{1,1} - P^p_{0,1})(1 + \frac{\beta\gamma}{1-\beta}) \leq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma 5} \tag{A.44}$$

$$\implies (P^p_{1,1} - P^p_{0,1})\frac{d(V_m(b))}{db} \leq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma 8} \tag{A.45}$$

$$\implies 1 + \beta\frac{d(V_m(\tau(b)))}{d(\tau b)}\frac{d(\tau(b))}{db} \leq 1 + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \text{ by Fact 5} \tag{A.46}$$

$$\implies \frac{d(V_m(b|a=0))}{d(b)} \leq \frac{d(V_m(b|a=1))}{d(b)} \tag{A.47}$$

$$\tag{A.48}$$

A.4 Threshold Conditions for Average Reward Case

First we define the concept of *value boundedness*[36]:

Definition 8 (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 $M(b)$:*

$$M(b) < V_\beta(b) - V_\beta(z) < M \tag{A.49}$$

We now prove that Thm. 6 and Thm. 7 hold respectively under the average reward criterion as $\beta \to 1$ using Dutta's Theorem as follows[36]. 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. 6 or Thm. 7 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 17. *Any Collapsing Bandit is value bounded.*

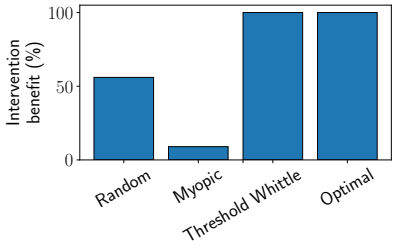
A.5 Example When the Myopic Policy Fails

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. A.1a.

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

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

(a)



(b)

Figure A.1: For the example transition matrices, Myopic performs worse than random, while Threshold Whittle is nearly optimal.

Fig. A.1b 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.

A.6 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 [150], which is a heuristic which has been shown to work well in practice and has been frequently used in the bandit literature [72]. 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 1.3. 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 $m : f_m^{(X_i,X_j)} = f_m^{(X_i+1,X_j)}$. These computations are repeated every day yielding overall complexity of $\mathcal{O}(|\Omega|T^2)$ per process.

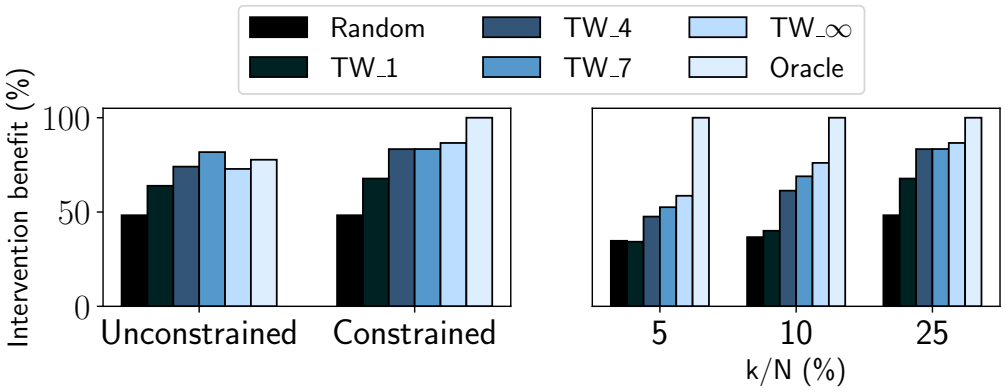


Figure A.2: (left) Constrained Thompson sampling improves learning. (right) buffer lengths of 4-7 perform well for various values of k/N, using constrained Thompson sampling. TW_X is the on-demand index algorithm run in tandem with Thompson sampling and a buffer length of X.

Fig. A.2 (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. A.2 (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.

A.7 Sensitivity Analysis

In Fig. A.3, 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.

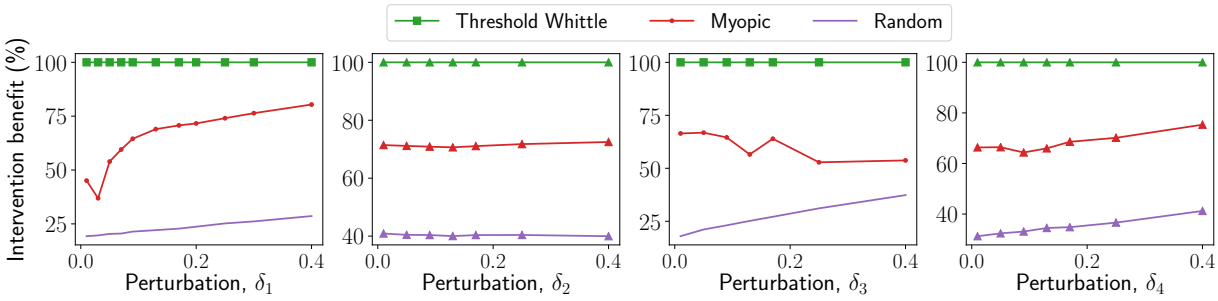


Figure A.3: Performance of Threshold Whittle is robust to perturbation of the transition matrix parameters. Note that 100% corresponds to the performance of Threshold Whittle for this plot only.

A.8 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. 6 and Thm. 7, respectively, using $\beta = 0.95$, then compute their indices with the Threshold Whittle algorithm. Figures. A.4a and A.4b show a few samples of these trajectories for reverse and forward threshold optimal processes, respectively. Via similar arguments from the proof in Appendix A.1, it can be shown that the true Whit-

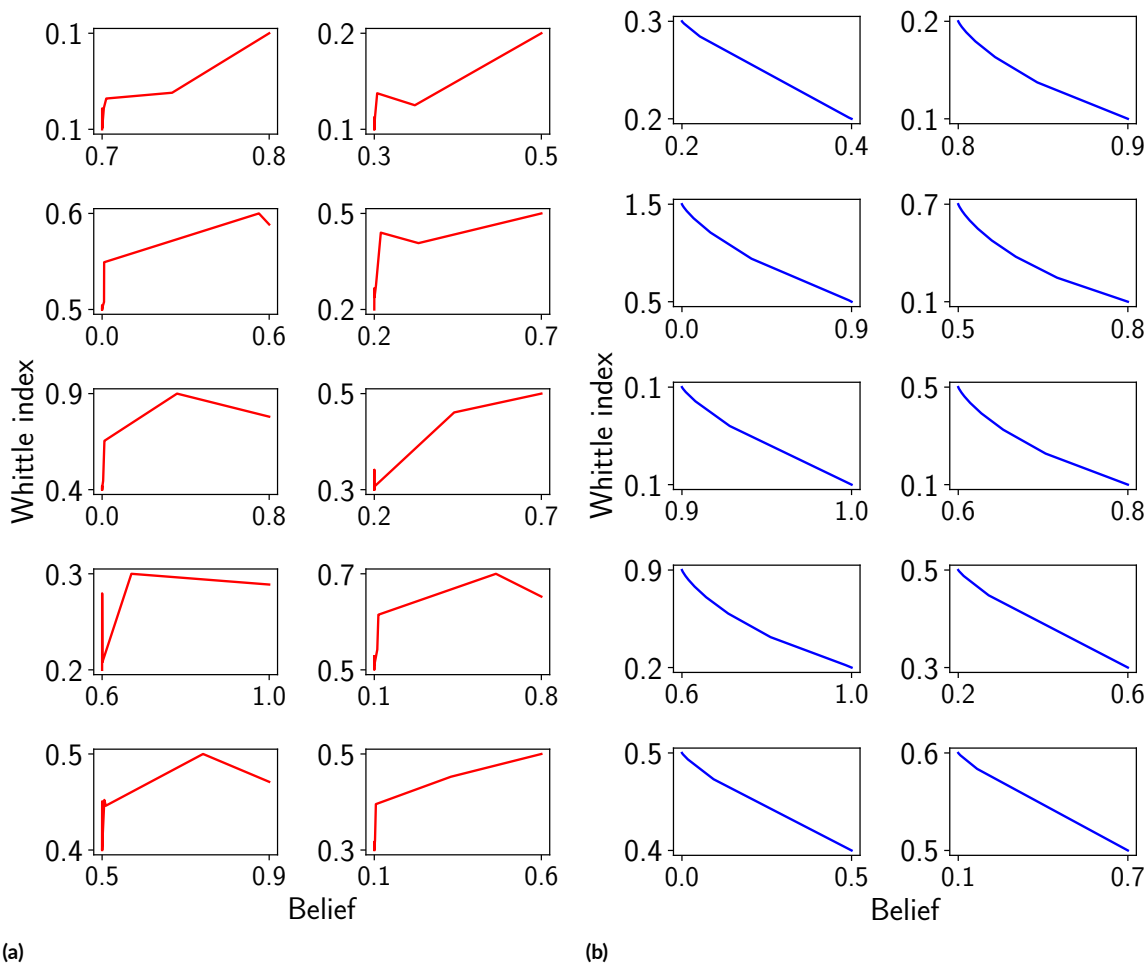


Figure A.4: (a) Threshold Whittle-computed indices vs. reachable beliefs for 10 randomly sampled reverse threshold optimal processes (one line per process). These indices tend to increase in belief, as expected for reverse threshold optimal processes according to the proof in Appendix A.1. (b) Threshold 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 A.1.

tle indices for reverse (forward) threshold optimal processes should always be increasing (decreasing) in belief. Fig. A.4a 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. A.4b shows that for forward threshold optimal policies, the indices computed by Threshold Whittle always decrease in belief as expected.)

B

Appendix to Chapter 2

B.1 Proof of Theorem 4

Theorem 4 (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:*

$$\frac{\Delta_p(1-\beta\max\{\Delta_p,\Delta_a\})}{\Delta_a(1-\beta\min\{\Delta_p,\Delta_a\})} \geq \frac{\rho'_{max}}{\rho'_{min}} \tag{2.3}$$

Proof. We start with presenting three facts and proving several lemmas that underpin the proofs of Thms. 4 and 5.

Fact 5. $\frac{d(\tau(b))}{db} = P^{p}_{11} - P^{p}_{01}$

Fact 6. $\forall b, b'$ s.t. $b \geq b', \tau(b) \geq \tau(b')$.

Fact 7. $\forall b, b'$ s.t. $b \geq b', \tau(b) - \tau(b') = (P^{p}_{11} - P^{p}_{01})(b - b')$.

Lemma 9. $V_m(b_1) - V_m(b_2) \geq \rho'_{min}(b_1 - b_2) \forall\, b_1, b_2$ *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 $V^1_m(b_1) - V^1_m(b_2) =$

$$m + \rho(b_1) - (m + \rho(b_2)) = \rho(b_1) - \rho(b_2) \qquad \text{passive} \tag{B.1}$$

$$\rho(b_1) - \rho(b_2) = \rho(b_1) - \rho(b_2) \qquad \text{active} \tag{B.2}$$

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):

$$\begin{aligned} &= m + \rho(b_1) + \beta V^t_m(\tau(b_1)) - (m + \rho(b_2) + \beta V^t_m(\tau(b_2))) \\ &= \rho(b_1) - \rho(b_2) + \beta\Big(V^t_m(\tau(b_1)) - V^t_m(\tau(b_2))\Big) \\ &\geq \rho(b_1) - \rho(b_2) + \beta\rho'_{min}(\tau(b_1) - \tau(b_2)) \\ &\geq \rho(b_1) - \rho(b_2) \end{aligned} \tag{B.3}$$

Case 2 (both active):

$$
\begin{aligned}
&= \rho(b_1) - \rho(b_2) + \beta\Big((b_1 - b_2)V^t_m(P^a_{1,1}) + (b_2 - b_1)V^t_m(P^a_{0,1})\Big) \\
&= \rho(b_1) - \rho(b_2) + \beta\Big((b_1 - b_2)(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1}))\Big) \\
&\geq \rho'_{min}(b_1 - b_2) + \beta\Big((b_1 - b_2)(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1}))\Big) \\
&\geq \rho'_{min}(b_1 - b_2)
\end{aligned}
\tag{B.4}
$$

□

Lemma 10. *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* $\alpha = \min\{\Delta_a, \Delta_p\}$*:*

$$
V_m(b_1) - V_m(b_2) \geq \rho'_{min}(1 + \beta\alpha\kappa)(b_1 - b_2) \tag{B.5}
$$

Proof. Using Eq. 2.2, we get:

$$
\frac{d(V_m(b))}{db} = \begin{cases} \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}\frac{d(\tau(b))}{db} & \text{passive} \\ \rho'(b) + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) & \text{active} \end{cases} \tag{B.6}
$$

Case 1 (passive):

$$
\begin{aligned}
&= \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau(b))}(P^p_{1,1} - P^p_{0,1}) && \text{(B.7)} \\
&= \rho'(b) + \beta\lim_{\delta\to 0}\frac{V_m(\tau(b) + \delta) - V_m(\tau(b))}{\tau(b) + \delta - \tau(b)}(P^p_{1,1} - P^p_{0,1}) && \text{(B.8)} \\
&\geq \rho'(b) + \beta\kappa\rho'_{min}(P^p_{1,1} - P^p_{0,1}) && \text{(B.9)} \\
&\geq \rho'_{min} + \beta\kappa\rho'_{min}(P^p_{1,1} - P^p_{0,1}) && \text{(B.10)} \\
&\geq \rho'_{min}(1 + \beta\alpha\kappa) && \text{(B.11)}
\end{aligned}
$$

