A Bayesian Approach to Online Learning for Contextual Restless Bandits with Applications to Public Health

Biyonka Liang[1] **Lily Xu**[2] **Aparna Taneja**[3] **Milind Tambe**[2] **Lucas Janson**[1]

Abstract

Restless multi-armed bandits (RMABs) are used to model sequential resource allocation in public health intervention programs. In these settings, the underlying transition dynamics are often unknown a priori, requiring online reinforcement learning (RL). However, existing methods in online RL for RMABs cannot incorporate properties often present in real-world public health applications, such as contextual information and non-stationarity. 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 a wide range of complex RMAB settings, such as contextual and non-stationary RMABs. A key contribution of our approach is its ability to leverage shared information within and between arms to learn unknown RMAB transition dynamics quickly in budget-constrained settings with relatively short time horizons. Empirically, we show that BCoR achieves substantially higher finite-sample performance than existing approaches over a range of experimental settings, including one constructed from a real-world public health campaign in India.

1. Introduction

Restless multi-armed bandits (RMABs) (Whittle, 1980) are extensions of stochastic multi-armed bandits where each arm represents a Markov Decision Process (MDP). The reward distribution depends on the state of the MDP and a budget-constrained subset of arms are pulled at each timestep (Villar et al., 2015; Borkar et al., 2017; Sun et al., 2022). RMABs are often used to model problems where the individual arms may change state regardless of whether they are pulled, and have been explored in areas such as autonomous driving (Li et al., 2021b), machine management (Iannello et al., 2012; Abbou & Makis, 2019), and especially healthcare. For public health programs in areas such as communicable disease management (Tuldrà et al., 1999; Killian et al., 2019), prenatal and infant care (Hegde & Doshi, 2016; Ope, 2020; Bashingwa et al., 2021), and cancer prevention (Wells et al., 2011; Lee et al., 2019), beneficiaries may at any time enter an adhering (e.g., following their treatment regimen) or non-adhering (e.g., missing a treatment) state. As adherence is often vital for ensuring certain health outcomes, programs may allocate resources or interventions to patients at risk of drop-out from the program due to continued non-adherence. We can model this problem as an RMAB by representing each beneficiary as an arm, their adherence status as the state of the corresponding MDP, and the allocation of an intervention as the action.

In these settings, the transition dynamics of the underlying MDPs (e.g., corresponding to the beneficiaries' adherence) are unknown a priori, and the (intervention) budget B is typically much smaller than the total number of arms N. For example, ARMMAN, a non-profit based in India with a maternal and infant healthcare program, can only provide interventions to $\sim 2\%$ of their beneficiaries each week (Hegde & Doshi, 2016). Additionally, the time horizon is naturally limited to the treatment period (e.g., the duration of a pregnancy), which is often small relative to the number of arms. Due to the scarce intervention budget and relatively short time horizon, at any given time point, many (or most) of the arms a learning algorithm has to choose from have never been pulled before. Thus, the algorithm must make a decision even though it has not observed the underlying outcome distributions for a potentially vast number of arms.

Previous works on public health programs also suggest 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 (Hegde & Doshi, 2016; Nishtala et al., 2020; Mohan et al., 2021; Mate et al., 2022; Verma et al., 2023). For instance, low-income beneficiaries are at higher risk of non-adherence, and thus, can benefit more from additional interventions (Mohan et al., 2021; Bashingwa et al., 2023).

[1] Department of Statistics, Harvard University, Cambridge, MA, USA [2] Department of Computer Science, Harvard University, Cambridge, MA, USA [3] Google Research India. Correspondence to: Biyonka Liang <biyonka@g.harvard.edu>.

However, both contextual information and non-stationarity are largely unaddressed by existing online RL methods for RMABs. Yet, accounting for these properties could improve an algorithm's ability to quickly learn an effective intervention allocation. Learning quickly is especially important in public health settings, where early intervention has been shown to improve long-term program adherence and overall health outcomes (Amagai et al., 2022; Mohan et al., 2022).

