Context in Public Health for Underserved Communities: A Bayesian Approach to Online Restless Bandits

Biyonka Liang[1], **Lily Xu**[1*], **Aparna Taneja**[2], **Milind Tambe**[1,2], **Lucas Janson**[1]
[1]Harvard University
[2]Google Research
biyonka@g.harvard.edu, lily.x@columbia.edu, aparnataneja@google.com, tambe@seas.harvard.edu, ljanson@fas.harvard.edu

Abstract

Public health programs often provide interventions to encourage program adherence, and effectively allocating interventions is vital for producing the greatest overall health outcomes, especially in underserved communities where resources are limited. Such resource allocation problems are often modeled as restless multi-armed bandits (RMABs) with unknown underlying transition dynamics, hence requiring online reinforcement learning (RL). We present Bayesian Learning for Contextual RMABs (BCoR), an online RL approach for RMABs that novelly combines techniques in Bayesian modeling with Thompson sampling to flexibly model the complex RMAB settings present in public health program adherence problems, namely context and non-stationarity. BCoR's key strength is the ability to leverage shared information within and between arms to learn the unknown RMAB transition dynamics quickly in intervention-scarce settings with relatively short time horizons, which is common in public health applications. Empirically, BCoR achieves substantially higher finite-sample performance over a range of experimental settings, including a setting using real-world adherence data that was developed in collaboration with ARMMAN, an NGO in India which runs a large-scale maternal mHealth program, showcasing BCoR practical utility and potential for real-world deployment.

1 Introduction

Public health programs in a wide array of areas such as communicable disease (Killian et al. 2019), prenatal care (Hegde and Doshi 2016; Ope 2020), and cancer prevention (Wells et al. 2011; Lee, Lavieri, and Volk 2019) often have many beneficiaries who are not adhering to their treatment. As adherence can be vital for ensuring positive health impacts, especially in the underserved communities these programs are designed to aid, programs frequently must decide how to allocate a scarce set of resources or *interventions* to beneficiaries at risk of drop-out of the program due to continued non-adherence. Such a constrained resource allocation problem is often modeled as a Restless Multi-Armed Bandit (RMAB) (Whittle 1980), which is an extension of stochastic multi-armed bandits where each arm represents a Markov Decision Process (MDP). Specifically, a beneficiary is represented as an arm, their adherence status as the state of the corresponding MDP, and the allocation of an intervention as the action. The reward is a function of the adherrence status of the beneficiary (i.e., the state of the MDP) and a budget-constrained subset of arms are given intervention (i.e., pulled) at each timestep. Most relevantly for our work, RMABs have been used to model this resource allocation problem by our collaborators at ARMMAN (ARMMAN 2022), an NGO in India which runs a large-scale maternal mHealth (mobile health) program that disseminates vital health information to pregnant beneficiaries via automated voice calls each week, with the goal of improving maternal and infant health outcomes (Mate et al. 2022; Verma et al. 2023; Wang et al. 2023). To encourage listenership (i.e., adherence), ARMMAN's healthcare workers can give live service calls (the intervention) to a subset of beneficiaries each week to troubleshoot potential barriers to adherence, thus improving overall listenership, which has been shown to result in to positive impact on the behavioural outcomes of the mothers (Dasgupta et al. 2024).

Developing an algorithm for resource allocation for mHealth programs like ARMMAN faces a few key challenges: (1) the transition dynamics of the underlying MDPs (e.g., corresponding to the beneficiaries' adherence) are unknown a priori; (2) these settings are often resource scarce, so the (intervention) budget B is typically much smaller than the total number of arms N; and (3) the time horizon T is naturally limited to the treatment period, which is often small relative to N. For example, ARMMAN can only provide interventions to $\sim 2\%$ of their beneficiaries each week (Hegde and Doshi 2016), and the time horizon is naturally limited by the duration of a pregnancy. Due to the scarce intervention budget and relatively short time horizon, at any given time point, many (or most) of the beneficiaries in the program have never been intervened on. Thus, the algorithm must make a resource allocation decision despite not observing the underlying outcome distributions for a potentially vast number of arms. Additionally, previous analyses on public health programs indicate that beneficiary adherence varies with contextual factors such as income and education, and that adherence rates can vary over time, suggesting non-stationarity in transition dynamics (Mate et al. 2022; Verma et al. 2023). However, these known properties of public health settings (e.g., short time horizon, contextual information, and non-stationarity) are largely unaddressed by existing online RL

*Now at Columbia University

methods for RMABs. Neglecting these valuable properties which could improve an algorithm's ability to quickly learn an effective intervention allocation.

1.1 Main Contributions

Driven by the needs of public health programs such as ARMMAN's maternal health program, we present **B**ayesian Learning for **Co**ntextual **R**MABs (BCoR), an online RL approach for RMABs which novelly combines techniques in hierarchical Bayesian modeling with Thompson sampling to account for the complexities of this application. In contrast with existing work on online RL for RMABs, which has largely focused on establishing asymptotic regret guarantees in T, we develop a learning approach that directly addresses an important real-world application, enabling known characteristics of the public health domain such as contextual information and non-stationarity to be incorporated for the first time. As such, a key component of our research approach is our direct collaboration with our stakeholders, ARMMAN, to understand the specific challenges around beneficiary adherence for public health program adherence. As adherence is crucial for many public health programs, which often share similar challenges as ARMMAN, our insights from ARMMAN enable us to precisely design and contextualize BCoR with respect to public health applications.

Our research approach is aligned with other works which design bandit algorithms for real-world applications (Mary, Gaudel, and Preux 2015; Shen et al. 2015; Ding, Li, and Liu 2019; Bouneffouf and Rish 2019) rather than providing improved theoretical results which, as we show in Section 5.2, do not always correspond to improved practical performance. In fact, part of our contribution is pushing beyond the simplifying assumptions previous works have made to facilitate their theoretical analysis, enabling us to incorporate the complex structure present in applied examples; by necessity, this complexity makes theory more challenging and we consider it beyond the scope of this work. In particular, BCoR's primary contribution is its strong empirical performance in experimental settings reflective of real-world settings: BCoR achieves substantially higher reward than existing approaches across a wide array of realistic experimental settings, including a setting based on real data from ARMMAN's maternal health program and various settings where our model is misspecified. Such results exhibit BCoR's robustness and ability to leverage key characteristics of its application area, making it more suitable than existing approaches for real-world application.