Case 2 (active):

$$= \rho'(b) + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \tag{B.12}$$

$$\geq \rho'(b) + \beta\kappa\rho'_{min}(P^a_{1,1} - P^a_{0,1}) \tag{B.13}$$

$$\geq \rho'_{min} + \beta\kappa\rho'_{min}(P^a_{1,1} - P^a_{0,1}) \tag{B.14}$$

$$\geq \rho'_{min}(1 + \beta\alpha\kappa) \tag{B.15}$$

Thus,

$$\begin{aligned} &\implies \frac{d(V_m(b))}{db} \geq \rho'_{min}(1 + \beta\alpha\kappa) \\ &\implies \int_{b_2}^{b_1} \frac{d(V_m(b))}{db} db \geq \int_{b_2}^{b_1} \rho'_{min}\Big(1 + \beta\alpha\kappa\Big) db \\ &\implies V_m(b_1) - V_m(b_2) \geq \rho'_{min}\big(1 + \beta\alpha\kappa\big)(b_1 - b_2) \end{aligned} \tag{B.16}$$

□

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

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

$$V_m(b_1) - V_m(b_2) \geq f^n(1)\rho'_{min}(b_1 - b_2) \forall n \in \mathbb{W},\ \forall\ b_1, b_2 \text{ s.t. } b_1 \geq b_2 \tag{B.17}$$

Consider the base case, $n = 0$. Eq. B.17 reduces to the statement of Lemma 9 with $f^0(1) = 1$, and thus holds true. For the inductive case, we assume Eq.B.17 to be true for some n and then show that it must also be true for $n + 1$, as follows:

$$\begin{aligned} &\text{If } V_m(b_1) - V_m(b_2) \geq f^n(1).\rho'_{min}(b_1 - b_2), \text{ then} \\ &\implies V_m(b_1) - V_m(b_2) \geq f\Big(f^n(1)\Big).\rho'_{min}(b_1 - b_2). \text{ using Lemma 10} \\ &\implies V_m(b_1) - V_m(b_2) \geq f^{n+1}(1)\rho'_{min}(b_1 - b_2) \end{aligned} \tag{B.18}$$

Thus we show Eq. B.17 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:

$$\begin{aligned}
&\text{Let the sequence converge to } f^\infty = \lim_{n\to\infty} f^n(1) \\
\implies & f(f^\infty) = f\big(\lim_{n\to\infty} f^n(1)\big) = \lim_{n\to\infty} f^{n+1}(1) = \lim_{n\to\infty} f^n(1) = f^\infty \\
\implies & 1 + \beta\alpha f^\infty = f^\infty \\
\implies & 1 = f^\infty(1-\beta\alpha) \\
\implies & f^\infty = \frac{1}{(1-\beta\alpha)}
\end{aligned} \tag{B.19}$$

Resubstituting f^∞ in place of $f^n(1)$ in Eq.B.17 finally gives us the required result. □

Corollary 4. $\frac{d(V_m(b))}{db} \geq \frac{\rho'_{min}}{1-\beta\alpha}$

Proof. This follows from Lemma 11 by setting $b_1 = b + \delta, b_2 = b$ under the limit $\delta \to 0$. □

Lemma 12. $V_m(b_1) - V_m(b_2) \leq \frac{\rho'_{max}(b_1-b_2)}{1-\beta} \forall\ b_1, b_2\ s.t.\ b_1 \geq b_2$

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 $V^t_m(b_1) - V^t_m(b_2) \leq \frac{\rho'_{max}(b_1-b_2)}{1-\beta}$. Then, $V^{t+1}_m(b_1) - V^{t+1}_m(b_2)$

Case 1 (both passive):

$$
\begin{aligned}
&= \big(m + \rho(b_1) + \beta V^t_m(\tau(b_1))\big) - \big(m + \rho(b_2) + \beta V^t_m(\tau(b_2))\big) \\
&= (\rho(b_1) - \rho(b_2)) + \beta\big(V^t_m(\tau(b_1)) - V^t_m(\tau(b_2))\big) \\
&\leq (\rho(b_1) - \rho(b_2)) + \beta\left(\frac{\rho'_{max}(\tau(b_1) - \tau(b_2))}{1-\beta}\right) \\
&\leq \rho'_{max}(b_1 - b_2) + \beta\left(\frac{\rho'_{max}(b_1 - b_2)}{1-\beta}\right) \text{ by Fact 7} \\
&= \frac{\rho'_{max}(b_1 - b_2)}{1-\beta}
\end{aligned}
\tag{B.20}
$$

Case 2 (both active):

$$
\begin{aligned}
&= \Big(\rho(b_1) + \beta\big(b_1 V^t_m(P^a_{1,1}) + (1-b_1)V^t_m(P^a_{0,1})\big)\Big) - \\
&\Big(\rho(b_2) + \beta\big(b_2 V^t_m(P^a_{1,1}) + (1-b_2)V^t_m(P^a_{0,1})\big)\Big) \\
&= (\rho(b_1) - \rho(b_2)) + \beta\Big(\big(b_1 - b_2\big)\big(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1})\big)\Big) \\
&\leq (\rho(b_1) - \rho(b_2)) + \beta\left((b_1 - b_2).\frac{\rho'_{max}(P^a_{1,1} - P^a_{0,1})}{1-\beta}\right) \\
&\leq \rho'_{max}(b_1 - b_2) + \beta\left(\frac{\rho'_{max}(b_1 - b_2)}{1-\beta}\right) \\
&= \frac{\rho'_{max}(b_1 - b_2)}{1-\beta}
\end{aligned}
\tag{B.21}
$$

□

Lemma 13. *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* $\gamma = \max\{\Delta_a, \Delta_p\}$*:*

$$
V_m(b_1) - V_m(b_2) \leq \rho'_{max}(1 + \beta\gamma\kappa)(b_1 - b_2) \tag{B.22}
$$

Proof. Using Equation 2.2, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} \rho'(b) + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))} \frac{d(\tau(b))}{db} & \text{passive} \\ \rho'(b) + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) & \text{active} \end{cases} \tag{B.23}$$

Case 1 (passive):

$$= \rho'(b) + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))}(P^p_{1,1} - P^p_{0,1}) \tag{B.24}$$

$$= \rho'(b) + \beta \lim_{\delta \to 0} \frac{V_m(\tau(b) + \delta) - V_m(\tau(b))}{\tau(b) + \delta - \tau(b)}(P^p_{1,1} - P^p_{0,1}) \tag{B.25}$$

$$\leq \rho'(b) + \beta\kappa\rho'_{max}(P^p_{1,1} - P^p_{0,1}) \tag{B.26}$$

$$\leq \rho'_{max}(1 + \beta\gamma\kappa) \tag{B.27}$$

Case 2 (active):

$$= \rho'(b) + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \tag{B.28}$$

$$\leq \rho'(b) + \beta\kappa\rho'_{max}(P^a_{1,1} - P^a_{0,1}) \tag{B.29}$$

$$\leq \rho'_{max} + \beta\rho'_{max}\gamma\kappa \tag{B.30}$$

$$\leq \rho'_{max}(1 + \beta\gamma\kappa) \tag{B.31}$$

Thus,

$$\begin{aligned} &\implies \frac{d(V_m(b))}{db} \leq \rho'_{max}(1 + \beta\gamma\kappa) \\ &\implies \int_{b_2}^{b_1} \frac{d(V_m(b))}{db} db \leq \int_{b_2}^{b_1} \rho'_{max}\Big(1 + \beta\gamma\kappa\Big) db \\ &\implies V_m(b_1) - V_m(b_2) \leq \rho'_{max}\big(1 + \beta\gamma\kappa\big)(b_1 - b_2) \end{aligned} \tag{B.32}$$

□

Lemma 14. $V_m(b_1) - V_m(b_2) \leq \frac{\rho'_{max}(b_1 - b_2)}{1 - \beta\gamma} \forall\, b_1, b_2 \text{ s.t. } b_1 \geq b_2$

Proof. We use an approach similar to the proof of Lemma 11. Consider the function, $g(x) = 1 + \beta\gamma x$ and let

$g^n(x) \coloneqq \underbrace{g(g(...g(x)))}_{g(.) \text{ applied } n \text{ times}}$. We show using induction that, $\forall n \in \mathbb{W},\ \forall\, b_1, b_2$ s.t. $b_1 \geq b_2$:

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

(B.33)

Consider the base case, $n = 0$. Eq. B.33 reduces to the statement of Lemma 19 with $g^0\big(\frac{1}{1-\beta}\big) = \frac{1}{1-\beta}$, and is thus true. For the inductive case, we assume Eq.B.33 to be true for some n and then show that it must also be true for $n+1$, as follows:

$$\begin{aligned}
&\text{If } V_m(b_1) - V_m(b_2) \leq g^n\Big(\frac{1}{1-\beta}\Big)\rho'_{max}(b_1 - b_2), \text{ then} \\
\implies &V_m(b_1) - V_m(b_2) \leq g\bigg(g^n\Big(\frac{1}{1-\beta}\Big)\bigg)\rho'_{max}(b_1 - b_2). \text{ by Lemma 20} \\
\implies &V_m(b_1) - V_m(b_2) \leq g^{n+1}\Big(\frac{1}{1-\beta}\Big)\rho'_{max}(b_1 - b_2)
\end{aligned} \tag{B.34}$$

Thus we show Eq. B.33 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:

$$\begin{aligned}
&\text{Let the sequence converge to } g^\infty = \lim_{n\to\infty} g^n(\frac{1}{1-\beta}) \\
\implies &g\big(g^\infty\big) = g\Big(\lim_{n\to\infty} g^n(\frac{1}{1-\beta})\Big) = \lim_{n\to\infty} g^{n+1}(\frac{1}{1-\beta}) = g^\infty \\
\implies &1 + \beta\gamma g^\infty = g^\infty \\
\implies &1 = g^\infty(1-\beta\gamma) \\
\implies &g^\infty = \frac{1}{(1-\beta\gamma)}
\end{aligned} \tag{B.35}$$

Resubstituting g^∞ in place of $g^n(\frac{1}{1-\beta})$ in Eq.B.33 finally gives us the required result. □

Corollary 5. $\frac{d(V_m(b))}{db} \leq \frac{\rho'_{max}}{1-\beta\gamma}$

Proof. This follows from Lemma 21 by setting $b_1 = b + \delta, b_2 = b$ under the limit $\delta \to 0$. □

Now we complete the proof for Thm.4 as follows:

$$Eq.2.3 \implies \Delta_p \geq \frac{\Delta_a(1-\beta\alpha)\rho'_{max}}{(1-\beta\gamma)\rho'_{min}} \tag{B.36}$$

$$\implies \frac{(P^p_{1,1} - P^p_{0,1})\rho'_{min}}{(1-\beta\alpha)} \geq \frac{\rho'_{max}(P^a_{1,1} - P^a_{0,1})}{(1-\beta\gamma)} \tag{B.37}$$

$$\implies \frac{(P^p_{1,1} - P^p_{0,1})\rho'_{min}}{(1-\beta\alpha)} \geq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma 21} \tag{B.38}$$

$$\implies (P^p_{1,1} - P^p_{0,1})\frac{d(V_m(b))}{db} \geq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Cor. 6} \tag{B.39}$$

$$\implies \frac{d(\tau(b))}{db}\frac{d(V_m(b))}{db} \geq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Fact 5} \tag{B.40}$$

$$\implies \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau b)}\frac{d(\tau(b))}{db} \geq \rho'(b) + \tag{B.41}$$

$$\beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \tag{B.42}$$

$$\implies \frac{d(V_m(b|a=0))}{d(b)} \geq \frac{d(V_m(b|a=1))}{d(b)} \tag{B.43}$$

□

B.2 Proof of Theorem 5

Theorem 5 (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:*

$$\frac{\Delta_p(1-\beta\min\{\Delta_p,\Delta_a\})}{\Delta_a(1-\beta\max\{\Delta_p,\Delta_a\})} \leq \frac{\rho'_{min}}{\rho'_{max}} \tag{2.4}$$