1.1. Main Contributions

We present **B**ayesian Learning for **Co**ntextual **R**MABs (BCoR), an online RL approach for RMABs which novelly combines techniques in Bayesian modeling with Thompson sampling to model complex RMAB settings, such as contextual and non-stationary RMABs. BCoR is, to our knowledge, the first approach that utilizes the flexibility of Bayesian modeling for online RL for RMABs. As the structure of Bayesian models can be tailored to known characteristics of its application areas (e.g., via the incorporation of contextual information), these models enable us to address a key problem in areas like public health: While domain expertise suggests the presence of certain RMAB properties (such as informative contextual information and non-stationarity), existing methods are unable to fully incorporate this knowledge. Using hierarchical Bayesian modeling, BCoR incorporates these properties by *sharing information within and across arms*, thus enabling BCoR to quickly learn effective resource allocations in the budget- and time-constrained settings often present in public health applications. This information sharing, which allows us to account for properties that are largely unaddressed by existing work in online RL for RMABs like contextual information and non-stationarity, is a novel contribution of our approach.

Empirically, BCoR achieves considerably higher reward than existing approaches across a wide array of experimental settings, even some misspecified settings where information sharing or non-stationarity is not completely present. BCoR also pays essentially no cost in terms of reward when these properties are completely absent, such as in stationary RMABs with no information sharing. We also exhibit the performance of BCoR on a setting constructed using real data from the ARMMAN maternal health program.

A key component of our research approach is direct collaboration with stakeholders. Authors of this paper have collaborated extensively with ARMMAN for many years to understand the specific challenges around beneficiary adherence for public health programs. As adherence is crucial for many public health programs, which often share similar challenges such as having scarce intervention budgets relative to their total beneficiaries, our insights from ARMMAN enable us to more precisely design and contextualize our method with respect to our stated application area.

2. Related Work

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 & 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, Wang et al. (2023) compute the Upper Confidence Bound (UCB) for each arm's state-action transitions, then uses the resulting bound to estimate the transition probabilities to plug into the Whittle index policy. As Thompson sampling (TS) exhibits superior empirical performance over UCB-based approaches in various applied examples with multi-armed bandits (Chapelle & Li, 2011; Dumitrascu et al., 2018; Russo et al., 2020; Neu et al., 2022), even when using approximate sampling methods for the posterior distribution (Phan et al., 2019; Huix et al., 2023), TS provides a natural alternative to UCB in the RMAB setting. Jung & Tewari (2019), Jung et al. (2019), and Akbarzadeh & Mahajan (2023) provide theoretical explorations of Bayesian regret bounds using Thompson sampling–based approaches. Other RMAB approaches utilize Q-learning combined with Whittle index policies (Biswas et al., 2021; Xiong & 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). As a result, they often only show empirical performance in relatively simple RMAB settings, such as when the number of arms is small (usually < 100) and the budget is high (usually $> 30\%$) relative to the number of arms. [1] While individually learning each arm's state-action transitions is feasible for smaller problems, most real-world public health programs operate in resource-limited communities where the program can only provide interventions to $\sim 2\%$ of the total number of beneficiaries at each timestep, and the number of beneficiaries may be in the tens of thousands (Mohan et al., 2022). Additionally, these programs often operate under naturally limited time horizons (e.g., the length of a treatment regimen) where early intervention can be vital for ensuring long-term adherence (Mohan et al., 2021; Amagai et al., 2022). Hence, it is crucial to determine an effective intervention allocation quickly. Some work in standard MABs also conforms to the structure of many users and relatively

[1] Works such as Biswas et al. (2021) which do consider larger N also simplify the RMAB instance by, e.g., assuming the arms are clustered, so there are only a few unique transition dynamics to learn.

short time horizons (Zhang et al., 2020), but online RL methods for RMABs are generally designed for long time horizons and tend to learn slowly in finite-sample settings, despite their asymptotic regret bounds (Wang et al., 2023).

Incorporating Contextual Information Context is often present in bandit settings (Hofmann et al., 2011; Jung et al., 2012; Bouneffouf et al., 2012; Intayoad et al., 2020; Bouneffouf et al., 2020) and can be highly informative (Bashingwa et al., 2021; Mohan et al., 2021; 2022). While the incorporation of contextual information has been heavily explored in standard multi-armed bandits (Auer, 2002; Langford & Zhang, 2007; Chu et al., 2011; Dudik et al., 2011; Li et al., 2021a; Kim et al., 2023), there are no works, to our knowledge, which address online RL of contextual RMABs. Mate et al. (2022) allow for some contextual dependence of the transition dynamics via a clustering approach, but only in an offline setting where the transition dynamics and cluster membership are assumed to be known a priori.