2 Related Work

Online RL for RMABs When the RMAB transition dynamics are *known*, the Whittle index policy (Whittle 1980), which pulls the top B arms with the highest estimated future value if pulled (called the *Whittle index*), asymptotically achieves the optimal time-averaged reward under certain conditions (Weber and Weiss 1990; Wang et al. 2019). Since RMAB dynamics are *unknown* in many applied settings, online RL approaches for RMABs generally use different learning techniques to quantify uncertainty on the transition probabilities, then apply a Whittle index policy, for instance, by computing the Upper Confidence Bound (UCB) for each arm's state-action transitions (Wang et al. 2023), utilizing Thompson sampling (Jung and Tewari 2019; Jung, Abeille, and Tewari 2019; Akbarzadeh and Mahajan 2023) or Q-learning (Biswas et al. 2021; Xiong and Li 2023). However, to our knowledge, all existing approaches for online RL in RMABs learn each arm's state-action transitions *individually*, without sharing information within or between arms (e.g., via contextual information). Furthermore, not only are the asymptotic regret guarantees of these methods limited to simplified conditions as discussed above, but in addition, they usually only show empirical performance in relatively simple RMAB settings, such as when the number of arms N is small (usually < 100) and the budget is high (usually $> 30\%$ of N). ARMMAN and other real-world public health programs operate on a much larger scale, and we show in Section 5.2 that empirical performance in the aforementioned simple settings often does not translate to such realistic settings.

Incorporating Contextual Information Context is often present in bandit settings and can be highly informative (Hofmann, Whiteson, and de Rijke 2011; Bouneffouf, Rish, and Aggarwal 2020). While contextual information has been heavily explored in standard multi-armed bandits, e.g., (Auer 2002; Langford and Zhang 2007; Chu et al. 2011; Li, Wu, and Wang 2021), there are no works, to our knowledge, which consider context for online learning in RMABs.

Allowing for Non-Stationarity in RMABs While existing online RL methods for RMABs assume stationary transition dynamics (Biswas et al. 2021; Gafni, Yemini, and Cohen 2022; Wang et al. 2023), this assumption may not well approximate real-world settings (Mate et al. 2022; Verma et al. 2023) and there is limited evidence to suggest that existing approaches are robust to non-stationarity (Biswas et al. 2021). Though non-stationarity in RMABs has been explored for RMABs with known transition dynamics (Zayas-Caban, Jasin, and Wang 2019; Ghosh et al. 2023; Zhang and Frazier 2022), such solutions generally rely on solving a linear program directly using the true transition dynamics. It is unclear how such results could be extended to online RL settings where the algorithm must learn the transition dynamics and determine a good policy simultaneously.

Other Related Learning Approaches Learning RMAB transition dynamics can be considered a specific case of multi-task reinforcement learning (MTRL), which aims to learn the transition dynamics of a set of MDPs, often by modeling the MDPs as having some shared structure between them by, for instance, clustering MDPs with similar transition probabilities (Wilson et al. 2007; Lazaric and Ghavamzadeh 2010; Yu et al. 2021). However, these MTRL approaches are largely designed for offline learning settings and, generally, do not consider contextual information and do not provide policy recommendations for regret minimization. For instance, past deployments of ARMMAN have largely focused on an offline learning approach precisely due to challenges this paper is set to address (Mate et al. 2020, 2022). Some online RL approaches focus on learning a single, sometimes partially observed, MDP However, these approaches are designed to

learn a single MDP cannot be used to determine a set of actions to apply to a collection of MDPs, as in the RMAB setting.

3 Problem Setting

Consider an RMAB instance with N arms. The learning algorithm interacts with the RMAB over T timesteps with an (intervention) budget of $B \ll N$ pulls at each timestep, where T is fixed and known in advance. Each arm is an MDP defined by the tuple $\left(\mathcal{S}, \mathcal{A}, R, \left\{P_i^{(t)}(s' \mid s, a)\right\}_{s',a,s,t}\right)$, where $\mathcal{S}$, $\mathcal{A}$, and $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are the shared state space, action space, and reward function, respectively, across all arms and timesteps. The standard formulation for RMABs sets $\mathcal{A} = \{0, 1\}$ where $a = 1$ represents a budget-constrained pull. The set of transition probabilities for arm i is $P_i \coloneqq \left\{P_i^{(t)}(s' \mid s, a)\right\}_{s',a,s,t}$, that is, for arm i in state $s \in \mathcal{S}$ that receives action $a \in \{0, 1\}$ at time $t \in [T]$, the transition probability to state $s' \in \mathcal{S}$ is $P_i^{(t)}(s' \mid s, a)$. Note that the transitions are indexed by time t, allowing for non-stationarity. We assume the reward function is known and the state of all arms is observable, even if they were not pulled. Importantly, we assume that all P_i are *unknown in advance* by the learning algorithm. Let $\boldsymbol{s}_t = (s_{1,t}, \ldots, s_{N,t})$ and $\boldsymbol{a}_t = (a_{1,t}, \ldots, a_{N,t})$ represent the tuple of states and actions across all arms at time t, respectively, where we are constrained by $\sum_{i=1}^{N} a_{i,t} \leq B$ for all $t \in [T]$. As in previous public health applications (Ong'ang'o et al. 2014; Newman et al. 2018; Ayer et al. 2019; Lee, Lavieri, and Volk 2019; Mate et al. 2022; Verma et al. 2023; Wang et al. 2023), we also model program adherence with $\mathcal{S} = \{0, 1\}$, where state $s_{i,t} = 0$ represents beneficiary i being in a non-adhering state, and $s_{i,t} = 1$ represents beneficiary i being in an adhering state. An action $a_{i,t} = 0$ represents no intervention on beneficiary i, and $a_{i,t} = 1$ represents an intervention. A reward of 1 is accrued when the beneficiary is in an adhering state and 0 otherwise, i.e., $R(s, a) = s$. Thus, the total reward at timestep t is a count of the number of beneficiaries who are in an adhering state, $\sum_{i=1}^{N} s_{i,t}$, and the time-averaged reward at timestep t (which we aim to maximize in this paper) is:

$$R^{(t)} = \frac{1}{t} \sum_{j=1}^{t} \sum_{i=1}^{N} s_{i,j}. \tag{1}$$

Note that reward is calculated across all arms, as an arm may generate reward even when not pulled, i.e., a beneficiary may be in an adhering state even when no intervention is applied. We assume the above state space, action space, and reward function in this paper, noting that the binary state and action spaces are largely for presentation purposes as it simplifies our notation and is standard in public health applications. BCoR's Bayesian modeling framework naturally extends to non-binary state and action spaces and general reward functions; see Appendix B.

4 The BCoR Algorithm

We introduce BCoR, which integrates a Bayesian model into Thompson sampling for the online learning of RMABs with complex structure. We use *hierarchical Bayesian modeling*, a Bayesian modeling approach where the prior distribution of some model parameters depends on other parameters, which are also assigned a priori. Hierarchical Bayesian models are flexible tools for modeling complex phenomena across broad application areas (Curry et al. 2013; Lawson 2018; Britten et al. 2021), as the hierarchical structure on the model parameters can be used to represent complex relationships and interactions between variables of the model.