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 4, 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 The-

orem 5 satisfy this required property:

$$Eq.2.4 \implies \Delta_p \leq \frac{\Delta_a(1-\beta\gamma)\rho'_{min}}{(1-\beta\alpha)\rho'_{max}} \tag{B.44}$$

$$\implies \frac{(P^p_{1,1}-P^p_{0,1})\rho'_{max}}{(1-\beta\gamma)} \leq \frac{\rho'_{min}(P^a_{1,1}-P^a_{0,1})}{(1-\beta\alpha)} \tag{B.45}$$

$$\implies \frac{(P^p_{1,1}-P^p_{0,1})\rho'_{max}}{(1-\beta\alpha)} \leq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Lemma 11} \tag{B.46}$$

$$\implies (P^p_{1,1}-P^p_{0,1})\frac{d(V_m(b))}{db} \leq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Cor. 6} \tag{B.47}$$

$$\implies \frac{d(\tau(b))}{db}\frac{d(V_m(b))}{db} \leq V_m(P^a_{1,1}) - V_m(P^a_{0,1}) \text{ by Fact 5} \tag{B.48}$$

$$\implies \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau b)}\frac{d(\tau(b))}{db} \leq \rho'(b) + \tag{B.49}$$

$$\beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) \tag{B.50}$$

$$\implies \frac{d(V_m(b|a=0))}{d(b)} \leq \frac{d(V_m(b|a=1))}{d(b)} \tag{B.51}$$

□

B.3 Proof of Theorem 6

Theorem 6. *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:*

$$\Delta_a \leq \Delta_p \text{ and } \Delta_a + \Delta_p \leq \frac{1}{\beta} \tag{2.6}$$

Proof. To prove this thoerem, we show that the condition of Eq. 2.6 satisfies the condition of Thm.4 when

$\rho(b) = b$. Note that $\rho'_{max} = \rho'_{min} = 1$.

$$
\begin{aligned}
Eq.2.6 &\implies (\Delta_p - \Delta_a)(\frac{1}{\beta} - (\Delta_p + \Delta_a)) \geq 0 \\
&\implies (\Delta_p - \Delta_a) - \beta(\Delta_p - \Delta_a)(\Delta_p + \Delta_a) \geq 0 \\
&\implies \Delta_p - \beta\Delta_p^2 - \Delta_a + \beta\Delta_a^2 \geq 0 \\
&\implies \Delta_p(1 - \beta\Delta_p) \geq \Delta_a(1 - \beta\Delta_a) \\
&\implies \frac{\Delta_p(1 - \beta \max\{\Delta_p, \Delta_a\})}{\Delta_a(1 - \beta \min\{\Delta_p, \Delta_a\})} \geq 1 (\because \Delta_p \geq \Delta_a)
\end{aligned}
$$

□

B.4 Proof of Theorem 7

Theorem 7. *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:*

$$\Delta_p \leq \Delta_a \text{ and } \Delta_p + \Delta_a \leq \frac{1}{\beta} \tag{2.7}$$

Proof. To prove this thoerem, we show that the condition of Eq. 2.7 satisfies the condition of Thm.5 when $\rho(b) = b$. Note that $\rho'_{max} = \rho'_{min} = 1$.

$$
\begin{aligned}
Eq.2.7 &\implies (\Delta_p - \Delta_a)(\frac{1}{\beta} - (\Delta_p + \Delta_a)) \leq 0 && \text{(B.52)} \\
&\implies (\Delta_p - \Delta_a) - \beta(\Delta_p - \Delta_a)(\Delta_p + \Delta_a) \leq 0 && \text{(B.53)} \\
&\implies (\Delta_p - \Delta_a) - \beta(\Delta_p^2 - \Delta_a^2) \leq 0 && \text{(B.54)} \\
&\implies \Delta_p - \beta\Delta_p^2 - \Delta_a + \beta\Delta_a^2 \leq 0 && \text{(B.55)} \\
&\implies \Delta_p(1 - \beta\Delta_p) \geq \Delta_a(1 - \beta\Delta_a) && \text{(B.56)} \\
&\implies \Delta_p \leq \frac{\Delta_a(1 - \beta\Delta_a)}{(1 - \beta\Delta_p)} && \text{(B.57)} \\
&\implies \frac{\Delta_p(1 - \beta \min\{\Delta_p, \Delta_a\})}{\Delta_a(1 - \beta \max\{\Delta_p, \Delta_a\})} \leq 1 (\because \Delta_p \leq \Delta_a) && \text{(B.58)}
\end{aligned}
$$

□

B.5 Value Boundedness Theorem

Definition 9 (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 $L(b)$:*

$$L(b) < V_\beta(b) - V_\beta(z) < U_0 \tag{B.59}$$

We use Dutta's Theorem [36] to prove that Thm. 4 and Thm. 5 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 18. *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 6. Therefore we can set $U_0 = 0$. Now according to Lemmas 11 and 21, we have:

$$V_{m,\beta}(P^a_{1,1}) - V_{m,\beta}(b) \leq \frac{P^a_{1,1} - b}{1 - \beta\gamma} \leq \frac{P^a_{1,1} - b}{1 - \gamma} \forall \beta \in [0, 1) \tag{B.60}$$

$$V_{m,\beta}(b) - V_{m,\beta}(P^a_{1,1}) \geq \frac{b - P^a_{1,1}}{1 - \beta\alpha} \geq \frac{b - P^a_{1,1}}{1} \forall \beta \in [0, 1) \tag{B.61}$$

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. □

B.6 Proof of Theorem 8

Theorem 8 (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:*

$$\frac{\Delta_p(1-\beta\max\{\Delta_p,(\Delta_a.\Delta_e)\})}{\Delta_a(1-\beta\min\{\Delta_p,(\Delta_a.\Delta_e)\})} \geq \frac{\rho'_{max}}{\rho'_{min}} \tag{2.11}$$

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:

$$V_m(b) = \max \begin{cases} m + \rho(b) + \beta V_m(\tau(b)) \text{...passive} \\ \rho(b) + \beta\big(\sum_\omega \Theta_\omega(b).V_m(P^a_\omega)\big) \text{...active} \end{cases} \tag{2.10}$$

Lemma 15.

$$\sum_{\omega=1}^{\|\Omega\|-1} \Theta'_\omega(b)\Big(V_m(P^a_\omega) - V_m(P^a_0)\Big) = \sum_\omega \Big(\Theta'_\omega(b)V_m(P^a_\omega)\Big) \tag{B.62}$$

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

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

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

$$m + \rho(b_1) - (m + \rho(b_2)) = \rho(b_1) - \rho(b_2) \qquad \text{passive} \tag{B.63}$$

$$\rho(b_1) - \rho(b_2) = \rho(b_1) - \rho(b_2) \qquad \text{active} \tag{B.64}$$

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):

$$\begin{aligned}
&= m + \rho(b_1) + \beta V^t_m(\tau(b_1)) - (m + \rho(b_2) + \beta V^t_m(\tau(b_2))) \\
&= \rho(b_1) - \rho(b_2) + \beta\Big(V^t_m(\tau(b_1)) - V^t_m(\tau(b_2))\Big) \\
&\geq \rho(b_1) - \rho(b_2) + \beta\rho'_{min}(\tau(b_1) - \tau(b_2)) \\
&\geq \rho(b_1) - \rho(b_2)
\end{aligned} \tag{B.65}$$

Case 2 (both active):

$$\begin{aligned}
&= \rho(b_1) - \rho(b_2) + \beta\Big((\Theta(b_1) - \Theta(b_2))V^t_m(P^a_{1,1}) + \\
&(\Theta(b_2) - \Theta(b_1))V^t_m(P^a_{0,1})\Big) \\
&= \rho(b_1) - \rho(b_2) + \beta\Big((\Theta(b_1) - \Theta(b_2))(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1}))\Big) \\
&\geq \rho'_{min}(b_1 - b_2) + \beta\Big((\Theta(b_1) - \Theta(b_2))(V^t_m(P^a_{1,1}) - V^t_m(P^a_{0,1}))\Big) \\
&\geq \rho'_{min}(b_1 - b_2)
\end{aligned} \tag{B.66}$$

□

Lemma 17. *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* $\alpha = \min\{\Delta_a, \sum_\omega \Delta_{e\omega}\Delta_{p\omega}\}$*:*

$$V_m(b_1) - V_m(b_2) \geq \rho'_{min}(1 + \beta\alpha\kappa)(b_1 - b_2) \tag{B.67}$$

Proof. Using Eq. 2.2, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} \rho'(b) + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))} \frac{d(\tau(b))}{db} & \text{passive} \\ \rho'(b) + \beta(\sum_\omega \Theta_\omega(b) V_m(P^a_\omega)) & \text{active} \end{cases} \tag{B.68}$$

Case 1 (passive):

$$= \rho'(b) + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))}(P^p_{1,1} - P^p_{0,1}) \tag{B.69}$$

$$= \rho'(b) + \beta \lim_{\delta \to 0} \frac{V_m(\tau(b) + \delta) - V_m(\tau(b))}{\tau(b) + \delta - \tau(b)}(P^p_{1,1} - P^p_{0,1}) \tag{B.70}$$

$$\geq \rho'(b) + \beta\kappa\rho'_{min}(P^p_{1,1} - P^p_{0,1}) \tag{B.71}$$

$$\geq \rho'_{min} + \beta\kappa\rho'_{min}(P^p_{1,1} - P^p_{0,1}) \tag{B.72}$$

$$\geq \rho'_{min}(1 + \beta\alpha\kappa) \tag{B.73}$$

Case 2 (active):

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

$$\geq \rho'(b) + \beta\kappa\rho'_{min}\Big(\sum_{\omega=1}^{\|\Omega\|-1} \Theta'_\omega(b)(P^a_\omega - P^a_0)\Big) \tag{B.75}$$

$$\geq \rho'_{min} + \beta\kappa\rho'_{min} \sum_{\omega=1}^{\|\Omega\|-1} (\Delta_{e\omega}\Delta_{p\omega}) \tag{B.76}$$

$$\geq \rho'_{min}(1 + \beta\alpha\kappa) \tag{B.77}$$

Thus,

$$\begin{aligned}
&\implies \frac{d(V_m(b))}{db} \geq \rho'_{min}(1+\beta\alpha\kappa) \\
&\implies \int_{b_2}^{b_1} \frac{d(V_m(b))}{db} db \geq \int_{b_2}^{b_1} \rho'_{min}\Big(1+\beta\alpha\kappa\Big) db \\
&\implies V_m(b_1) - V_m(b_2) \geq \rho'_{min}\big(1+\beta\alpha\kappa\big)(b_1 - b_2)
\end{aligned} \tag{B.78}$$

□

Lemma 18. $V_m(b_1) - V_m(b_2) \geq \frac{\rho'_{min}(b_1-b_2)}{1-\beta\alpha} \forall\ b_1, b_2\ s.t.\ b_1 \geq b_2$

Proof. This proof is exactly same as the proof for Lemma. 11. □

Corollary 6. $\frac{d(V_m(b))}{db} \geq \frac{\rho'_{min}}{1-\beta\alpha}$

Proof. This follows from Lemma 18 by setting $b_1 = b + \delta, b_2 = b$ under the limit $\delta \to 0$. □

Lemma 19. $V_m(b_1) - V_m(b_2) \leq \frac{\rho'_{max}(b_1-b_2)}{1-\beta} \forall\ b_1, b_2\ s.t.\ b_1 \geq b_2$

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 $V^t_m(b_1) - V^t_m(b_2) \leq \frac{\rho'_{max}(b_1-b_2)}{1-\beta}$. Then, $V^{t+1}_m(b_1) - V^{t+1}_m(b_2)$

Case 1 (both passive):

$$
\begin{aligned}
&= \big(m + \rho(b_1) + \beta V^{\pi}_{m}(\tau(b_1))\big) - \big(m + \rho(b_2) + \beta V^{\pi}_{m}(\tau(b_2))\big) \\
&= (\rho(b_1) - \rho(b_2)) + \beta\big(V^{\pi}_{m}(\tau(b_1)) - V^{\pi}_{m}(\tau(b_2))\big) \\
&\leq (\rho(b_1) - \rho(b_2)) + \beta\left(\frac{\rho'_{max}(\tau(b_1) - \tau(b_2))}{1-\beta}\right) \\
&\leq \rho'_{max}(b_1 - b_2) + \beta\left(\frac{\rho'_{max}(b_1 - b_2)}{1-\beta}\right) \text{ by Fact 7} \\
&= \frac{\rho'_{max}(b_1 - b_2)}{1-\beta}
\end{aligned}
\tag{B.79}
$$

Case 2 (both active):

$$
\begin{aligned}
&= \Big(\rho(b_1) + \beta\big(\Theta(b_1) V^{\pi}_{m}(P^{a}_{1,1}) + (1 - \Theta(b_1)) V^{\pi}_{m}(P^{a}_{0,1})\big)\Big) - \\
&\Big(\rho(b_2) + \beta\big(\Theta(b_2) V^{\pi}_{m}(P^{a}_{1,1}) + (1 - \Theta(b_2)) V^{\pi}_{m}(P^{a}_{0,1})\big)\Big) \\
&= (\rho(b_1) - \rho(b_2)) + \beta\Big(\big(\Theta(b_1) - \Theta(b_2)\big)\big(V^{\pi}_{m}(P^{a}_{1,1}) - V^{\pi}_{m}(P^{a}_{0,1})\big)\Big) \\
&\leq (\rho(b_1) - \rho(b_2)) + \beta\left((\Theta(b_1) - \Theta(b_2)).\frac{\rho'_{max}(P^{a}_{1,1} - P^{a}_{0,1})}{1-\beta}\right) \\
&\leq \rho'_{max}(b_1 - b_2) + \beta\left(\frac{\rho'_{max}(\Theta(b_1) - \Theta(b_2))}{1-\beta}\right) \\
&\leq \rho'_{max}(b_1 - b_2) + \beta\left(\frac{\rho'_{max}(b_1 - b_2)}{1-\beta}\right) \\
&= \frac{\rho'_{max}(b_1 - b_2)}{1-\beta}
\end{aligned}
\tag{B.80}
$$