Allowing for Non-Stationarity While existing online RL methods for RMABs assume stationary transition dynamics (Biswas et al., 2021; Gafni et al., 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 et al., 2019; Ghosh et al., 2023; Zhang & 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 & 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*. Some online RL approaches focus on learning a single, sometimes partially observed, MDP rather than jointly learning a set of MDPs (Abbasi-Yadkori & Neu, 2014; Golowich et al., 2022; Jafarnia Jahromi et al., 2022; Xiong et al., 2022), and hence, are not designed to determine a combinatorial 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 $P_i := \left\{P_i^{(t)}(s' \mid s, a)\right\}_{s',a,s,t}$, 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 $\boldsymbol{s}_t = (s_{1,t}, ..., s_{N,t})$ and $\boldsymbol{a}_t = (a_{1,t}, ..., a_{N,t})$ represent the tuple of states and actions across all arms at time t, respectively, where we must have $\sum_{i=1}^{N} a_{i,t} \leq B$ 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 et al., 2019; Mate et al., 2022; Verma et al., 2023; Wang et al., 2023), we also model the public health program adherence setting with $\mathcal{S} = \{0, 1\}$, where a state of 0 represents a beneficiary being in a non-adhering state, and a state of 1 represents a beneficiary being in an adhering state. An action of 0 represents no intervention, and an action of 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 from now on. However, BCoR's Bayesian modeling framework can be extended to non-binary state and action spaces and general reward functions.

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 prior. Hierarchical Bayesian models are flexible tools for modeling complex phenomenons 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 will model 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 $i \in [N]$.[2] 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.

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 are parameters of this Bayesian model, and, as is standard in Bayesian models of this form (Chapter 3, 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:

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

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 $\boldsymbol{\mu_\beta}$, which is a new parameter in our model, on which we place a (normally distributed) prior; see the third and sixth lines of Model (3) for the explicit forms of $\boldsymbol{\mu_\beta}$ and the $\boldsymbol{\beta}^{(s,a)}$'s. Intuitively, our posterior updates of $\boldsymbol{\mu_\beta}$ 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*.

[2] Note, $P_i^{(t)}(0 \mid s, a) = 1 - P_i^{(t)}(1 \mid s, a)$ deterministically, so we only need to model the $P_i^{(t)}(1 \mid s, a)$'s.

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 et al., 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:

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

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 (3) 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. We now explicitly state the full model. 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.

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

$$
\begin{aligned}
b_0 &\sim \mathcal{N}\left(0, \tau^2_{b_0}\right) \\
b_1 &\sim \mathcal{N}\left(0, \tau^2_{b_1}\right) \\
\boldsymbol{\mu}_{\boldsymbol{\beta}} &\sim \mathcal{N}\left(\mathbf{0}_k, \tau^2_{\boldsymbol{\mu}} I_{k \times k}\right) \\
\tau^2_{\alpha^{(s,a)}} &\sim \text{Inv-Gamma}(\tau_0, \sigma_0) \qquad \forall s, a \\
\alpha_i^{(s,a)} &\sim \mathcal{N}\left(0, \tau^2_{\alpha^{(s,a)}}\right) \qquad \forall s, a \\
\boldsymbol{\beta}^{(s,a)} &\sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \tau^2_{\boldsymbol{\beta}^{(s,a)}} I_{k \times k}\right) \qquad \forall s, a \qquad (3) \\
\boldsymbol{\eta}^{(s,a)} &\sim \mathcal{N}\left(\mathbf{0}_d, \tau^2_{\boldsymbol{\eta}^{(s,a)}} I_{d \times d}\right) \qquad \forall s, a \\
P_i^{(t)}(1 \mid s, 0) &= \Phi\left(X_i \boldsymbol{\beta}^{(s,0)} + M_t \boldsymbol{\eta}^{(s,0)} + \alpha_i^{(s,0)}\right) \\
P_i^{(t)}(1 \mid s, 1) &= \Phi\Big(X_i \boldsymbol{\beta}^{(s,1)} + M_t \boldsymbol{\eta}^{(s,1)} + \alpha_i^{(s,1)} \\
&\qquad\qquad + b_0 \alpha_i^{(0,0)} + b_1 \alpha_i^{(1,0)}\Big),
\end{aligned}
$$

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 A.2 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, 4–9, and 15–16, suggest that BCoR is primarily learning from the data and hence, the specific input values had little impact on the performance of the algorithm.

We reiterate the key advantages of such a model: 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 utilized in the online learning for RMABs literature. Through our modeling approach, BCoR shares information within an arm and between arms to more efficiently learn unknown transition dynamics, thus enabling it to achieve higher reward over existing approaches in budget- and time-constrained settings; see Section 5 for empirical results.

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.