4.1 Learning the Transition Dynamics

To apply Thompson sampling, we must specify a Bayesian model of the RMAB's reward distribution. Since our rewards equal our states, we focus on modeling the state transition distribution, i.e., the $P_i^{(t)}(1 \mid s, a)$'s, for all $s \in \{0, 1\}, a \in \{0, 1\}$, $t \in [T]$, and $i \in [N]$.[1] To specify this model, we will first consider the simple non-contextual RMAB with stationary transitions and incrementally add complexity, separately explaining each addition until our full model is presented. The flexibility of hierarchical Bayesian modeling often comes with a corresponding computational cost, and hence the components we discuss in this section are carefully chosen not just to incorporate properties of our applied setting, but also to maintain computational tractability; see Appendix B.1 for further details.

Sharing Information Within an Arm A simple and natural choice of Bayesian model for this simple RMAB is to treat $P_i(1 \mid s, a)$ as drawn independently from some distribution (e.g., Unif$[0, 1]$), where we remove the superscript for time (for now). Hence, this model aims to learn each arm's state-action transitions *individually* — requiring the model to learn $4N$ different transition probabilities (i.e., each of the four state-action pairs for each arm). This learning approach may not be very effective because, as discussed in Section 1, the budget and time horizon may be small relative to N, and hence, there may be many arms for which the algorithm never observes behavior under $a = 1$. However, since the vast majority of arms will receive $a = 0$ at each timestep, we expect to observe a relatively large set of outcomes for each arm when $a = 0$ over time. Through discussions with ARMMAN, we also expect that, for a given arm j, its transition dynamics when $a = 0$ have some relationship to its transition dynamics when $a = 1$. Thus, it can be useful to *share information within an arm* to better estimate an arm's active ($a = 1$) transition probabilities, for which there is very little data, based on its passive ($a = 0$) transition data, for which there is much more data. Hence, we propose to model this relationship as:

$$\begin{aligned} P_i(1 \mid s, 0) &= \Phi\left(\alpha_i^{(s,0)}\right) \\ P_i(1 \mid s, 1) &= \Phi\left(\alpha_i^{(s,1)} + b_0 \alpha_i^{(0,0)} + b_1 \alpha_i^{(1,0)}\right) \end{aligned} \tag{2}$$

for all $s \in \mathcal{S}$, where Φ is the standard normal cumulative distribution function. Here, b_0, b_1, and each of the $\alpha_i^{(s,a)}$'s

[1] Note, $P_i^{(t)}(0 \mid s, a) = 1 - P_i^{(t)}(1 \mid s, a)$ deterministically, so we only need to model the $P_i^{(t)}(1 \mid s, a)$ transitions.

are parameters of this Bayesian model, and, as is standard in Bayesian models of this form (Gelman et al. 2013), we will set their priors as zero-centered Normal distributions. Hence, we can interpret the $\alpha_i^{(s,0)}$'s as representing each arm's individual passive transitions ($a = 0$), the $\alpha_i^{(s,1)}$'s as representing the active transitions ($a = 1$), and b_0 and b_1 as representing the informativeness of an arm's transition dynamics under passive actions for inferring its dynamics under active action. Hence, the parameters b_0 and b_1 enable us to use information about passive actions, of which we observe many, to inform our inference on active actions, of which we observe very few. As we set the prior on b_0 and b_1 to be zero-centered, which corresponds to no information sharing, our model will only learn to share information if the data suggests that such a relationship exists.

Sharing Information Across Arms The ideas presented above deal with sharing information *within* a given arm. As RMAB problems often come with contextual information for each arm, such as age, education, and other demographic factors, it is desirable to use this information to share information *across* arms. For instance, we may reasonably expect that arms with similar covariates will have similar behavior (e.g., low-income beneficiaries tend to have lower adherence (Mohan et al. 2021)). Given a covariate matrix $\boldsymbol{X} \in \mathbb{R}^{N \times k}$ where the row vectors X_i represent feature vectors for each of the arms, we incorporate $\boldsymbol{X}$ into Model (2) by adding a parameter $\boldsymbol{\beta}^{(s,a)} \in \mathbb{R}^k$ for each state-action pair and modeling the transitions as:

$$\begin{aligned} P_i(1 \mid s, 0) &= \Phi\left(X_i \boldsymbol{\beta}^{(s,0)} + \alpha_i^{(s,0)}\right) \\ P_i(1 \mid s, 1) &= \Phi\Big(X_i \boldsymbol{\beta}^{(s,1)} + \alpha_i^{(s,1)} \\ &\qquad + b_0 \alpha_i^{(0,0)} + b_1 \alpha_i^{(1,0)}\Big), \end{aligned} \tag{3}$$

where, similar to b_0 and b_1, we set zero-centered Normal priors on the $\boldsymbol{\beta}^{(s,a)}$'s. Note, the four $\boldsymbol{\beta}^{(s,a)}$ vectors are shared *across* all arms for *each* state-action pair. Since we may not observe many transitions when $a = 1$ due to budget constraints, it will be harder to learn the $\boldsymbol{\beta}^{(s,a=1)}$'s. To facilitate learning the $\boldsymbol{\beta}^{(s,a=1)}$'s, we can *add a level of hierarchy* to the $\boldsymbol{\beta}^{(s,a)}$'s by modeling all four of the $\boldsymbol{\beta}^{(s,a)}$ vectors as having the same mean vector $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, which is a new parameter in our model, on which we place a (normally distributed) prior; see the third and sixth lines of Model (4.1) for the explicit forms of $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and the $\boldsymbol{\beta}^{(s,a)}$'s. Intuitively, our posterior updates of $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ would use data across all arms' state-action transitions to learn where the covariate effects $\boldsymbol{\beta}^{(s,a)}$'s are "centered" and our posterior updates of each $\boldsymbol{\beta}^{(s,a)}$ would use data across arms specifically in state s that receive action a to learn how far that particular (s, a) pair deviates from the center. Hence, *this hierarchy facilitates greater information sharing*.