□

Lemma 20. *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* $\gamma = \max\{\Delta_p, \sum_{\omega} \Delta_{a\omega}\Delta_{e\omega}\}$*:*

$$
V_m(b_1) - V_m(b_2) \leq \rho'_{max}(1 + \beta\gamma\kappa)(b_1 - b_2) \tag{B.81}
$$

Proof. Using Equation 2.2, we get:

$$\frac{d(V_m(b))}{db} = \begin{cases} \rho'(b) + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))} \frac{d(\tau(b))}{db} & \text{passive} \\ \rho'(b) + \beta(V_m(P^a_{1,1}) - V_m(P^a_{0,1})) & \text{active} \end{cases} \tag{B.82}$$

Case 1 (passive):

$$= \rho'(b) + \beta \frac{d(V_m(\tau(b)))}{d(\tau(b))}(P^p_{1,1} - P^p_{0,1}) \tag{B.83}$$

$$= \rho'(b) + \beta \lim_{\delta \to 0} \frac{V_m(\tau(b) + \delta) - V_m(\tau(b))}{\tau(b) + \delta - \tau(b)}(P^p_{1,1} - P^p_{0,1}) \tag{B.84}$$

$$\leq \rho'(b) + \beta\kappa\rho'_{max}(P^p_{1,1} - P^p_{0,1}) \tag{B.85}$$

$$\leq \rho'_{max}(1 + \beta\gamma\kappa) \tag{B.86}$$

Case 2 (active):

$$= \rho'(b) + \beta(\sum_{\omega} \Theta'_{\omega} V_m(P^a_{\omega})) \tag{B.87}$$

$$= \rho'(b) + \beta(\sum_{\omega} \Theta'_{\omega}(V_m(P^a_{\omega}) - V_m(P^a_0))) \tag{B.88}$$

$$\leq \rho'(b) + \beta\kappa\rho'_{max}(\sum_{\omega} \Theta'_{\omega}(P^a_{\omega} - P^a_0)) \tag{B.89}$$

$$\leq \rho'_{max} + \beta\rho'_{max}\gamma\kappa \tag{B.90}$$

$$\leq \rho'_{max}(1 + \beta\gamma\kappa) \tag{B.91}$$

Thus,

$$\begin{aligned} &\implies \frac{d(V_m(b))}{db} \leq \rho'_{max}(1 + \beta\gamma\kappa) \\ &\implies \int_{b_2}^{b_1} \frac{d(V_m(b))}{db} db \leq \int_{b_2}^{b_1} \rho'_{max}\Big(1 + \beta\gamma\kappa\Big) db \\ &\implies V_m(b_1) - V_m(b_2) \leq \rho'_{max}\big(1 + \beta\gamma\kappa\big)(b_1 - b_2) \end{aligned} \tag{B.92}$$

□

Lemma 21. $V_m(b_1) - V_m(b_2) \leq \frac{\rho'_{max}(b_1-b_2)}{1-\beta\gamma} \forall\, b_1, b_2 \; s.t. \; b_1 \geq b_2$

Proof. The proof is same as the proof of Lemma 21. □

Corollary 7. $\frac{d(V_m(b))}{db} \leq \frac{\rho'_{max}}{1-\beta\gamma}$

Proof. This follows from Lemma 21 by setting $b_1 = b + \delta, b_2 = b$ under the limit $\delta \to 0$. □

Now we complete the proof for Thm.8 as follows:

$$Eq.2.11 \implies \Delta_p \geq \frac{\Delta_a(1-\beta\alpha)\rho'_{max}}{(1-\beta\gamma)\rho'_{min}} \tag{B.93}$$

$$\implies \frac{(P^p_{1,1} - P^p_{0,1})\rho'_{min}}{(1-\beta\alpha)} \geq \frac{\rho'_{max}(P^a_{1,1} - P^a_{0,1})}{(1-\beta\gamma)} \tag{B.94}$$

$$\implies \frac{(P^p_{1,1} - P^p_{0,1})\rho'_{min}}{(1-\beta\alpha)} \geq V_m(P^a_1) - V_m(P^a_0) \text{ by Lemma 21} \tag{B.95}$$

$$\implies (P^p_{1,1} - P^p_{0,1})\frac{d(V_m(b))}{db} \geq V_m(P^a_1) - V_m(P^a_0) \text{ by Cor. 6} \tag{B.96}$$

$$\implies \frac{d(\tau(b))}{db}\frac{d(V_m(b))}{db} \geq V_m(P^a_1) - V_m(P^a_0) \text{ by Fact 5} \tag{B.97}$$

$$\implies \rho'(b) + \beta\frac{d(V_m(\tau(b)))}{d(\tau b)}\frac{d(\tau(b))}{db} \geq \rho'(b) + \tag{B.98}$$

$$\beta\Big(V_m(P^a_1) - V_m(P^a_0)\Big) \tag{B.99}$$

$$\implies \frac{d(V_m(b|a=0))}{d(b)} \geq \frac{d(V_m(b|a=1))}{d(b)} \tag{B.100}$$

□

B.7 Proof of Theorem 9

Theorem 9 (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:*

$$\frac{\Delta_p(1-\beta\min\{\Delta_p, (\Delta_a.\Delta_c)\})}{\Delta_a(1-\beta\max\{\Delta_p, (\Delta_a.\Delta_c)\})} \leq \frac{\rho'_{min}}{\rho'_{max}} \tag{2.12}$$

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. 5 using the value function bounds for imprecise observations. □

C

Appendix to Chapter 3

C.1 Proofs

Theorem 19. *The belief state transition model for a 2-state Streaming Bandit arm with deterministic arrival time T_1 and departure time T_2 can be reduced to a belief state model for a restless bandit arm with $T_2 + (T_2 - T_1)^2$ states.*

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

probabilities are constructed as follows:

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

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

$$T_1 - 1 + 2(1 + \ldots + (T_2 - T_1)) + 1 \tag{C.1}$$

$$= T_1 + (T_2 - T_1)(T_2 - T_1 + 1) \tag{C.2}$$

$$= T_2 + (T_2 - T_1)^2 \tag{C.3}$$

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

Lemma 22. *If a forward (or reverse) threshold policy π is optimal for a subsidy m for the belief states MDP of the infinite horizon problem, then π is also optimal for the augmented belief state MDP.*

Proof. First, we define the value function for the modified belief states.

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

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

where b' is the next belief state.

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

$$V^{p}_{m_U}(\langle U,t\rangle) = V^{a}_{m_U}(\langle U,t\rangle) \tag{C.4}$$

$$\Rightarrow U + m_U + V_{m_U}(\langle b', t+1\rangle) = U + V_{m_U}(\langle b', t+1\rangle) \tag{C.5}$$

$$\Rightarrow m_U = 0. \tag{C.6}$$

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

Let us assume that the minimum subsidy $m_b = 0$ for a belief state $b \neq U$. Then,

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

$\because P^a_{x1} > P^p_{x1}$ and $V_m(b)$ is non-decreasing (Corollary 1 in [101]).

The last inequality contradicts the fact that $V_m(b)$ is a convex function of b [145]. Hence, the minimum subsidy required at any belief state $b \neq U$ to make the passive action more valuable is strictly greater than 0. □

Theorem 20. *A Streaming Bandits instance is indexable when there exists an optimal policy, for each arm and every value of $m \in \mathbb{R}$, that is forward (or reverse) threshold optimal policy.*

Proof. Using Theorem 1 and Lemma 1, it is straightforward to see that an optimal threshold policy for infinite horizon problem can be translated to a threshold policy for Streaming bandits instance. Moreover, using the fact that the existence of threshold policies for each subsidy m and each arm $i \in N$ is sufficient for indexability to hold (Theorem 1 of [101]), we show that the Streaming bandit problem is also indexable. □

Theorem 21 (Index Decay)**.** *Let $V^p_{m,T}(b)$ and $V^a_{m,T}(b)$ be the T-step passive and active value functions for a belief state b with passive subsidy m. Let m_T be the value of subsidy m, that satisfies the equation $V^p_{m,T}(b) = V^a_{m,T}(b)$ (i.e. m_T is the Whittle Index for a horizon T). Assuming indexability holds, we show that the Whittle index decays for short horizons: $\forall T > 1\colon\ m_T > m_1 > m_0 = 0$.*

Proof. We provide our argument for a more general reward criterion than the total reward introduced in Section 3.3. Consider a discounted reward criterion with discount factor $\beta \in [0, 1]$ (where $\beta = 1$ corresponds to total

reward). m_0 is simply the m that satisfies: $V^p_{m,0}(b) = V^a_{m,0}(b)$ i.e., $b + m = b$, thus $m_0 = 0$. Similarly, m_1 can be solved by equating $V^p_{m_1,1}(b)$ and $V^a_{m_1,1}(b)$ as follows:

$$\begin{aligned}
&\implies b + m_1 + \beta\big(bP^p_{11} + (1-b)P^p_{01}\big) = b + \beta\big(b(P^a_{11}) + (1-b)(P^a_{01})\big) \\
&\implies m_1 = \beta\big(b(P^a_{11} - P^p_{11}) + (1-b)(P^a_{01} - P^p_{01})\big) \qquad \text{(C.7)}
\end{aligned}$$

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

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

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

$$\begin{aligned}
V^p_{m_1,T}(b_0) &= b_0 + m_1 + \beta\rho^p(b_0,1) + \sum_{t=2}^{T}\beta^t\rho^p(b_0,t) \qquad \text{(C.8)} \\
&\qquad + \Big(\sum_{t=1}^{T}\beta^t m_1.1_{\{\pi^p(t)=passive\}}\Big) \\
&= b_0 + \beta\rho^a(b_0,1) + \sum_{t=2}^{T}\beta^t\rho^p(b_0,t) + \Big(\sum_{t=1}^{T}\beta^t m_1.1_{\{\pi^p(t)=passive\}}\Big) \\
&< b_0 + \beta\rho^a(b_0,1) + \sum_{t=2}^{T}\beta^t\rho^a(b_0,t) + \Big(\sum_{t=1}^{T}\beta^t m_1.1_{\{\pi^p(t)=passive\}}\Big) \qquad \text{(C.9)} \\
&\qquad\qquad \text{(by Lemma 2)} \\
&\leq V^a_{m_1,T}(b_0)
\end{aligned}$$

□

C.2 Robustness Checks

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

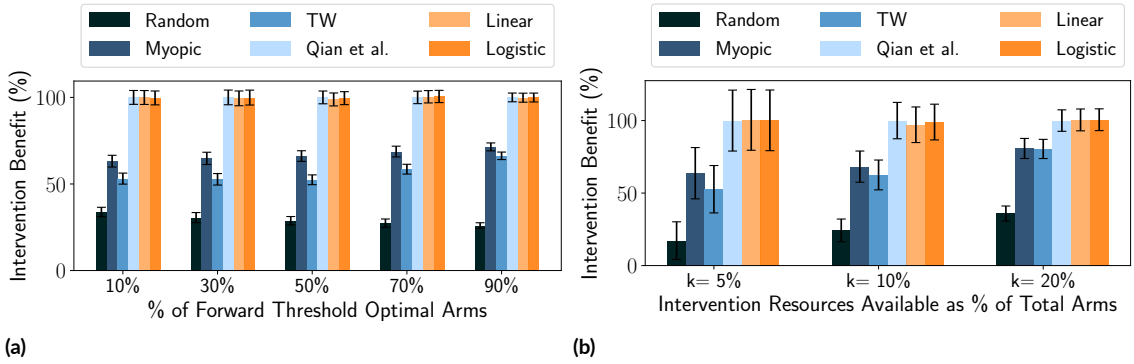


Figure C.1: (a)Performance of our algorithm remains robust even when population is composed of a varying fraction of forward threshold optimal arms (b) Performance of our algorithm remains robust under varying levels of available resources

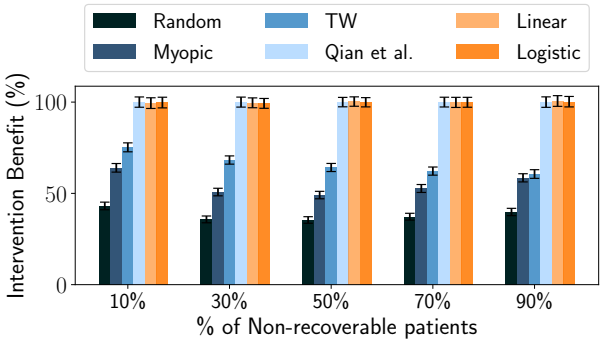


Figure C.2: Non-recoverable patients are those that remain in the bad state with high probability, even after receiving an intervention. Performance of Threshold Whittle begins to dwindle when the fraction of non-recoverable patients in the cohort increases, but our interpolation algorithms remain robust.