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 (3). Specifically, for all $s \in \{0, 1\}$, $a \in \{0, 1\}$, $i \in [N]$,

$$
\begin{aligned}
\tilde{P}_i^{(t)}(1|s, 0) &\coloneqq \Phi\Big(X_i \tilde{\boldsymbol{\beta}}^{(s,0)(t)} + M_t \tilde{\boldsymbol{\eta}}^{(s,0)(t)} + \tilde{\alpha}_i^{(s,0)(t)}\Big) \\
\tilde{P}_i^{(t)}(1 \mid s, 1) &\coloneqq \Phi\Big(X_i \tilde{\boldsymbol{\beta}}^{(s,1)(t)} + M_t \tilde{\boldsymbol{\eta}}^{(s,1)(t)} \qquad (4) \\
&\qquad + \tilde{\alpha}_i^{(s,1)(t)} + \tilde{b}_0^{(t)} \tilde{\alpha}_i^{(0,0)(t)} + \tilde{b}_1^{(t)} \tilde{\alpha}_i^{(1,0)(t)}\Big).
\end{aligned}
$$

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 top B arms with the highest Whittle index; see Appendix A.1 for a formal definition of the Whittle index and computational details.

4.3. The BCoR Algorithm

Algorithm 1 provides the full BCoR algorithm.

Algorithm 1 BCoR

```
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 $\{\tau_0, \sigma_0, \tau^2_{\boldsymbol{\mu}}, \tau^2_{b_0}, \tau^2_{b_1}, \tau^2_{\boldsymbol{\beta}^{(s,a)}}, \tau^2_{\boldsymbol{\eta}^{(s,a)}}\}$ for all $s \in \{0,1\}, a \in \{0,1\}$.
for timestep $t \in \{1, \ldots, T\}$ do
    Observe $\boldsymbol{s}_t$ and use all historical data to compute the posterior distribution of Model (3)'s parameters.[3]
    From the posterior distribution computed in the previous step, generate $\tilde{P}_i^{(t)}(1 \mid s,a)$ as in Equation (4) for all $s \in \{0,1\}, a \in \{0,1\}, i \in [N]$.
    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 top $B$ arms with the highest Whittle index.
```

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 A.2, with details specific to Section 5.2 and Section 5.3 in Appendix A.3 and A.4, respectively. The code, data, and instructions needed to reproduce the results of Section 5.2, as well as all additional simulations in Appendix A.3.1, are available via `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 & 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 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 (3) 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 (3) 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 (3) 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 (3) to generate the RMAB instances (and in each setting; all components not explicitly zero'ed out are left as in Model (3)). 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 (3) 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

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

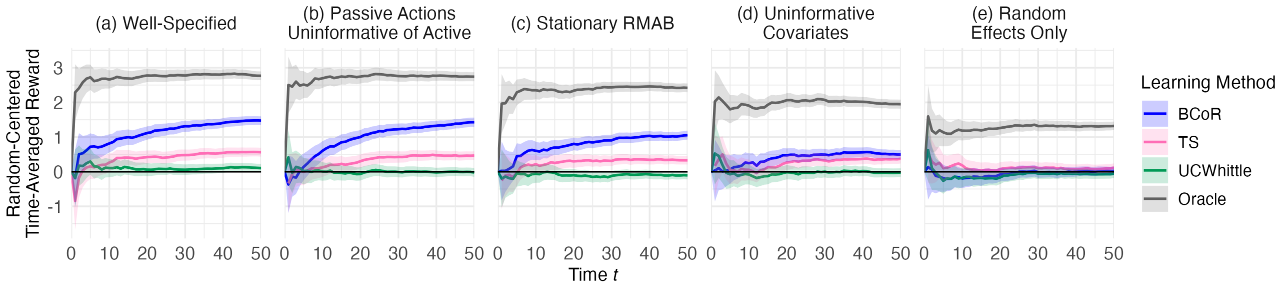


Figure 1. We generate all RMAB instances using $N = 400, T = 50$, and $B = 10$, i.e., B is 2.5% of N, across 1,000 random seeds. The covariate matrix $\boldsymbol{X}$ is randomly generated with $k = 4$ (two continuous covariates and two categorical) across the random seeds. The various RMAB simulation settings are detailed in Section 5.2 and can be summarized as (a) a well-specified setting (no components of Model (3) 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 ($\boldsymbol{\mu}_{\boldsymbol{\beta}} = 0$, $\boldsymbol{\beta}^{(s,a)} = 0, \forall s, a$), 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.