Addressing Non-Stationarity As is standard in the RMAB literature, we have so far treated the transition dynamics as stationary or fixed over time. However, in real-world scenarios, the arms often exhibit non-stationary transition dynamics (Mate et al. 2022; Verma et al. 2023). To model these time effects, we use spline regression, a common approach for flexibly modeling non-linear effects (Hastie, Tibshirani, and Friedman 2009). Given a spline basis matrix $\boldsymbol{M} \in \mathbb{R}^{T \times d}$ with rows M_t, where T is the time horizon and d is the dimension of the spline basis, we can incorporate non-stationarity into our model as:

$$\begin{aligned} P_i^{(t)}(1 \mid s, 0) &= \Phi\left(X_i \boldsymbol{\beta}^{(s,0)} + M_t \boldsymbol{\eta}^{(s,0)} + \alpha_i^{(s,0)}\right) \\ P_i^{(t)}(1 \mid s, 1) &= \Phi\Big(X_i \boldsymbol{\beta}^{(s,1)} + M_t \boldsymbol{\eta}^{(s,1)} + \alpha_i^{(s,1)} \\ &\qquad + b_0 \alpha_i^{(0,0)} + b_1 \alpha_i^{(1,0)}\Big), \end{aligned} \tag{4}$$

where we set zero-centered Normal priors on the $\boldsymbol{\eta}^{(s,a)}$'s and we now have superscripts on the $P_i^{(t)}(s' \mid s, a)$ to denote time-varying transition dynamics. Hence, $\boldsymbol{\eta}^{(s,a)}$ represents the magnitude of the time effects on the transition dynamics.

Lastly, we place a prior on the variance of the $\alpha_i^{(s,a)}$'s, which we denote $\tau^2_{\alpha^{(s,a)}}$; see the fourth and fifth lines of Model (4.1) for the explicit forms of the $\tau^2_{\alpha^{(s,a)}}$'s and $\alpha_i^{(s,a)}$'s. We do so because, without this prior, the $\alpha_i^{(s,a)}$'s would be modeled *per arm*, while all other parameters are shared across arms. Hence, we cannot directly use the posteriors of the $\alpha_i^{(s,a)}$'s to infer anything about new arms since they only represent information about a single arm's state-action pair. Adding a prior on the variance enables us to share information across arms for all parameters. See Definition (4.1) for the full statement of the BCoR learning model. We let $\mathbf{0}_k \in \mathbb{R}^k$ be the k-dimensional vector with all 0 entries and $I_{k \times k}$ be the $k \times k$ identity matrix.

Hence, the user-specified values which are required as inputs to our model are: $\tau_0, \sigma_0, \tau^2_{\boldsymbol{\mu}}, \tau^2_{b_0}, \tau^2_{b_1}, \tau^2_{\boldsymbol{\beta}^{(s,a)}}$ and $\tau^2_{\boldsymbol{\eta}^{(s,a)}}$, for all $s \in \{0, 1\}, a \in \{0, 1\}$. Such user inputs are common in Bayesian modeling, and, as more data is observed, the posterior distributions of the parameters will be most strongly influenced by the actual data rather than these specific inputs (van der Vaart 2000; Gelman et al. 2013). The input values used in all experimental results in this paper were chosen to ensure they were reasonably default values given the problem setting; see Appendix B.1 for the exact specification and further discussion. Importantly, we used the *same* input values for *all* experimental results presented in this paper, including our example constructed from real ARMMAN data and many misspecified simulation settings where these input values do not correctly reflect the RMAB's true structure, across various configurations of N, T and B. Our experimental results across these various settings, shown in Figures 1–2 and Figures 4–15, suggest that BCoR is primarily learning from the data and hence, the specific input values had little impact on the performance of the algorithm.

Hence, using insights from our collaboration with domain experts, we carefully incorporate characteristics of our application area into the structure of BCoR using hierarchical Bayesian modeling, which has not previously been used for online learning in the RMAB literature. Though it may seem limiting to assume a linear model, Bayesian linear models have an extensive history of being sufficiently expressive and empirically effective across a wide range of complex

settings that almost surely do not satisfy linearity (Gelman, Fagan, and Kiss 2007; Hilbe 2009; Curry et al. 2013; Gelman et al. 2013; Lawson 2018; Britten et al. 2021). From a bias-variance tradeoff perspective, a linear model is more suitable for the small sample size settings we consider compared to to a more complex model, which would have higher variance. Notably, the real data-based example of Section 5.3 is highly misspecified and strongly violates the linearity assumptions, and BCoR is still achieves higher reward than existing approaches.

4.2 Online Arm Selection

To apply Thompson sampling, we can update the posterior distribution of our model parameters at each timestep, and take a draw from the posterior. As we observe more data over time, we expect the posterior distributions of our model parameters to concentrate around values that best fit the data, and hence, so will our estimates of the transition dynamics.

Definition 4.1 (The BCoR Learning Model)**.**

$$
\begin{aligned}
b_0 &\sim \mathcal{N}\left(0, \tau_{b_0}^2\right) \\
b_1 &\sim \mathcal{N}\left(0, \tau_{b_1}^2\right) \\
\boldsymbol{\mu}_{\boldsymbol{\beta}} &\sim \mathcal{N}\left(\mathbf{0}_k, \tau_{\boldsymbol{\mu}}^2 I_{k\times k}\right) \\
\tau_{\alpha^{(s,a)}}^2 &\sim \text{Inv-Gamma}(\tau_0, \sigma_0) \qquad \forall s, a \\
\alpha_i^{(s,a)} &\sim \mathcal{N}\left(0, \tau_{\alpha^{(s,a)}}^2\right) \qquad \forall s, a \\
\boldsymbol{\beta}^{(s,a)} &\sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \tau_{\boldsymbol{\beta}^{(s,a)}}^2 I_{k\times k}\right) \qquad \forall s, a \\
\boldsymbol{\eta}^{(s,a)} &\sim \mathcal{N}\left(\mathbf{0}_d, \tau_{\boldsymbol{\eta}^{(s,a)}}^2 I_{d\times d}\right) \qquad \forall s, a
\end{aligned} \tag{5}
$$

$$
P_i^{(t)}(1 \mid s, 0) = \Phi\left(X_i \boldsymbol{\beta}^{(s,0)} + M_t \boldsymbol{\eta}^{(s,0)} + \alpha_i^{(s,0)}\right)
$$

$$
\begin{aligned}
P_i^{(t)}(1 \mid s, 1) = \Phi\Big(&X_i \boldsymbol{\beta}^{(s,1)} + M_t \boldsymbol{\eta}^{(s,1)} + \alpha_i^{(s,1)} \\
&+ b_0 \alpha_i^{(0,0)} + b_1 \alpha_i^{(1,0)}\Big),
\end{aligned} \tag{6}
$$