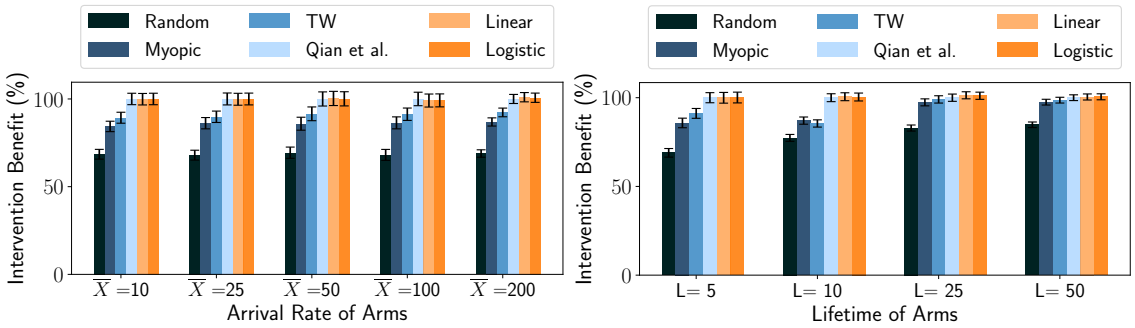


Figure C.3: Tests on the ARMMAN domain reveal that the large speedup is achieved while virtually maintaining the same good quality of performance

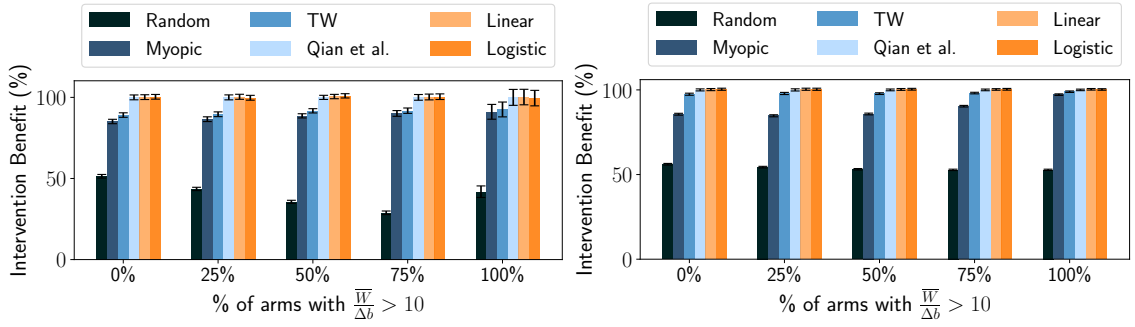


Figure C.4: We generate corner cases consisting of varying proportion of patients with high value of $\frac{\overline{W}}{\Delta b}$. We test the algorithms under two situations corresponding to lifetime of arms smaller/larger than the ratio $\frac{\overline{W}}{\Delta b}$ and find that our algorithm still show good performance throughout.

D

Appendix to Chapter 4

D.1 Clarification on Statistical Analysis

It has been brought to our attention that the standard statistical test employed in the paper, while being popular, is still imperfect for the unique nature of the adaptively collected data analyzed in our study. Designing a valid statistical test for concluding from adaptively collected data from trials such as ours, is still an area of modern research [51,174].

The unique challenge is that the outcomes of individual beneficiaries within an experimental arm, are not independent owing to the budget constraint. In fact, we expect the outcomes to be negatively correlated under the assumption that interventions are never detrimental — reason being that while allocating an intervention to a

beneficiary tends to improve their outcomes, it robs some other beneficiary of the intervention resource, leading to a decline in their expected outcome. This negative correlation is expected to translate to a lower sample variation in the measured outcomes than what is accounted for under the independent-outcomes assumption. Our discussion[67] suggests that this is likely to result in a less heavy tail than expected, and consequently, our computation should likely yield conservative p-values. We justify this intuition further via an illustrative example involving a simpler setting below.

Because the computed p-values are conservative, the conclusions drawn in our study should still remain valid.

D.1.1 Justification of why negative correlation leads to p-value inflation:

We provide intuition through an example, for why negative correlation leads to inflated p-values.

Consider a simplified linear model, similar to the one employed in the paper, without contributions from intermediate terms such as k, T_i, x_{ij}, retaining only β, ε. Consider the null hypothesis $H_0 := \beta = 0$. To reject this hypothesis, consider employing the z-test that would construct a z-statistic as: z-stat $= \frac{\sqrt{n}\cdot\bar{Y}}{\sigma}$. If the Y_i's were assumed to be independent, we would obtain the p-value as α in terms of the tail distribution Z of a standard normal distribution as:

$$|\frac{\sqrt{n}\cdot\bar{Y}}{\sigma}| = Z_{1-\frac{\alpha}{2}}$$

However, in reality, since the Y_i's are not independent, but in fact, negatively correlated, the actual variance in $\bar{Y}$ is smaller. This is because the second term in the expansion of the variance of the mean outcomes, contributes negatively as in:

$$var(\bar{Y}) = \frac{1}{n^2}\Big[\sum var(Y_i) + \sum cov(Y_i, Y_j)\Big]$$

This leads to σ^2 being an inflated version of the actual variance and equivalently, this leads to an under-inflated value of $|\frac{\sqrt{n}\cdot\bar{Y}}{\sigma}|$. This translates to a larger mass on the right tail of the Z distribution, leading to a conservative (i.e. larger) p-value.

E

Appendix to Chapter 7

E.1 Complete Proofs to Theoretical Results

E.1.1 Proof of Lemma 23

Lemma 23. *The relation † is an equivalence relation and the family of sets defined by $\mathcal{C}^{\dagger}(\cdot)$ forms a partition over $\mathcal{C}$.*

Proof. To prove that † is an equivalence relation, we show that it is reflexive, symmetric and transitive. † is reflexive because $\forall \mathfrak{C} \in \mathcal{C}, \mathfrak{C} \in \mathcal{C}^{\dagger}(\mathfrak{C})$ by definition. Furthermore, † is also trivially symmetric because if $\mathfrak{C}_2 \in \mathcal{C}^{\dagger}(\mathfrak{C}_1)$, then by definition, the allocations received by all individuals at all times are identical under both $\mathfrak{C}_1$ and $\mathfrak{C}_2$. Hence $\mathfrak{C}_1 \in \mathcal{C}^{\dagger}(\mathfrak{C}_2)$. Finally, † is also transitive because if all allocations received by all individuals at all times

are identical in $\mathfrak{C}_1$ and $\mathfrak{C}_2$ as well as in $\mathfrak{C}_2$ and $\mathfrak{C}_3$, that means the allocations are also identical in $\mathfrak{C}_1$ and $\mathfrak{C}_3$. Thus formally, if $\mathfrak{C}_2 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$ and $\mathfrak{C}_3 \in \mathcal{C}^\dagger(\mathfrak{C}_2)$, then $\mathfrak{C}_3 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$. Thus $\dagger$ is an equivalence relation over $\mathcal{C}$ and consequently, partitions $\mathcal{C}$ into a family of equivalence classes $\mathcal{C}^\dagger(\cdot)$ such that every element $\mathfrak{C} \in \mathcal{C}$ lies in exactly one partition [39]. □

E.1.2 Proof of Theorem 14

Theorem 14. *$Eval^\dagger(\pi_m)$ is an unbiased estimate of the expected value of the performance, $Eval^*(\pi_m)\ \forall m \in [M]$, defined in equation 7.1. i.e.*

$$\mathbb{E}_{S_m \sim P^*} \mathbb{E}_{\mathfrak{C} \sim \mathcal{C}}[Eval^\dagger(\pi_m)] = Eval^*(\pi_m)\ \forall m \in [M]$$

Proof.

$$\begin{aligned}
\text{Eval}^*(\pi_m) =& \mathbb{E}_{S_m \sim P^*}\Big[\mathbb{E}_{\mathfrak{C} \sim \mathcal{C}}\big[\text{Eval}(\pi_m)\big]\Big] \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\sum_{\mathfrak{C} \in \mathcal{C}} \text{Prob}[\mathfrak{C}] \cdot \text{Eval}(\pi_m|\mathfrak{C})\Big] \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\frac{1}{|\mathcal{C}|}\sum_{\mathfrak{C} \in \mathcal{C}} \text{Eval}(\pi_m|\mathfrak{C})\Big] && (\because \text{all } \mathfrak{C} \text{ equally likely}) \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\frac{1}{|\mathcal{C}|}\Big[\sum_{\mathfrak{C} \in \mathcal{P}_1} \text{Eval}(\pi_m|\mathfrak{C}) + \cdots + \sum_{\mathfrak{C} \in \mathcal{P}_\eta} \text{Eval}(\pi_m|\mathfrak{C})\Big]\Big] \\
& \text{where } \{\mathcal{P}_1, \ldots, \mathcal{P}_\eta\} \text{ defines partition of } \mathcal{C} \text{ induced by } \dagger. \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\sum_{j \in [\eta]} \frac{|\mathcal{P}_j|}{|\mathcal{C}|} \cdot \frac{1}{|\mathcal{P}_j|}.\Big[\sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Eval}(\pi_m|\mathfrak{C})\Big]\Big] \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\sum_{j \in [\eta]} \frac{|\mathcal{P}_j|}{|\mathcal{C}|} \cdot \Big[\text{Eval}^\dagger(\pi_m|\mathfrak{C})\Big]\Big] && (\forall \mathfrak{C} \in \mathcal{P}_j) \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\sum_{j \in [\eta]} \text{Prob}[\mathcal{P}_j] \cdot \Big[\text{Eval}^\dagger(\pi_m|\mathfrak{C})\Big]\Big] && (\forall \mathfrak{C} \in \mathcal{P}_j) \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\sum_{j \in [\eta]} \sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Prob}[\mathfrak{C}] \cdot \Big[\text{Eval}^\dagger(\pi_m|\mathfrak{C})\Big]\Big] && \Big(\because \text{Prob}[\mathcal{P}_j] = \sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Prob}[\mathfrak{C}]\Big) \\
=& \mathbb{E}_{S_m \sim P^*} \mathbb{E}_{\mathfrak{C} \sim \mathcal{C}}[\text{Eval}^\dagger(\pi_m)]
\end{aligned}$$

□

E.1.3 Proof of Theorem 15

Theorem 15. *The sample variance of our estimator, $Eval^{\dagger}(\pi)$ is smaller than the standard estimator, $Eval(\pi)$:*

$$(Eval(\pi)) - (Eval^{\dagger}(\pi)) =$$
$$\frac{1}{|\mathcal{C}|} \cdot \sum_{j \in [\eta]} \left[\sum_{\mathfrak{C} \in \mathcal{P}_j} Eval^2(\pi|\mathfrak{C}) - \frac{\left(\sum_{\mathfrak{C} \in \mathcal{P}_j} Eval(\pi|\mathfrak{C}) \right)^2}{|\mathcal{P}_j|} \right]$$

≥ 0, *where* $\{\mathcal{P}_1, \ldots, \mathcal{P}_\eta\}$ *is the partition of* $\mathcal{C}$ *induced by* $\dagger$.