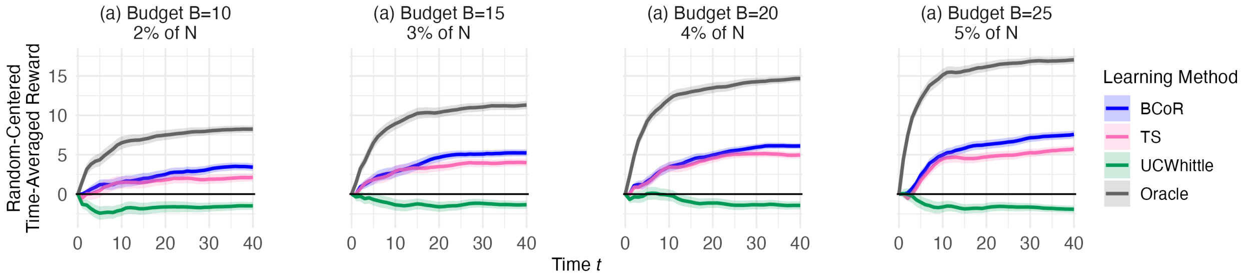


Figure 2. Performance of various methods on the ARMMAN data-driven example described in Section 5.3 with $N = 500, T = 40$, and varying budget B, where all $B \leq 5\%$ of N to reflect the magnitude of real-world budget constraints. Lines represent the time-averaged reward of each method averaged over 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. Observe that UCWhittle performs worse than random across all settings, which is not uncommon for UCWhittle when the budget is relatively small and the time horizon is short, though it recovers over a longer time horizon; see Figure 17.

RMAB instances have no information sharing and are stationarity. Hence, this setting is particularly challenging for BCoR. See Appendix A.2 and A.3 for further descriptions and visualizations of 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. 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 nonstationarity, 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 A.3.1. 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.[4]

In summary, the results of our simulation study exhibit how existing approaches can perform poorly over short time horizons when N is large, even in stationary settings. While some existing methods have asymptotic guarantees in T, they struggle with estimation when N is large because they must learn each arm's state-action transitions *individually*. In comparison, BCoR learns across N by sharing information within and between arms, and various studies exhibit evidence that there is such information to share (Hegde & Doshi, 2016; Bashingwa et al., 2021; Mohan et al., 2021; 2022). 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. This efficient learning is especially important for our applied setting to maximize health outcomes within the timespan of a program.

5.3. Experiment using Real Data from ARMMAN

As it is impossible to evaluate algorithm performance on the true transition dynamics of ARMMAN's beneficiaries without actual deployment of the algorithm, 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 in 2022, where the beneficiaries consented to have their anonymized data collected and used for research purposes; for further details on responsible data usage, see Section 7. The covariates included various demographic and program enrollment metrics such as education, income, and gestational age. ARMMAN assigns each beneficiary a risk score based on her demographic information, where higher risk scores indicate a higher risk of non-adherence with the program. Through previous analyses on prior program runs, ARMMAN has generated distributional estimates of the transition dynamics for each risk score. To create the data-driven simulator, we first mapped each beneficiary's covariates to her risk score and drew from the estimated distribution of her transition dynamics based on her risk score. This gives us a baseline stationary estimate of each beneficiary's transition dynamics. We then added non-stationarity via a spline basis model that was parameterized differently from the one provided to BCoR (i.e., has different knots and degrees of freedom), so that the model of non-stationarity is still misspecified for BCoR; see Appendix A.4.2 for further details. For our experimental results, we take a random sample of $N = 500$ beneficiaries and assess all methods under comparison over the first $T = 40$ timesteps for varying budgets all below $\leq 5\%$ of N to emulate the true applied setting. We average methods over 100 random seeds, where the initial state provided for all methods is randomized over the seeds. Note, we *did not* deploy BCoR or any other method to ARMMAN's actual beneficiaries. Our use of the anonymized beneficiary data was approved by ARMMAN's ethics review committee and we complied with all ARMMAN data privacy and ethics protocols. See Appendix A.4 for additional discussion and further implementation details.

The performance on the real-data example is shown in Figure 2. In this example, BCoR outperforms the TS and UCWhittle approaches across all budget values. For instance, at the end of the time horizon, the (random-centered) time-averaged reward of BCoR is 61% greater than that of TS in the $B = 10$ setting. BCoR achieves a similar increase in cumulative reward; see Figure 16 in Appendix A.4.2 for the corresponding cumulative reward plot. Such an increase in overall patient adherence could potentially lead to life-saving health outcomes in high-stakes public health settings. Though real-world public health programs occur on a larger scale than the examples shown in Figure 2, we believe our method's performance in these examples is representative of our applied settings. We specifically choose N, T, and B to reflect the learning challenges present in the ARMMAN setting, e.g., we choose $T = 40$ because that is the approximate length of a pregnancy in weeks, and the varying budget values are reflective of ARMMAN's true budget constraints.