Concretely, for all $s \in \{0,1\}$, $a \in \{0,1\}$, $i \in [N]$, let $\tilde{b}_0^{(t)}, \tilde{b}_1^{(t)}, \tilde{\alpha}_i^{(s,a)(t)}, \tilde{\boldsymbol{\eta}}^{(s,a)(t)}, \tilde{\boldsymbol{\beta}}^{(s,a)(t)}$ represent a draw from the posterior distributions of $b_0, b_1, \alpha_i^{(s,a)}, \boldsymbol{\eta}^{(s,a)}, \boldsymbol{\beta}^{(s,a)}$ at time t, respectively. We can generate estimates of the transition probabilities $\tilde{P}_i^{(t)}(1 \mid s, a)$ by plugging these posterior draws into the last two lines of Model (4.1). Specifically, for all $s \in \{0,1\}$, $a \in \{0,1\}$, $i \in [N]$,

$$
\tilde{P}_i^{(t)}(1|s,0) \coloneqq \Phi\Big(X_i \tilde{\boldsymbol{\beta}}^{(s,0)(t)} + M_t \tilde{\boldsymbol{\eta}}^{(s,0)(t)} + \tilde{\alpha}_i^{(s,0)(t)}\Big) \tag{7}
$$

$$
\begin{aligned}
\tilde{P}_i^{(t)}(1 \mid s,1) \coloneqq \Phi\Big(&X_i \tilde{\boldsymbol{\beta}}^{(s,1)(t)} + M_t \tilde{\boldsymbol{\eta}}^{(s,1)(t)} + \tilde{\alpha}_i^{(s,1)(t)} \\
&+ \tilde{b}_0^{(t)} \tilde{\alpha}_i^{(0,0)(t)} + \tilde{b}_1^{(t)} \tilde{\alpha}_i^{(1,0)(t)}\Big).
\end{aligned} \tag{8}
$$

Using the $\tilde{P}_i^{(t)}(1 \mid s, a)$'s, we implement a Whittle index policy (Whittle 1980), which computes the Whittle index using the set of all $\tilde{P}_i^{(t)}(1 \mid s, a)$'s and pulls the B arms with the highest Whittle indices. See Definition B.1 in Appendix B.1 for a formal definition of the Whittle index (Whittle 1980). The Whittle index is computable via an efficient binary search approach presented in (Qian et al. 2016). Further details on the implementation and computational efficiency of BCoR are in Appendix B.1.

Algorithm 1: BCoR

1: **Input:** N arms, budget B, time horizon T, covariate matrix $\boldsymbol{X} \in \mathbb{R}^{N\times k}$, spline basis matrix $\boldsymbol{M} \in \mathbb{R}^{T\times d}$, model inputs $\{\tau_0, \sigma_0, \tau_{\boldsymbol{\mu}}^2, \tau_{b_0}^2, \tau_{b_1}^2, \tau_{\boldsymbol{\beta}^{(s,a)}}^2, \tau_{\boldsymbol{\eta}^{(s,a)}}^2\}$ for all $s \in \{0,1\}, a \in \{0,1\}$.
2: **for** timestep $t \in \{1, \ldots, T\}$ **do**
3: Observe $\boldsymbol{s}_t$ and use all historical data to compute the posterior distribution of Model (4.1)'s parameters.[2]
4: From the posterior distribution computed in the previous step, generate $\tilde{P}_i^{(t)}(1 \mid s, a)$ as in Equation (7) for all $s \in \{0,1\}, a \in \{0,1\}, i \in [N]$.
5: Using the $\tilde{P}_i^{(t)}(1|s,a)$'s generated in the previous step, compute the Whittle index for all $i \in [N]$ and pull the B arms with the highest Whittle indices.

5 Experiments

We show that BCoR consistently achieves high reward across various experimental settings, even in challenging settings where the data generating model is misspecified. We also evaluate performance in a setting constructed from a real-world public health campaign, namely ARMMAN's maternal healthcare program. General implementation details for all experiments in this paper are in Appendix B.1, with details specific to Sections 5.2 and 5.3 in Appendices B.2 and B.3, respectively. The code, data, and instructions needed to reproduce the results of Section 5.2, as well as all additional simulations in Appendix B.2, are available via our Github: `https://github.com/biyonka/BCoR`.

5.1 Methods Under Comparison

We evaluate the BCoR algorithm as described in Algorithm 1. For comparison, we consider the *UCWhittle* approach of Wang et al. (2023) (denoted *UCW-Value* in their paper). This approach, which computes a UCB for each arm's state-action transitions and selects an "optimistic" value within the confidence bound to plug into the Whittle index policy, exhibits superior empirical performance over other existing approaches such as Biswas et al. (2021) and Wang et al. (2019). We also consider a Thompson sampling–based approach based on Akbarzadeh and Mahajan (2023), denoted *TS*, which performs Thompson sampling on the *individual* arm's state-action pairs (i.e., it models each arm's state-action transitions individually with no information sharing), then plugs the estimated transitions into the Whittle index policy. For baselines, the *Random* algorithm assigns $a = 1$ to B arms uniformly at random at each timestep, providing a lower baseline for the reward without the use of any learning approach. We also

[2] At time $t = 1$, before having observed transitions, the posterior remains the prior.

implement a Whittle index *oracle* approach, which executes the Whittle index policy using the true transition dynamics.

5.2 Simulation Experiments

Given a fixed number of arms N, time horizon T, and budget B, we use Model (4.1) to generate simulated RMAB instances over $1{,}000$ random seeds. For each instance, we run all algorithms and calculate the time-averaged reward (Equation 1) at each timestep $t \in [T]$. The initial state provided for each algorithm is randomized across the seeds. We plot the average performance across the $1{,}000$ seeds, as shown in Figure 1.

We explore various parameterizations of Model (4.1) to generate a well-specified setting and various misspecified settings for the BCoR learning model. For the well-specified setting of Figure 1(a), the parameterization of Model (4.1) used to generate the RMAB instances is the same as the prior used for BCoR. The misspecified settings shown in Figures 1(b–d) each represent zeroing out just one component of Model (4.1) to generate the RMAB instances (and in each setting; all components not explicitly zero'ed out are left as in Model (4.1)). For instance, Figure 1(b) and (c) represent a setting where the transitions are truly stationary (i.e., we set $\boldsymbol{\eta}^{(s,a)} = 0, \forall s, a$ when generating the RMAB instances across the random seeds), but the prior for BCoR never changes from the one used in the well-specified setting. Hence, the prior allows for properties like non-stationarity and informative contextual information, and BCoR must learn from the data that some of these properties are not present. In particular, Figure 1(e) represents a highly misspecified setting where the transition probabilities are generated by zeroing out all components of Model (4.1) and just leaving the random effects $\alpha_i^{(s,a)} \sim \mathcal{N}(0, \sigma^2)$ (note we also remove the prior on the variance of the $\alpha_i^{(s,a)}$'s), so that the transition dynamics are just generated as $P_i(1 \mid s, a) = \Phi\left(\alpha_i^{(s,a)}\right)$ for all s, a. In such a setting, the transitions are *stationary* and there is *no information sharing* within an arm or across arms. While existing approaches such as UCWhittle and TS are implicitly designed for this setting (since they learn each arm's state-action transitions individually), BCoR must learn that the RMAB instances have no information sharing and are stationarity. Hence, this setting is particularly challenging for BCoR. See Appendix B.1 and B.2 for further details about the simulation environment.