Proof. We compute the sample variance by first conditioning over the partition $\mathcal{P}_j$ (of the equivalence sets defined by $\dagger$) that an instance of an assignment, $\mathfrak{C}$ belongs to and then accounting for the variance stemming from the candidate assignments $\mathfrak{C}$ within the partition. Thus we get:

$$\begin{aligned}
(\text{Eval}(\pi)) &= \frac{1}{|\mathcal{C}|} \sum_{\mathfrak{C} \in \mathcal{C}} \left(\text{Eval}(\pi|\mathfrak{C}) - \text{Eval}^*(\pi) \right)^2 \\
&= \frac{1}{|\mathcal{C}|} \sum_{j \in [\eta]} \sum_{\mathfrak{C} \in \mathcal{P}_j} \left(\text{Eval}(\pi|\mathfrak{C}) - \text{Eval}^*(\pi) \right)^2 \\
&= \frac{1}{|\mathcal{C}|} \sum_{j \in [\eta]} \left(\sum_{\mathfrak{C} \in \mathcal{P}_j} \left(\text{Eval}(\pi|\mathfrak{C}) \right)^2 - 2\, \text{Eval}^*(\pi) \sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Eval}(\pi|\mathfrak{C}) + |\mathcal{P}_j| \left(\text{Eval}^*(\pi) \right)^2 \right) \qquad \text{(E.1)}
\end{aligned}$$

Similarly, we compute the variance of our estimator $\text{Eval}^{\dagger}$ as:

$$\begin{aligned}
(\text{Eval}^{\dagger}(\pi)) &= \frac{1}{|\mathcal{C}|} \sum_{j \in [\eta]} \sum_{\mathfrak{C} \in \mathcal{P}_j} \left(\text{Eval}^{\dagger}(\pi|\mathfrak{C}) - \text{Eval}^*(\pi) \right)^2 \\
&= \frac{1}{|\mathcal{C}|} \sum_{j \in [\eta]} \left(|\mathcal{P}_j| \cdot \left\{ \frac{\sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Eval}(\pi|\mathfrak{C})}{|\mathcal{P}_j|} - \text{Eval}^*(\pi) \right\}^2 \right) \\
&= \frac{1}{|\mathcal{C}|} \sum_{j \in [\eta]} \left(|\mathcal{P}_j| \cdot \left\{ \left(\frac{\sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Eval}(\pi|\mathfrak{C})}{|\mathcal{P}_j|} \right)^2 - 2\, \text{Eval}^*(\pi) \left(\frac{\sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Eval}(\pi|\mathfrak{C})}{|\mathcal{P}_j|} \right) + \left(\text{Eval}^*(\pi) \right)^2 \right\} \right) \\
&= \frac{1}{|\mathcal{C}|} \sum_{j \in [\eta]} \left\{ \frac{\left(\sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Eval}(\pi|\mathfrak{C}) \right)^2}{|\mathcal{P}_j|} - 2\, \text{Eval}^*(\pi) \sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Eval}(\pi|\mathfrak{C}) + |\mathcal{P}_j| \left(\text{Eval}^*(\pi) \right)^2 \right\} \qquad \text{(E.2)}
\end{aligned}$$

Subtracting expression in Equation E.2 from the expression in Equation E.1 gives:

$$(\text{Eval}(\pi)) - (\text{Eval}^{\dagger}(\pi)) = \frac{1}{|\mathcal{C}|} \sum_{j \in [\eta]} \left(\sum_{\mathfrak{C} \in \mathcal{P}_j} \left(\text{Eval}(\pi|\mathfrak{C})\right)^2 - \frac{\left(\sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Eval}(\pi|\mathfrak{C})\right)^2}{|\mathcal{P}_j|} \right) \tag{E.3}$$

We can show that the expression for variance contraction derived in Equation E.3 is non-negative as a direct consequence of the Cauchy-Schwarz inequality. The Cauchy-Schwarz inequality states that for two vectors $\mathbf{u}$ and $\mathbf{v}$, $|\langle \mathbf{u}, \mathbf{v} \rangle|^2 \leq \langle \mathbf{u}, \mathbf{u} \rangle \cdot \langle \mathbf{v}, \mathbf{v} \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the inner product. Setting $\mathbf{u} = \underbrace{\left[\text{Eval}(\pi|\mathfrak{C}_1), \ldots, \text{Eval}(\pi|\mathfrak{C}_{|\mathcal{P}_j|})\right]}_{|\mathcal{P}_j| \text{entries}})$ and $\mathbf{v} = \underbrace{\left[1, \ldots, 1\right]}_{|\mathcal{P}_j| \; 1's}$ yields the desired result: $(\text{Eval}(\pi)) - (\text{Eval}^{\dagger}(\pi)) \geq 0$. □

E.1.4 Efficient Algorithm

Lemma 24. *The relation $\dagger_\Upsilon$ is an equivalence relation over both $\mathcal{C}$ as well as each set in the family $\mathcal{C}^{\dagger}(\cdot)$ and the family of sets defined by $\mathcal{C}^{\dagger}_{\Upsilon}(\cdot)$ forms a partition over $\mathcal{C}$.*

Proof. Similar to Lemma 23, we prove that $\dagger_\Upsilon$ is an equivalence relation by showing that it is reflexive, symmetric and transitive. $\dagger_\Upsilon$ is reflexive because $\forall \mathfrak{C} \in \mathcal{C}, \mathfrak{C} \in \mathcal{C}^{\dagger}_{\Upsilon}(\mathfrak{C})$ by definition. Furthermore, $\dagger$ is also trivially symmetric because if $\mathfrak{C}_2 \in \mathcal{C}^{\dagger}(\mathfrak{C}_1)$, then by definition, the index thresholds, as well as the allocations received by all individuals at all times, are identical under both $\mathfrak{C}_1$ and $\mathfrak{C}_2$. Hence $\mathfrak{C}_1 \in \mathcal{C}^{\dagger}(\mathfrak{C}_2)$. Finally, $\dagger$ is also transitive because if all index threshold and allocations received by all individuals at all times are identical in $\mathfrak{C}_1$ and $\mathfrak{C}_2$ as well as in $\mathfrak{C}_2$ and $\mathfrak{C}_3$, that means the same are also identical in $\mathfrak{C}_1$ and $\mathfrak{C}_3$. Thus formally, if $\mathfrak{C}_2 \in \mathcal{C}^{\dagger}(\mathfrak{C}_1)$ and $\mathfrak{C}_3 \in \mathcal{C}^{\dagger}(\mathfrak{C}_2)$, then $\mathfrak{C}_3 \in \mathcal{C}^{\dagger}(\mathfrak{C}_1)$. Thus $\dagger_\Upsilon$ is an equivalence relation over $\mathcal{C}$ and consequently, partitions $\mathcal{C}$ into a family of equivalence classes $\mathcal{C}^{\dagger}(\cdot)$ such that every element $\mathfrak{C} \in \mathcal{C}$ lies in exactly one partition [39]. Similar reasoning also shows that $\dagger_\Upsilon$ is an equivalence relation over each set in the family $\mathcal{C}^{\dagger}(\cdot)$. □

Corollary 8. *$Eval^{\dagger}_{\Upsilon}(\pi_m)$ is an unbiased estimate of the expected value of the performance, $Eval^{*}(\pi)$, defined in equation 7.1. i.e. $\mathbb{E}_{S_m \sim P^*} \mathbb{E}_{\mathfrak{C} \sim \mathcal{C}}[Eval^{\dagger}(\pi_m)] = Eval^{*}(\pi_m) \; \forall m \in [M]$*

Proof. Using Lemma 24, we apply similar arguments as Theorem 14 on the parition defined by $\dagger_\Upsilon$ to show that $\text{Eval}^{\dagger}_{\Upsilon}$ yields an unbiased estimate. □

Theorem 16. *$Eval^{\dagger}_{\Upsilon}(\cdot)$ computed as per Equation 7.6 computes the average of $Eval(|\mathfrak{C})$ over all assignments in $\mathcal{C}^{\dagger}_{\Upsilon}$. i.e. $Eval^{\dagger}_{\Upsilon}(\pi) = \frac{\sum_{\mathfrak{C} \in \mathcal{C}^{\dagger}_{\Upsilon}} Eval(\pi|\mathfrak{C})}{|\mathcal{C}^{\dagger}_{\Upsilon}|}$*

Proof. The key to showing that the two are equivalent is in interpreting the summation of (modified) rewards over individuals in the view of average group rewards over assignments. Mathematically, starting from the definition of $\text{Eval}^{\dagger}_{\Upsilon}(\pi_m)$, the key lies in moving the summation operation over assignments in $\mathcal{C}^{\dagger}_{\Upsilon}$ from outside the $\text{Eval}^{\dagger}_{\Upsilon}()$ term to inside, applying it individually on each contributing participant. Formally, we can rewrite the

expression of $\text{Eval}^{\dagger}_{\Upsilon}(\pi_m)$ as:

$$\text{Eval}^{\dagger}_{\Upsilon}(\pi_j) = \frac{\sum_{\mathfrak{C}\in\mathcal{C}^{\dagger}_{\Upsilon}} \text{Eval}(\pi_j|\mathfrak{C})}{|\mathcal{C}^{\dagger}_{\Upsilon}|} \tag{E.4}$$

$$= \frac{\sum_{\mathfrak{C}\in\mathcal{C}^{\dagger}_{\Upsilon}} \sum_{i\in C_j} r(S[i], A[i])}{|\mathcal{C}^{\dagger}_{\Upsilon}|} \tag{E.5}$$

$$= \sum_{\mathfrak{C}\in\mathcal{C}^{\dagger}_{\Upsilon}} \frac{\sum_{k\in\kappa} \sum_{i\in\mathbf{G}_k} 1_{\{i\in C_j\}} \cdot r(S[i], A[i]) + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i])}{|\mathcal{C}^{\dagger}_{\Upsilon}|} \tag{E.6}$$

(splitting the summation over groups $\mathbf{G}_k$ and other individuals that can't be swapped) (E.7)

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\sum_{\mathfrak{C}\in\mathcal{C}^{\dagger}_{\Upsilon}} \frac{1_{\{i\in C_j\}}\cdot r(S[i], A[i])}{|\mathcal{C}^{\dagger}_{\Upsilon}|}\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{E.8}$$

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\sum_{\mathfrak{C}\in\mathcal{C}^{\dagger}_{\Upsilon}} \frac{1_{\{i\in C_j\}}}{|\mathcal{C}^{\dagger}_{\Upsilon}|}\cdot r(S[i], A[i])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{E.9}$$

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\Pr(i\in C_j|\mathcal{C}^{\dagger}_{\Upsilon})\cdot r(S[i], A[i])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{E.10}$$

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\frac{|\{\iota : \iota\in(\mathbf{G}_{\varphi(i)}\cap C_j)\}|}{|\mathbf{G}_{\varphi(i)}|}\cdot r(S[i,0:T], A[i,1:T])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{E.11}$$

$$= \sum_{k\in\kappa}|\{\iota : \iota\in(\mathbf{G}_{\varphi(i)}\cap C_j)\}|\cdot\sum_{i\in\mathbf{G}_k}\left[\frac{1}{|\mathbf{G}_{\varphi(i)}|}\cdot r(S[i,0:T], A[i,1:T])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{E.12}$$

$$= \sum_{k\in\kappa}|\{\iota : \iota\in(\mathbf{G}_{\varphi(i)}\cap C_j)\}|\cdot\tilde{r}_k + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{E.13}$$

$$= \sum_{i\in C_j}\tilde{r}_{\varphi(i)}\cdot\Lambda_i + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i], 1:T) \tag{E.14}$$

□

Table E.1: Sample variance in Measured Performance Lift

T	B	π_1 v π_0	RAW	PERMUTED	IPW	n-VAL
1	3%	π_{WI} v π_{GR}	49.09	4.94	**0.48**	9
1	10%	π_{WI} v π_{GR}	49.86	15.11	**6.66**	3
1	25%	π_{WI} v π_{GR}	49.45	**19.94**	78.12	2
10	3%	π_{WI} v π_{WI}	2381	**916**	NA	3
10	3%	π_{WI} v π_{GR}	2348	**728**	NA	4
10	3%	π_{GR} v π_{GR}	26356	**1860**	NA	13
10	10%	π_{GR} v π_{GR}	25983	**3808**	NA	7
10	25%	π_{GR} v π_{GR}	23619	**5477**	NA	5

E.2 Casting Resource Allocation Policies as Index Policies

Control: A control group that sees no interventions can be handled by using any randomly generated index matrix with finite entries. Setting the index threshold $\Upsilon_i = \infty \, \forall \, i$ ensures that no individual assigned the control policy gets picked for intervention.

Round Robin: Common policies such as 'round robin', that operate by selecting individuals cyclically for intervention in a set order can also be represented as index policies. The index for each individual at each time step, can be determined in two stages. First, we consider the feature used for ranking the N individuals and we start by setting $\Upsilon_i(t) := r$, for $r \in \{1, \ldots N\}$ where r denotes the priority rank of the individual (highest rank picked first). Next, each time an individual receives an action $a = 1$, we want to push them at the bottom of the queue, so we subtract N from their index for all future timesteps, repeating this process for each instance of $a = 1$.

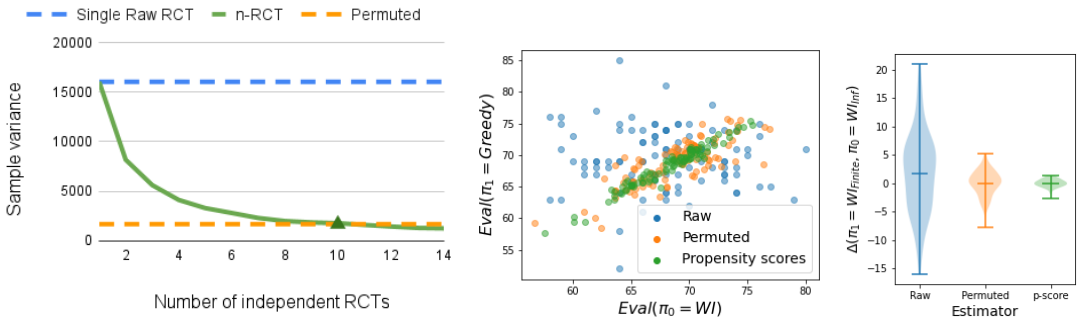


Figure E.2: (Left:) Variance reduces by running and averaging over n-independently run trials. (Center, Right:) Single shot setup

E.3 Additional Experimental Results

E.3.1 Synthetic Data Generation in Section 7.6.1 Explained

$$P^{p,1} = \begin{bmatrix} 0.97 & 0.03 \\ 0.03 & 0.97 \end{bmatrix} \quad P^{a,1} = \begin{bmatrix} 0.96 & 0.04 \\ 0.01 & 0.99 \end{bmatrix}$$
$$P^{p,2} = \begin{bmatrix} 0.25 & 0.75 \\ 0.03 & 0.97 \end{bmatrix} \quad P^{a,2} = \begin{bmatrix} 0.23 & 0.77 \\ 0.01 & 0.99 \end{bmatrix}$$

Figure E.1: Probability values forming the matrix P_1 and P_2

We reproduce the transition probabilities P_1 and P_2 used in our simulation, adopted from [101] in Figure E.1. Each P comprises of a set of probabilities under each of the two actions ($a = 0$, denoted as 'p', for passive and $a = 1$ denoted as 'a', for active).