6. Conclusion

We present BCoR, the first online RL approach for contextual and non-stationary RMABs using a novel combination of techniques in Bayesian hierarchical modeling combined with Thompson sampling. Through an extensive empirical evaluation, including an example constructed from a real-world public health campaign, we show that BCoR quickly outperforms existing approaches in budget-constrained settings over relatively short time horizons. Thus, BCoR provides a tailored approach for intervention allocation in public health programs, which requires fast learning in order to provide early interventions to at-risk patients and maintain

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

better overall program adherence.

7. Broader Impact

BCoR provides an online learning approach to intervention allocation in public health programs for improved beneficiary adherence. Adherence in public health programs has been shown to improve target measures of health in areas such as sexual/reproductive health (Elazan et al., 2016), cardiac disease management (Corotto et al., 2013), cancer prevention (Wells et al., 2011), and infant and maternal health in the case of ARMMAN (Hegde & Doshi, 2016; Mohan et al., 2021; Bashingwa et al., 2021; Mate et al., 2022). In particular, these programs often operate in low-income communities where healthcare workers may be scarce or lack the resources to provide individualized care to a large proportion of their beneficiaries. Approaches like BCoR can be used to improve intervention allocation for these programs to generate larger overall health improvements for their target populations.

One advantage of BCoR over black-box models such as Mate et al. (2020) is that the user can generate diagnostic measures from the posterior distribution to assess how strongly the individual covariates and potential nonstationarity are impacting the model's predictions. These diagnostic measures enable administrators to interpret the model outputs (e.g., seeing which covariates most strongly influence the final prediction), thus allowing them to screen for potentially unfair intervention allocations. Such results could be used to inform improvements, and even new iterations, of BCoR, thus enabling more fair and responsible deployment. Additionally, since BCoR shares information across all arms for all parameters, it can immediately use the posterior distribution of the parameters based on previously observed arms to infer the transition dynamics of new arms. This can aid in resource allocation in certain enrollment settings where new beneficiaries join partway through ongoing programs; see Appendix A.5 for additional discussion.

To contextualize the potential real-world impact of our approach, we provide some additional details about ARMMAN. ARMMAN is a non-profit based in India that runs mMitra, a 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. To encourage listenership, ARMMAN's community healthcare workers can give live service calls (the intervention) to a subset of beneficiaries each week. These live calls allow the community healthcare workers to troubleshoot potential barriers to information access, thus improving the beneficiaries' listenership (i.e., adherence) to the program.

In the context of deployment for ARMMAN, the BCoR algorithm would be used to enhance the allocation of live service calls to beneficiaries, which is *in addition* to the automated calls all ARMMAN beneficiaries already receive. Hence, in practice, our algorithm *would not withhold any health information* from beneficiaries. As described in Wang et al. (2023), all beneficiaries receive the same weekly health information via ARMMAN's automated calls, regardless of who receives a live service call, and no additional vital health information is provided in live service calls that is not already provided in the automated service calls. Hence, the same health information is available to all beneficiaries regardless of who receives a live service call. In addition to these scheduled live service calls, beneficiaries can request service calls themselves via a free missed call service provided by ARMMAN. If deployed, our algorithm would only help with the *scheduled* live call allocation and would not hinder access to these requested service calls.

We, the authors, complied with all ARMMAN data privacy and ethics protocols in the use of the anonymized beneficiary data; see Appendix A.4.1 for additional details. In accordance with ARMMAN's data privacy policy, the anonymized covariate information, risk scores, and all scripts used to generate the real data simulator cannot be made publicly available. However, the code, data, and instructions needed to reproduce the fully simulated results present in Section 5.2 and Appendix A.3 are available via our Github link, including all implementations of the methods under comparison which were used for our ARMMAN data-driven example.

References

Abbasi-Yadkori, Y. and Neu, G. Online learning in MDPs with side information, 2014.

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

Akbarzadeh, N. and Mahajan, A. On learning Whittle index policy for restless bandits with scalable regret, 2023.

Amagai, S., Pila, S., Kaat, A. J., Nowinski, C. J., and Gershon, R. C. Challenges in participant engagement and retention using mobile health apps: literature review. *Journal of medical Internet research*, 24(4):e35120, 2022.