In the well-specified setting of Figure 1(a) and the partially misspecified setting of Figures 1(b–c), BCoR achieves significantly higher reward than other approaches. For instance, in Figure 1(c), BCoR's final time-averaged reward is *more than double* that of the next-best solution, TS, with even larger engagement wins in the settings of Figures 1(a)–(b), potentially correspond to significant positive impacts in overall health outcomes in real-world deployment. In particular, Figures 1(a–c) have covariate structure, showing that even when the RMAB is stationary, which is the setting TS and UCWhittle are designed for, ignoring informative covariate information when present can significantly decrease performance. In Figure 1(d), BCoR achieves slightly higher reward than the other approaches by accounting for the non-stationarity, even though it has the additional challenge of learning that the covariates are completely uninformative. In Figure 1(e), it is only possible to learn each arm's state-action transitions individually. In this setting, none of the methods perform significantly better than random over the entire time horizon. These results show that if no structure is present, and the time horizon T and budget B are small relative to N, the learning problem is too challenging for essentially any approach to significantly outperform random. Intuitively, this is because the only way to learn about a particular state-action transition is to directly observe it, but having small T and B relative to N means that any algorithm will only observe a small portion of the RMAB's dynamics. Hence, in applied settings such as ARMMAN's maternal health program, where we expect similar configurations of T, B, and N, it is essential to use a learning algorithm that can leverage properties present in the RMAB instance. *We repeated this experiment for different RMAB configurations, varying the number of arms N, the time horizon T, the budget B, and the number of covariates k, which are in Appendix B.2. Those results show similar trends as in Figure 1, exhibiting BCoR's robustness to these different experimental settings.* Additionally, recall from Section 4.1 that we used the *same* prior for *all* experimental results in this paper, where each plot represents an average over $1{,}000$ different RMAB instances. BCoR's performance in misspecified settings across these many RMAB instances suggests that it is not very sensitive to the specific prior used and is effectively learning from the data.[3] In summary, these experiments exhibit how existing approaches, which have ostensibly strong theoretical guarantees, can perform close to, and sometimes no better than, a random selection algorithm in moderate sample regimes, even in simplified settings amenable to their theoretical guarantees such as when stationarity is present. These empirical results exhibit that the types of existing theoretical guarantees found in restless bandit papers are insufficient to ensure high performance in public health applications. While all methods will improve as they observe more samples (over T), only BCoR can use information over the N arms, thus allowing it to learn more quickly and efficiently in challenging settings where T and B are limited.

5.3 Experiment Using Real Data From ARMMAN

It is essential to evaluate the performance of BCoR experimentally on a data-driven simulator before running it on actual ARMMAN beneficiaries in order to confirm its expected performance before actual deployment in real-world settings. Hence, we construct a data-driven simulator that approximates the true dynamics based on real historical ARMMAN covariate data, leveraging our extensive collaborations with ARMMAN to inform the design of the simulator. ARMMAN provided anonymized covariate information from $24{,}011$ beneficiaries enrolled in their maternal health program, collected

[3] For instance, if BCoR was highly sensitive to its prior, it would not be able to effectively learn from the data when the RMAB is, e.g., stationary. However, BCoR achieves high reward in this setting (see Figure 1(b)) suggesting BCoR is not highly sensitive.

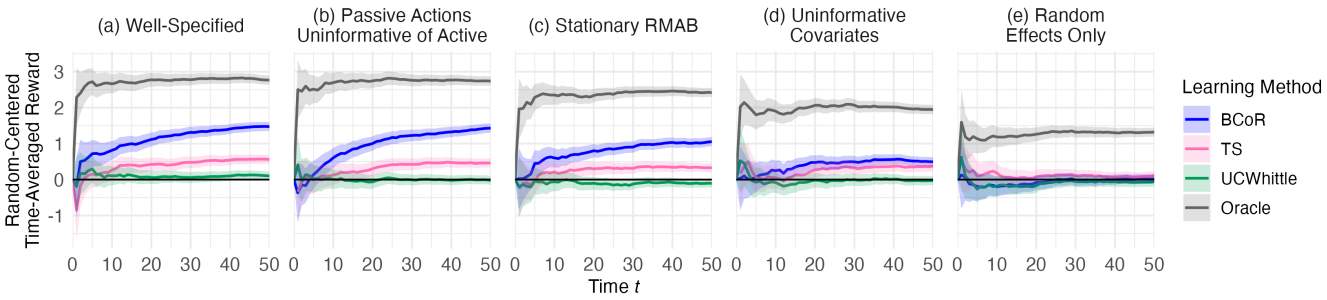

Figure 1: We generate all RMAB instances using $N = 400, T = 50$, and $B = 10$, i.e., B is 2.5% of N, across 1,000 random seeds. The covariate matrix $\boldsymbol{X}$ is randomly generated with $k = 4$ (two continuous covariates and two categorical) across the random seeds. The various RMAB simulation settings are detailed in Section 5.2 and can be summarized as (a) a well-specified setting (no components of Model (4.1) are zero'ed out), (b) a setting where passive actions are uninformative of active actions ($b_0 = b_1 = 0$), (c) a stationary setting ($\boldsymbol{\eta}^{(s,a)} = \mathbf{0}, \forall s, a$), (d) a setting with uninformative covariate information ($\boldsymbol{\mu_\beta} = 0$, $\boldsymbol{\beta}^{(s,a)} = 0, \forall s, a$), and (e) a highly misspecified setting, i.e., one where the RMAB instances are stationary with no information sharing between or within the arms. Lines represent the time-averaged reward of each method averaged over the 1,000 random seeds with the Random baseline subtracted out. Error bars depict ± 2 SEs.

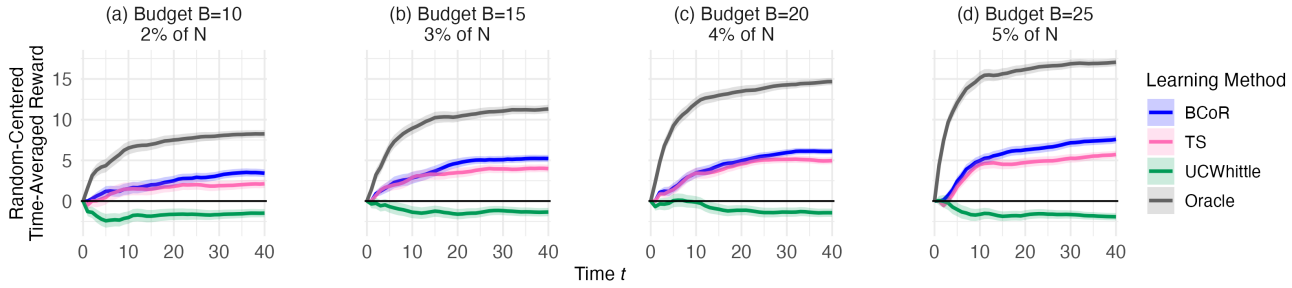

Figure 2: Performance of various methods on the ARMMAN data-driven example described in Section 5.3 with $N = 500, T = 40$, and varying budget B, where all $B \leq 5\%$ of N to reflect the magnitude of real-world budget constraints. Lines represent the time-averaged reward of each method averaged over 100 random seeds with the Random baseline subtracted out. Note, the grey line is an oracle approach with access to the true transitions. Error bars depict ± 2 SEs. UCWhittle performs worse than random across all settings, which can sometimes occurs when the budget is relatively small and the time horizon is short, though it recovers over a longer time horizon; see Figure 17.