Intuition is that P_1 has a very small $P^a_{0,1}$ and $P^p_{0,1}$ and is thus difficult to revive once it enters state $s = 0$, even with an intervention ($a = 1$), making it important to keep intervening to stop the individual from ever entering $s = 0$. On the other hand, P_2 has a large $P^p_{0,1}$, making it self-correcting, meaning the individual is likely to return to $s = 1$ quickly even without intervention.

E.3.2 Single-shot RCTs

E.3.3 Sequential RCTs

We run more comparisons using $N = 100$ individuals per arm, simulating 500 instances of trials for $T = 10$ timesteps. The $n-$values are listed in Table E.1.

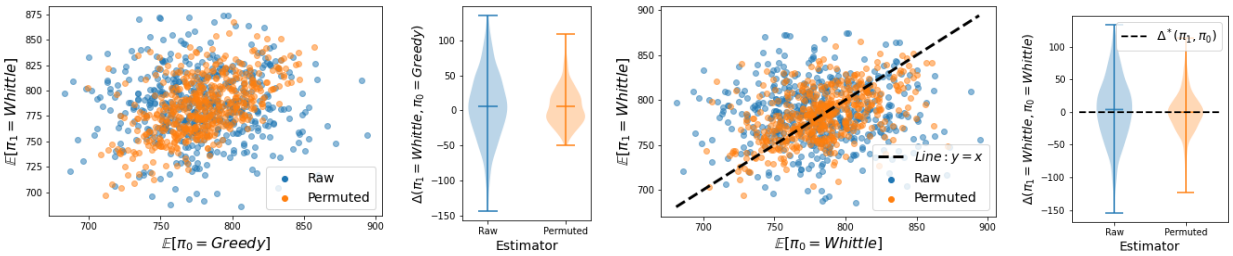


Figure E.3: Whittle vs Greedy (left two panels) and Whittle vs whittle (right two panels)

SETUP: GREEDY VS GREEDY For more granular analysis, we consider the state trajectories of individuals participating in the trial. This uses $N = 1000$ individuals per arm and simulates 30 instances of trials for $T = 10$ timesteps. We see that orange trajectories (after permutation) is closer to the expected value than blue trajectories (blue)

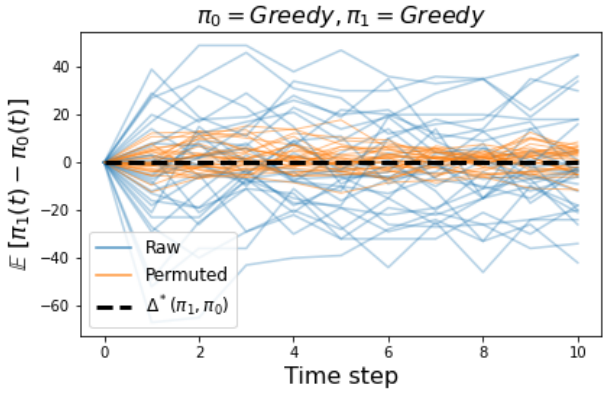


Figure E.4: Time trajectories of Greedy v Greedy

References

[1] Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2022). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1), 1–35.

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

[3] Akbarzadeh, N. & Mahajan, A. (2019a). Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *IEEE Conference on Decision and Control*.

[4] Akbarzadeh, N. & Mahajan, A. (2019b). Restless bandits with controlled restarts: Indexability and computation of whittle index. In *2019 IEEE Conference on Decision and Control*: IEEE.

[5] Akbarzadeh, N. & Mahajan, A. (2020). Conditions for indexability of restless bandits and an o (k^ 3) algorithm to compute whittle index. *arXiv preprint arXiv:2008.06111*.

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

[7] Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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

[9] ARMMAN (2008). About armman. https://armman.org/about-us. Accessed: 2022-08-12.

[10] ARMMAN (2019). Assessing the impact of mobile-based intervention on health literacy among pregnant women in urban india. https://armman.org/wp-content/uploads/2019/09/Sion-Study-Abstract.pdf. Accessed: 2022-08-12.

[11] ARMMAN (2020a). ARMMAN helping mothers and children: mMitra. https://armman.org/mmitra/.

[12] ARMMAN (2020b). mmitra. https://armman.org/mmitra/.

[13] Athey, S. & Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1), 133–161.

[14] Austin, P. C. & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28), 3661–3679.

[15] Avrachenkov, K. & Borkar, V. S. (2020). Whittle index based q-learning for restless bandits with average reward. *arXiv preprint arXiv:2004.14427*.

[16] Ayer, T., Zhang, C., Bonifonte, A., Spaulding, A. C., & Chhatwal, J. (2019). Prioritizing hepatitis c treatment in us prisons. *Operations Research*, 67(3), 853–873.

[17] Besbes, O., Gur, Y., & Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27.

[18] Bhattacharya, B. (2018). Restless bandits visiting villages: A preliminary study on distributing public health services. In *COMPASS*.

[19] Bian, J., Tian, D., Tang, Y., & Tao, D. (2018). A survey on trajectory clustering analysis. *arXiv preprint arXiv:1802.06971*.

[20] Biele, G., Erev, I., & Ert, E. (2009). Learning, risk attitude and hot stoves in restless bandit problems. *Journal of mathematical psychology*, 53(3), 155–167.

[21] Biswas, A., Aggarwal, G., Varakantham, P., & Tambe, M. (2021a). Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. *arXiv preprint arXiv:2105.07965*.

[22] Biswas, A., Aggarwal, G., Varakantham, P., & Tambe, M. (2021b). Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. In Z. Zhou (Ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021* (pp. 4039–4046).: ijcai.org.

[23] Biswas, A., Aggarwal, G., Varakantham, P., & Tambe, M. (2021c). Learning index policies for restless bandits with application to maternal healthcare. *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*.

[24] Biswas, A., Jain, S., Mandal, D., & Narahari, Y. (2015). A truthful budget feasible multi-armed bandit mechanism for crowdsourcing time critical tasks. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems* (pp. 1101–1109).

[25] Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C., & Shi, W. (2018). Federated learning of predictive models from federated electronic health records. *International journal of medical informatics*, 112, 59–67.

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

[27] Brownstein, J. N., Chowdhury, F. M., Norris, S. L., Horsley, T., Jack Jr, L., Zhang, X., & Satterfield, D. (2007b). Effectiveness of community health workers in the care of people with hypertension. *American journal of preventive medicine*, 32(5), 435–447.

[28] Chang, A. H., Polesky, A., & Bhatia, G. (2013a). 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.

[29] Chang, A. H., Polesky, A., & Bhatia, G. (2013b). 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.

[30] Chen, R. & Paschalidis, I. (2018). Learning optimal personalized treatment rules using robust regression informed K-NN. In *NIPS Machine Learning for Health Workshop*.

[31] Chen, R., Santo, K., Wong, G., Sohn, W., Spallek, H., Chow, C., & Irving, M. (2021). Mobile apps for dental caries prevention: Systematic search and quality evaluation. *JMIR mHealth and uHealth*, 9.

[32] Christopher, J. B., Le May, A., Lewin, S., & Ross, D. A. (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).

[33] Corotto, P. S., McCarey, M. M., Adams, S., Khazanie, P., & Whellan, D. J. (2013). Heart failure patient adherence: epidemiology, cause, and treatment. *Heart failure clinics*, 9(1), 49–58.

[34] Dahiya, K. & Bhatia, S. (2015). Customer churn analysis in telecom industry. In *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)* (pp. 1–6).: IEEE.

[35] Dil, Y., Strachan, D., Cairncross, S., Korkor, A., & Hill, Z. (2012). Motivations and challenges of community-based surveillance volunteers in the northern region of Ghana. *J. Commun. Health*, 37(6), 1192–1198.

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

[37] Eagle, N., Macy, M., & Claxton, R. (2010). Network diversity and economic development. *Science*, 328(5981), 1029–1031.

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

[39] Enderton, H. B. (1977). *Elements of set theory*. Academic press.

[40] Eysenbach, G. (2005). The law of attrition. *J Med Internet Res*, 7(1), e11.

[41] Floridi, L., Cowls, J., King, T., & Taddeo, M. (2020). How to design ai for social good: Seven essential factors. *Science and Engineering Ethics*, 26.

[42] Galstyan, A., Czajkowski, K., & Lerman, K. (2004). Resource allocation in the grid using reinforcement learning. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, volume 1 (pp. 1314–1315).: IEEE Computer Society.

[43] Gentile, C., Li, S., & Zappella, G. (2014). Online clustering of bandits. In *International Conference on Machine Learning* (pp. 757–765).: PMLR.

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

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

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

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

[48] Google (2018). Artificial intelligence at google: Our principles. https://ai.google/principles. Accessed: 2022-08-12.

[49] Green, B. & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 90–99).

[50] Gupta, K., Roy, S., Poonia, R., Nayak, S. R., Kumar, R., Alzahrani, K., Alnefaie, M., & Al-Wesabi, F. (2022). Evaluating the usability of mhealth applications on type 2 diabetes mellitus using various mcdm models. *Healthcare*, 10, 4.

[51] Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., & Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15), e2014602118.

[52] Haines, A., Sanders, D., Lehmann, U., Rowe, A. K., Lawn, J. E., Jan, S., Walker, D. G., & Bhutta, Z. (2007). Achieving child survival goals: potential contribution of community health workers. *The lancet*, 369(9579), 2121–2131.

[53] Hawkins, J. T. (2003). *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. PhD thesis, Massachusetts Institute of Technology.

[54] Hayes, R. J. & Moulton, L. H. (2017). *Cluster randomised trials*. Chapman and Hall/CRC.

[55] Health, U. M. (2020). UNFPA united nations population fund. https://www.unfpa.org/maternal-health#readmore-expand.

[56] HelpMum (2021). Preventing maternal and infant mortality in nigeria. https://helpmum.org/.

[57] Hernán, M. A. & Robins, J. M. (2010). Causal inference.

[58] Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 159–166).

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

[60] Hsu, Y.-P. (2018b). Age of information: Whittle index for scheduling stochastic arrivals. In *IEEE International Symposium on Information Theory*.

[61] Hu, W. & Frazier, P. (2017). An asymptotically optimal index policy for finite-horizon restless bandits.

[62] Hudgens, M. G. & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482), 832–842.

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

[64] Imai, K. & Li, M. L. (2021). Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association*, (pp. 1–15).

[65] Jakob, R., Harperink, S., Rudolf, A. M., Fleisch, E., Haug, S., Mair, J. L., Salamanca-Sanabria, A., & Kowatsch, T. (2022). Factors influencing adherence to mhealth apps for prevention or management of noncommunicable diseases: Systematic review. *J Med Internet Res*, 24(5), e35371.

[66] Jamieson, K., Malloy, M., Nowak, R., & Bubeck, S. (2014). lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory* (pp. 423–439).: PMLR.

[67] Janson, L. (2022). Private communication.

[68] Jiang, N. & Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning* (pp. 652–661).: PMLR.

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

[70] Johnson, J. . (2017). Momconnect: Connecting women to care, one text at a time. https://www.jnj.com/our-giving/momconnect-connecting-women-to-care-one-text-at-a-time.

[71] Jung, Y. H. & Tewari, A. (2019a). Regret bounds for thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems*.

[72] Jung, Y. H. & Tewari, A. (2019b). Regret bounds for Thompson sampling in episodic restless bandit problems. In *NeurIPS*.

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

[74] Kanade, V., McMahan, H. B., & Bryan, B. (2009). Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *Artificial Intelligence and Statistics* (pp. 272–279).: PMLR.

[75] Kaur, J., Kaur, M., Chakrapani, V., Webster, J., Santos, J., & Kumar, R. (2020). Effectiveness of information technology–enabled 'smart eating' health promotion intervention: A cluster randomized controlled trial. *PLOS ONE*, 15, e0225892.

[76] Kenya, S., Chida, N., Symes, S., & Shor-Posner, G. (2011). 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.

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

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

[79] Killian, J., Wilder, B., Sharma, A., Choudhary, V., Dilkina, B., & Tambe, M. (2019a). Learning to prescribe interventions for tuberculosis patients using digital adherence data. In *KDD*.

[80] Killian, J. A., Biswas, A., Shah, S., & Tambe, M. (2021a). Q-learning lagrange policies for multi-action restless bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 871–881).

[81] Killian, J. A., Perrault, A., & Tambe, M. (2021b). 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)*.

[82] Killian, J. A., Wilder, B., Sharma, A., Choudhary, V., Dilkina, B., & Tambe, M. (2019b). Learning to prescribe interventions for tuberculosis patients using digital adherence data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[83] Kleinberg, R., Niculescu-Mizil, A., & Sharma, Y. (2010). Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2), 245–272.