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

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

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. 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), 2021.

Bashingwa, J. J. H., Mohan, D., Chamberlain, S., Scott, K., Ummer, O., Godfrey, A., Mulder, N., Moodley, D., and LeFevre, A. E. Can we design the next generation of digital health communication programs by leveraging the power of artificial intelligence to segment target audiences, bolster impact and deliver differentiated services? a machine learning analysis of survey data from rural India. *BMJ open*, 13(3):e063354, 2023.

Biswas, A., Aggarwal, G., Varakantham, P., and Tambe, M. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare, 2021.

Borkar, V. S., Kasbekar, G. S., Pattathil, S., and Shetty, P. Y. Opportunistic scheduling as restless bandits. *IEEE Transactions on Control of Network Systems*, 5(4):1952–1961, 2017.

Bouneffouf, D., Bouzeghoub, A., and Gançarski, A. L. A contextual-bandit algorithm for mobile context-aware recommender system. In Huang, T., Zeng, Z., Li, C., and Leung, C. S. (eds.), *Neural Information Processing*, pp. 324–331, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-34487-9.

Bouneffouf, D., Rish, I., and Aggarwal, C. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, 2020. doi: 10.1109/CEC48606.2020.9185782.

Britten, G. L., Mohajerani, Y., Primeau, L., Aydin, M., Garcia, C., Wang, W.-L., Pasquier, B., Cael, B., and Primeau, F. W. 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, 2021.

Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. *Advances in Neural Information Processing Systems*, 24, 2011.

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

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

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

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2011.

Dumitrascu, B., Feng, K., and Engelhardt, B. E. PG-TS: Improved Thompson Sampling for logistic contextual bandits. In *Advances in Neural Information Processing Systems*, 2018.

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

Gafni, T., Yemini, M., and Cohen, K. Restless multi-armed bandits under exogenous global Markov process. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

Gasparrini, A. Distributed lag linear and non-linear models in R: the package dlnm. *Journal of Statistical Software*, 43(8):1–20, 2011. doi: 10.18637/jss.v043.i08. URL https://doi.org/10.18637/jss.v043.i08.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955.

Ghosh, A., Nagaraj, D., Jain, M., and Tambe, M. Indexability is not enough for Whittle: Improved, near-optimal algorithms for restless bandits. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, 2023.

Golowich, N., Moitra, A., and Rohatgi, D. Learning in observable POMDPs, without computationally intractable oracles. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1458–1473. Curran Associates, Inc., 2022.

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

Hegde, A. and Doshi, R. P. Assessing the impact of mobile-based intervention on health literacy among pregnant women in urban India. In *American Medical Informatics Association Annual Symposium*, 2016.

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

Huix, T., Zhang, M., and Durmus, A. Tight regret and complexity bounds for Thompson sampling via Langevin Monte Carlo. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8749–8770. PMLR, 25–27 Apr 2023.

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

Intayoad, W., Kamyod, C., and Temdee, P. Reinforcement learning based on contextual bandits for personalized online learning recommendation systems. *Wireless Personal Communications*, 115(4):2917–2932, 2020.

Jafarnia Jahromi, M., Jain, R., and Nayyar, A. Online learning for unknown partially observable MDPs. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 1712–1732. PMLR, 28–30 Mar 2022.

Jung, T., Martin, S., Ernst, D., and Leduc, G. Contextual multi-armed bandits for web server defense. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2012.

Jung, Y.-H. and Tewari, A. Regret bounds for Thompson sampling in episodic restless bandit problems. In *Advances in Neural Information Processing Systems*, 2019.

Jung, Y.-H., Abeille, M., and Tewari, A. Thompson sampling in non-episodic restless bandits. In *arXiv*, 2019.

Killian, J. A., Wilder, B., Sharma, A., Choudhary, V., Dilkina, B., and Tambe, M. Learning to prescribe interventions for tuberculosis patients using digital adherence data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, 2019.

Kim, J., Yun, S.-Y., Jeong, M., Nam, J. H., Shin, J., and Combes, R. Contextual linear bandits under noisy features: Towards Bayesian oracles. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

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

Lazaric, A. and Ghavamzadeh, M. Bayesian multi-task reinforcement learning. In *The 27th International Conference on Machine Learning*, pp. 599–606. Omnipress, 2010.

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

Li, C., Wu, Q., and Wang, H. Unifying clustered and non-stationary bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021a.