in 2022. We generate the data-driven simulator by using ARMMAN's internal estimates of the true transition probabilities given a beneficiary's covariate information. We choose N, T, and B to reflect the learning challenges present in the ARMMAN setting, e.g., $T = 40$ because that is the approximate length of a pregnancy in weeks, and the varying budget values are reflective of ARMMAN's true budget constraints.

As shown in Figure 2, BCoR achieves the highest reward across all budget constraints, translating to significant increases in overall engagement compared to the next best performing method; for instance, a 61% *increase in engagement* in the tightest, and hence most realistic, budget setting ($B = 10$). Hence, BCoR enables RMABs to be applied to realistic public health settings with significantly better performance at unprecedented scale in N and shorter horizon T, potentially leading to life-saving health outcomes in real-world mHealth settings. See Appendix B.3 for further details on implementation and data privacy protocols.

6 Conclusion and Ethical Considerations

We present BCoR, the first online RL approach for contextual and non-stationary RMABs designed for mHealth in close collaboration with domain experts at ARMMAN. Using a novel combination of techniques in Bayesian hierarchical modeling combined with Thompson sampling, BCoR outperforms existing approaches across a wide range of challenging empirical settings reflective of real-world applications, significantly increasing the potential social impact of RMABs for real-world resource allocation. Importantly, BCoR is designed to improve listenership in mHealth settings and hence, *would not* withhold any health information from beneficiaries. Our data-driven simulator is considered secondary analysis, was approved by ARMMAN's ethics board, and was performed with fully anonymized data collected with consent prior to data collection. See Appendix A and B.3 for detailed discussions on ethical considerations and data privacy.

Acknowledgements

L.X. was supported by a Google PhD fellowship and B.L. was partially supported by the NSF Graduate Research Fellowship Program. L.J. was partially supported by the NSF grant CBET-2112085. The authors would like to thank Jun Liu for helpful discussion on our Bayesian model and Nathan Cheng and anonymous reviewers for their thoughtful comments.

References

Akbarzadeh, N.; and Mahajan, A. 2023. On Learning Whittle Index Policy for Restless Bandits with Scalable Regret. arXiv:2202.03463.

ARMMAN. 2022. ARMMAN Helping Mothers and Children. https://armman.org/. Accessed: 2022-05-19.

Auer, P. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov): 397–422.

Ayer, T.; Zhang, C.; Bonifonte, A.; Spaulding, A. C.; and Chhatwal, J. 2019. Prioritizing hepatitis C treatment in US prisons. *Operations Research*, 67(3): 853–873.

Bashingwa, J. J. H.; Mohan, D.; Chamberlain, S.; Arora, S.; Mendiratta, J.; Rahul, S.; Chauhan, V.; Scott, K.; Shah, N.; Ummer, O.; Ved, R.; Mulder, N.; and LeFevre, A. E. 2021. Assessing exposure to Kilkari: a big data analysis of a large maternal mobile messaging service across 13 states in India. *BMJ Global Health*, 6(Suppl 5).

Betancourt, M. 2017. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Biswas, A.; Aggarwal, G.; Varakantham, P.; and Tambe, M. 2021. Learn to Intervene: An Adaptive Learning Policy for Restless Bandits in Application to Preventive Healthcare. In *IJCAI*, 4039–4046.

Bouneffouf, D.; and Rish, I. 2019. A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*.

Bouneffouf, D.; Rish, I.; and Aggarwal, C. 2020. Survey on Applications of Multi-Armed and Contextual Bandits. In *IEEE Congress on Evolutionary Computation*, 1–8.

Britten, G. L.; Mohajerani, Y.; Primeau, L.; Aydin, M.; Garcia, C.; Wang, W.-L.; Pasquier, B.; Cael, B.; and Primeau, F. W. 2021. Evaluating the benefits of Bayesian hierarchical methods for analyzing heterogeneous environmental datasets: A case study of marine organic carbon fluxes. *Frontiers in Environmental Science*, 9: 491636.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 208–214. JMLR Workshop and Conference Proceedings.

Corotto, P. S.; McCarey, M. M.; Adams, S.; Khazanie, P.; and Whellan, D. J. 2013. Heart failure patient adherence: epidemiology, cause, and treatment. *Heart Failure Clinics*, 9(1): 49–58.

Curry, D. J.; Cochran, J. J.; Radhakrishnan, R.; and Pinnell, J. 2013. Hierarchical Bayesian prediction methods in election politics: introduction and major test. *Journal of Political Marketing*, 12(4): 275–305.

Dasgupta, A.; Boehmer, N.; Madhiwalla, N.; Hedge, A.; Wilder, B.; Tambe, M.; and Taneja, A. 2024. Preliminary Study of the Impact of AI-Based Interventions on Health and Behavioral Outcomes in Maternal Health Programs. arXiv:2407.11973.

Ding, K.; Li, J.; and Liu, H. 2019. Interactive anomaly detection on attributed networks. In *ACM International Conference on Web Search and Data Mining*, 357–365.

Elazan, S. J.; Higgins-Steele, A. E.; Fotso, J. C.; Rosenthal, M. H.; and 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 Journal of Community Medicine*, 41(1): 34.

Gafni, T.; Yemini, M.; and Cohen, K. 2022. Restless Multi-Armed Bandits under Exogenous Global Markov Process. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Gasparrini, A. 2011. Distributed lag linear and non-linear models in R: the package dlnm. *Journal of Statistical Software*, 43(8): 1–20.

Gelman, A.; Carlin, J.; Stern, H.; Dunson, D.; Vehtari, A.; and Rubin, D. 2013. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Gelman, A.; Fagan, J.; and Kiss, A. 2007. An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102(479): 813–823.

Ghosh, A.; Nagaraj, D.; Jain, M.; and Tambe, M. 2023. Indexability is Not Enough for Whittle: Improved, Near-Optimal Algorithms for Restless Bandits. In *AAMAS*.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. ISBN 9780387848846.