[84] Kuss, O., Blettner, M., & Börgermann, J. (2016). Propensity score: An alternative method of analyzing treatment effects: Part 23 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 113(35-36), 597.

[85] Le Ny, J., Dahleh, M., & Feron, E. (2008). Multi-uav dynamic routing with partial observations using restless bandit allocation indices. In *2008 American Control Conference* (pp. 4220–4225).: IEEE.

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

[87] Leskovec, J., Adamic, L. A., & H., B. A. (2007). The dynamics of viral marketing. *ACM Trans. Web*, 1(1), 5–es.

[88] Li, C., Wu, Q., & Wang, H. (2021). Unifying clustered and non-stationary bandits. In *International Conference on Artificial Intelligence and Statistics* (pp. 1063–1071).: PMLR.

[89] Li, L., Munos, R., & Szepesvári, C. (2015). Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics* (pp. 608–616).: PMLR.

[90] Li, S., Chen, W., & Leung, K.-S. (2019). Improved algorithm on online clustering of bandits. *arXiv preprint arXiv:1902.09162*.

[91] Liao, P., Greenewald, K., Klasnja, P., & Murphy, S. (2019). Personalized HeartSteps: A reinforcement learning algorithm for optimizing physical activity. In *JSM*.

[92] Liao, P., Greenewald, K., Klasnja, P., & Murphy, S. (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1), 1–22.

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

[94] Little, J. D. (1961). A proof for the queuing formula: L= λ w. *Operations research*, 9(3), 383–387.

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

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

[97] Löwe, B., Unützer, J., Callahan, C., Perkins, A., & Kroenke, K. (2004a). Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical Care*, (pp. 1194–1201).

[98] Löwe, B., Unützer, J., Callahan, C. M., Perkins, A. J., & Kroenke, K. (2004b). Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical care*, (pp. 1194–1201).

[99] Martin, L. R., Williams, S. L., Haskard, K. B., & DiMatteo, M. R. (2005). The challenge of patient adherence. *Therapeutics and clinical risk management*, 1(3), 189.

[100] Mate, A., Biswas, A., Siebenbrunner, C., & Tambe, M. (2022a). Efficient algorithms for finite horizon and streaming restless multi-armed bandit problems. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

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

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

[103] Mate, A., Madaan, L., Taneja, A., Madhiwalla, N., Verma, S., Singh, G., Hegde, A., Varakantham, P., & Tambe, M. (2022c). Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36 (pp. 12017–12025).

[104] Mate, A., Perrault, A., & Tambe, M. (2021a). Risk-aware interventions in public health: Planning with restless multi-armed bandits. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

[105] Mate, A., Perrault, A., & Tambe, M. (2021b). Risk-aware interventions in public health: Planning with restless multi-armed bandits. In *International Conference on Autonomous Agents and Multiagent Systems*.

[106] Mate, A., Wilder, B., Taneja, A., & Tambe, M. (2023). Improved policy evaluation for randomized trials of algorithmic resource allocation. *International Conference on Machine Learning (ICML) 2023, Honolulu, Hawaii, USA*.

[107] Meh, C., Sharma, A., Ram, U., Fadel, S., Correa, N., Snelgrove, J., Shah, P., Begum, R., Shah, M., Hana, T., Fu, H., Raveendran, L., Mishra, B., & Jha, P. (2022). Trends in maternal mortality in india over two decades in nationally-representative surveys. *BJOG: An International Journal of Obstetrics & Gynaecology*, 129.

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

[109] Meuleau, N., Hauskrecht, M., Kim, K.-E., Peshkin, L., Kaelbling, L. P., Dean, T. L., & Boutilier, C. (1998). Solving very large weakly coupled markov decision processes. In *AAAI/IAAI* (pp. 165–172).

[110] Meyerowitz-Katz, G., Ravi, S., Arnolda, L., Feng, X., Maberly, G., & Astell-Burt, T. (2020). Rates of attrition and dropout in app-based interventions for chronic disease: Systematic review and meta-analysis. *J Med Internet Res*, 22(9), e20283.

[111] Mintz, Y., Aswani, A., K., P., Flowers, E., & Fukuoka, Y. (2020). Nonstationary bandits with habituation and recovery dynamics. *Operations Research*, 68(5), 1493–1516.

[112] Mukerjee, R., Dasgupta, T., & Rubin, D. B. (2018). Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, 113(522), 868–881.

[113] Mundorf, C., Shankar, A., Moran, T., Heller, S., Hassan, A., Harville, E., & Lichtveld, M. (2018a). 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.

[114] Mundorf, C., Shankar, A., Moran, T., Heller, S., Hassan, A., Harville, E., & Lichtveld, M. (2018b). 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.

[115] Nakhleh, K., Ganji, S., Hsieh, P.-C., Hou, I., Shakkottai, S., et al. (2021). Neurwin: Neural whittle index network for restless bandits via deep rl. *Advances in Neural Information Processing Systems*, 34.

[116] Newman, P., Franke, M., Arrieta, J., Carrasco, H., Elliott, P., Flores, H., Friedman, A., Graham, S., Martinez, L., Palazuelos, L., et al. (2018a). 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.

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

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

[119] Nishtala, S., Kamarthi, H., Thakkar, D., Narayanan, D., Grama, A., Padmanabhan, R., Madhiwalla, N., Chaudhary, S., Ravindra, B., & Tambe, M. (2020). Missed calls, automated calls and health support: Using ai to improve maternal health outcomes by increasing program engagement. *arXiv preprint arXiv:2006.07590*.

[120] NITI AAYOG (2021). Responsible ai: Approach document for india. https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf. Accessed: 2022-08-12.

[121] Organization, W. H. 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.

[122] Organization, W. H. et al. (2017). Tracking universal health coverage: 2017 global monitoring report.

[123] Ou, H.-C., Siebenbrunner, C., Killian, J., Brooks, M. B., Kempe, D., Vorobeychik, Y., & Tambe, M. (2022). Networked restless multi-armed bandits for mobile interventions. *arXiv preprint arXiv:2201.12408*.

[124] Papadimitriou, C. H. & Tsitsiklis, J. N. (1994). The complexity of optimal queueing network control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory* (pp. 318–322).: IEEE.

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

[126] Pfammatter, A., Spring, B., Saligram, N., Davé, R., Gowda, A., Blais, L., Arora, M., Ranjani, H., Ganda, O., Hedeker, D., Reddy, S., & Ramalingam, S. (2016). mhealth intervention to improve diabetes risk behaviors in india: A prospective, parallel group cohort study. *Journal of Medical Internet Research*, 18, e207.

[127] Pilote, L., Tulsky, J. P., Zolopa, A. R., Hahn, J. A., Schecter, G. F., & Moss, A. R. (1996). Tuberculosis Prophylaxis in the Homeless: A Trial to Improve Adherence to Referral. *Archives of Internal Medicine*, 156(2), 161–165.

[128] Pollack, M. E., Brown, L., Colbry, D., Orosz, C., Peintner, B., Ramakrishnan, S., Engberg, S., Matthews, J. T., Dunbar-Jacob, J., McCarthy, C. E., et al. (2002a). Pearl: A mobile robotic assistant for the elderly. In *AAAI workshop on automation as eldercare*, volume 2002: AAAI, 2002, Edmonton, Alberta, Canada.

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

[130] Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley.

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

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

[133] Qian, Y., Zhang, C., Krishnamachari, B., & Tambe, M. (2016b). Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In C. M. Jonker, S. Marsella, J. Thangarajah, & K. Tuyls (Eds.), *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016* (pp. 123–131).: ACM.

[134] Rahedi Ong'ang'o, J., Mwachari, C., Kipruto, H., & Karanja, S. (2014a). 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), e88937.

[135] Rahedi Ong'ang'o, J., Mwachari, C., Kipruto, H., & Karanja, S. (2014b). 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), e88937.

[136] Raj, V. & Kalyani, S. (2017). Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*.

[137] Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322–331.

[138] Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). Re-imagining algorithmic fairness in india and beyond. *CoRR*, abs/2101.09995.

[139] Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. (2012). A proposed churn prediction model. *International Journal of Engineering Research and Applications*, 2(4), 693–697.

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

[141] Sombabu, B., Mate, A., Manjunath, D., & Moharir, S. (2020a). Whittle index for aoi-aware scheduling. In *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)* (pp. 630–633).: IEEE.

[142] Sombabu, B., Mate, A., Manjunath, D., & Moharir, S. (2020b). Whittle index for AoI-aware scheduling. In *COMSNETS*: IEEE.

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

[144] Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S., & Lee, S.-K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare informatics research*, 16(4), 253–259.

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

[146] Sutton, R. S., Maei, H., & Szepesvári, C. (2008). A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. *Advances in neural information processing systems*, 21.

[147] Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., & Zitouni, I. (2017). Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30.

[148] Tesauro, G., Jong, N. K., Das, R., & Bennani, M. N. (2006). A hybrid reinforcement learning approach to autonomic resource allocation. In *2006 IEEE International Conference on Autonomic Computing* (pp. 65–73).: IEEE.

[149] Thirumurthy, H. & Lester, R. T. (2012). M-health for health behaviour change in resource-limited settings: applications to hiv care and beyond. *Bulletin of the World Health Organization*, 90, 390–392.

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

[151] Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D., Ezer, D., Haert, F., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake, M., Othman, M., Glasmachers, T., Wever, W., & Clopath, C. (2020). Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11, 2468.

[152] Tripathi, V. & Modiano, E. (2019). A whittle index approach to minimizing functions of age of information. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (pp. 1160–1167).: IEEE.

[153] Tshikomana, R. S. & Ramukumba, M. M. (2022). Implementation of mhealth applications in community-based health care: Insights from ward-based outreach teams in south africa. *PLOS ONE*, 17(1), 1–15.

[154] Tuldrà, A., Ferrer, M. J., Fumaz, C. R., Bayés, R., Paredes, R., Burger, D. M., & Clotet, B. (1999). Monitoring Adherence to HIV Therapy. *Archives of Internal Medicine*, 159(12), 1376–1377.

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

[156] Ustun, B. & Rudin, C. (2019). Learning optimized risk scores. *J. Mach. Learn. Res.*, 20(150), 1–75.

[157] Vaswani, S., Lakshmanan, L., Schmidt, M., et al. (2015). Influence maximization with bandits. *arXiv preprint arXiv:1503.00024*.

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

[159] Wang, Y.-X., Agarwal, A., & Dudık, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning* (pp. 3589–3597).: PMLR.

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

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

[162] Wells, K. J., Luque, J. S., Miladinovic, B., Vargas, N., Asvat, Y., Roetzheim, R. G., & Kumar, A. (2011). Do community health worker interventions improve rates of screening mammography in the United States? A systematic review. *Cancer Epidem. Biomar.*, 20(8), 1580–1598.

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

[164] Whittle, P. (1988b). Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, (pp. 287–298).

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

[166] Wilder, B., Onasch-Vera, L., Diguiseppi, G., Petering, R., Hill, C., Yadav, A., Rice, E., & Tambe, M. (2021). Clinical trial of an ai-augmented intervention for hiv prevention in youth experiencing homelessness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 (pp. 14948–14956).

[167] Witmer, A., Seifer, S. D., Finocchio, L., Leslie, J., & O'Neil, E. H. (1995). Community health workers: integral members of the health care work force. *American journal of public health*, 85(8_Pt_1), 1055–1058.

[168] World Bank, . (2020). *Poverty and shared prosperity 2020: Reversals of fortune*. The World Bank.

[169] World Health Organization (WHO) (2020). Newborns: improving survival and well-being. https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality#:~:text=Neonates,in%20child%20survival%20since%201990. Accessed: 2022-08-12.

[170] Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3, Part 1), 5445 – 5449.

[171] Xu, J., Chen, L., & Tang, O. (2020). An online algorithm for the risk-aware restless bandit. *European Journal of Operational Research*.

[172] Yang, L., Liu, B., Lin, L., Xia, F., Chen, K., & Yang, Q. (2020). Exploring clustering of bandits for online recommendation system. In *Fourteenth ACM Conference on Recommender Systems* (pp. 120–129).

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

[174] Zhang, K. W., Janson, L., & Murphy, S. A. (2022). Statistical inference after adaptive sampling in non-markovian environments. *arXiv preprint arXiv:2202.07098*.

[175] Zhao, P., Zhang, L., Jiang, Y., & Zhou, Z.-H. (2020). A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics* (pp. 746–755).: PMLR.

[176] Zhou, M., Mintz, Y., Fukuoka, Y., Goldberg, K., Flowers, E., Kaminsky, P., Castillejo, A., & Aswani, A. (2018). Personalizing mobile fitness apps using reinforcement learning. In *CEUR workshop proceedings*, volume 2068: NIH Public Access.

[177] Zhou, Y., Matsouaka, R. A., & Thomas, L. (2020). Propensity score weighting under limited overlap and model misspecification. *Statistical methods in medical research*, 29(12), 3721–3756.