Hegde, A.; and Doshi, R. P. 2016. Assessing the Impact of Mobile-based Intervention on Health Literacy among Pregnant Women in Urban India. In *American Medical Informatics Association Annual Symposium*.

Hilbe, J. 2009. Data analysis using regression and multilevel/hierarchical models. *Journal of Statistical Software*, 30: 1–5.

Hofmann, K.; Whiteson, S.; and de Rijke, M. 2011. Contextual bandits for information retrieval. In *NIPS Workshop on Bayesian Optimization, Experimental Design, and Bandits*.

Jung, Y.-H.; Abeille, M.; and Tewari, A. 2019. Thompson Sampling in Non-Episodic Restless Bandits. In *arXiv*.

Jung, Y.-H.; and Tewari, A. 2019. Regret Bounds for Thompson Sampling in Episodic Restless Bandit Problems. In *NeurIPS*.

Killian, J. A.; Wilder, B.; Sharma, A.; Choudhary, V.; Dilkina, B.; and Tambe, M. 2019. Learning to Prescribe Interventions for Tuberculosis Patients Using Digital Adherence Data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Langford, J.; and Zhang, T. 2007. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In *NeurIPS*, volume 20.

Lawson, A. B. 2018. *Bayesian Disease Mapping:Hierarchical Modeling in Spatial Epidemiology*. CRC Press.

Lazaric, A.; and Ghavamzadeh, M. 2010. Bayesian multi-task reinforcement learning. In *International Conference on Machine Learning*, 599–606. Omnipress.

Lee, E.; Lavieri, M. S.; and Volk, M. 2019. Optimal screening for hepatocellular carcinoma: A restless bandit model. *Manufacturing and Service Operations Management*, 21(1): 198–212.

Li, C.; Wu, Q.; and Wang, H. 2021. Unifying Clustered and Non-stationary Bandits. In *International Conference on Artificial Intelligence and Statistics*.

Mary, J.; Gaudel, R.; and Preux, P. 2015. Bandits and recommender systems. In *Machine Learning, Optimization, and Big Data Workshop*, 325–336. Springer.

Matamoros, I. A. A. 2020. An introduction to computational complexity in Markov Chain Monte Carlo methods. *arXiv preprint arXiv:2004.07083*.

Mate, A.; Killian, J.; Xu, H.; Perrault, A.; and Tambe, M. 2020. Collapsing bandits and their application to public health intervention. *NeurIPS*, 33: 15639–15650.

Mate, A.; Madaan, L.; Taneja, A.; Madhiwalla, N.; Verma, S.; Singh, G.; Hegde, A.; Varakantham, P.; and Tambe, M. 2022. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *AAAI Conference on Artificial Intelligence*, volume 36.

Mohan, D.; Scott, K.; Shah, N.; Bashingwa, J. J. H.; Chakraborty, A.; Ummer, O.; Godfrey, A.; Dutt, P.; Chamberlain, S.; and LeFevre, A. E. 2021. Can health information through mobile phones close the divide in health behaviours among the marginalised? An equity analysis of Kilkari in Madhya Pradesh, India. *BMJ Global Health*, 6.

Neal, R. M.; et al. 2011. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11): 2.

Newman, P. M.; Franke, M. F.; Arrieta, J.; Carrasco, H.; Elliott, P.; Flores, H.; Friedman, A.; Graham, S.; Martinez, L.; Palazuelos, L.; et al. 2018. Community health workers improve disease control and medication adherence among patients with diabetes and/or hypertension in Chiapas, Mexico: an observational stepped-wedge study. *BMJ Global Health*, 3(1): e000566.

Nishtala, S.; Kamarthi, H.; Thakkar, D.; Narayanan, D.; Grama, A.; Hegde, A.; Padmanabhan, R.; Madhiwalla, N.; Chaudhary, S.; Ravindran, B.; and Tambe, M. 2020. Missed calls, Automated Calls and Health Support: Using AI to improve maternal health outcomes by increasing program engagement. In *AI for Social Good Workshop*.

Ong'ang'o, J. R.; Mwachari, C.; Kipruto, H.; and Karanja, S. 2014. The effects on tuberculosis treatment adherence from utilising community health workers: a comparison of selected rural and urban settings in Kenya. *PLoS One*, 9(2): e88937.

Ope, B. W. 2020. Reducing maternal mortality in Nigeria: addressing maternal health services' perception and experience. *Journal of Global Health Reports*, 4: e2020028.

Qian, Y.; Zhang, C.; Krishnamachari, B.; and Tambe, M. 2016. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *AAMAS*, 123–131.

Shen, W.; Wang, J.; Jiang, Y.-G.; and Zha, H. 2015. Portfolio choices with orthogonal bandit learning. In *IJCAI*.

Stan Development Team. 2024. RStan: the R interface to Stan. R package version 2.32.5.

van der Vaart, A. 2000. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press. ISBN 9780521784504.

Verma, S.; Mate, A.; Wang, K.; Madhiwalla, N.; Hegde, A.; Taneja, A.; and Tambe, M. 2023. Restless Multi-Armed Bandits for Maternal and Child Health: Results from Decision-Focused Learning. In *AAMAS*, 1312–1320.

Wang, K.; Xu, L.; Taneja, A.; and Tambe, M. 2023. Optimistic Whittle Index Policy: Online Learning for Restless Bandits. In *AAAI*.

Wang, K.; Yu, J.; Chen, L.; Zhou, P.; Ge, X.; and Win, M. Z. 2019. Opportunistic scheduling revisited using restless bandits: Indexability and index policy. *IEEE Transactions on Wireless Communications*, 18(10): 4997–5010.

Weber, R. R.; and Weiss, G. 1990. On an Index Policy for Restless Bandits. *Journal of Applied Probability*, 27(3): 637–648.

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

Whittle, P. 1980. Multi-Armed Bandits and the Gittins Index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2): 143–149.

Wilson, A.; Fern, A.; Ray, S.; and Tadepalli, P. 2007. Multi-task reinforcement learning: a hierarchical Bayesian approach. In *International Conference on Machine learning*, 1015–1022.

Xiong, G.; and Li, J. 2023. Finite-Time Analysis of Whittle Index based Q-Learning for Restless Multi-Armed Bandits with Neural Network Function Approximation. arXiv:2310.02147.

Yu, T.; Kumar, A.; Chebotar, Y.; Hausman, K.; Levine, S.; and Finn, C. 2021. Conservative Data Sharing for Multi-Task Offline Reinforcement Learning. In *NeurIPS*.

Zayas-Caban, G.; Jasin, S.; and 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.

Zhang, X.; and Frazier, P. I. 2022. Near-optimality for infinite-horizon restless bandits with many arms. *arXiv preprint arXiv:2203.15853*.