Li, M., Gao, J., Zhao, L., and Shen, X. Adaptive computing scheduling for edge-assisted autonomous driving. *IEEE Transactions on Vehicular Technology*, 70(6):5318–5331, 2021b.

Mate, A., Killian, J., Xu, H., Perrault, A., and Tambe, M. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

Mate, A., Madaan, L., Taneja, A., Madhiwalla, N., Verma, S., Singh, G., Hegde, A., Varakantham, P., and Tambe, M. 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, 2022.

Mohan, D., Scott, K., Shah, N., Bashingwa, J. J. H., Chakraborty, A., Ummer, O., Godfrey, A., Dutt, P., Chamberlain, S., and LeFevre, A. E. 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, 2021.

Mohan, D., Bashingwa, J. J. H., Scott, K., Arora, S., Rahul, S., Mulder, N., Chamberlain, S., and LeFevre, A. E. Optimising the reach of mobile health messaging programmes: an analysis of system generated data for the Kilkari programme across 13 states in India. *BMJ Global Health*, 6 (Suppl 5):e009395, 2022.

Neu, G., Olkhovskaia, I., Papini, M., and Schwartz, L. Lifting the information ratio: An information-theoretic analysis of Thompson sampling for contextual bandits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9486–9498, 2022.

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

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

Ong'ang'o, J. R., Mwachari, C., Kipruto, H., and Karanja, S. 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, 2014.

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

Phan, M., Abbasi Yadkori, Y., and Domke, J. Thompson sampling and approximate inference. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Russo, D., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. A Tutorial on Thompson Sampling, 2020.

Stan Development Team. RStan: the R interface to Stan, 2024. URL https://mc-stan.org/. R package version 2.32.5.

Sun, J., Zhao, Y., Zhang, N., Chen, X., Hu, Q., and Song, J. A dynamic distributed energy storage control strategy for providing primary frequency regulation using multi-armed bandits method. *IET Generation, Transmission & Distribution*, 16(4):669–679, 2022.

Tuldrà, A., Ferrer, M. J., Fumaz, C. R., Bayés, R., Paredes, R., Burger, D. M., and Clotet, B. Monitoring adherence to HIV therapy. *Archives of Internal Medicine*, 159(12): 1376–1377, 1999.

van der Vaart, A. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000. ISBN 9780521784504. URL https://books.google.com/books?id=SYlmEAAAQBAJ.

Verma, S., Mate, A., Wang, K., Madhiwalla, N., Hegde, A., Taneja, A., and Tambe, M. Restless multi-armed bandits for maternal and child health: Results from decision-focused learning. In *AAMAS*, pp. 1312–1320, 2023.

Villar, S. S., Bowden, J., and Wason, J. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

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

Wang, K., Xu, L., Taneja, A., and Tambe, M. Optimistic Whittle index policy: Online learning for restless bandits. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 2023.

Weber, R. R. and Weiss, G. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990. ISSN 00219002.

Wells, K. J., Luque, J. S., Miladinovic, B., Vargas, N., Asvat, Y., Roetzheim, R. G., and Kumar, A. 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, 2011.

Whittle, P. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):143–149, 1980. ISSN 00359246.

Wilson, A., Fern, A., Ray, S., and Tadepalli, P. Multi-task reinforcement learning: a hierarchical Bayesian approach. In *Proceedings of the 24th International Conference on Machine learning*, pp. 1015–1022, 2007.

Xiong, G. and Li, J. Finite-time analysis of Whittle index based Q-learning for restless multi-armed bandits with neural network function approximation, 2023.

Xiong, Y., Chen, N., Gao, X., and Zhou, X. Sublinear regret for learning POMDPs. *Production and Operations Management*, 31(9):3491–3504, 2022.

Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Levine, S., and Finn, C. Conservative data sharing for multi-task offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

Zayas-Caban, G., Jasin, S., and Wang, G. An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits. *Advances in Applied Probability*, 51(3):745–772, 2019.

Zhang, K., Janson, L., and Murphy, S. Inference for batched bandits. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, 2020.

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

A. Appendix

A.1. The Whittle Index

We provide the formal definition of the Whittle index (Whittle, 1980), as adapted from Wang et al. (2023):

Definition A.1 (Whittle index)**.** Given transition probabilities P_i, a state s, and a discount factor $\gamma \in (0, 1)$, the *Whittle index* $W_i(P_i, s)$ of arm i in state s is defined as:

$$W_i(P_i, s) \coloneqq \inf_{m_i} \{m_i : Q^{m_i}(s, 0) = Q^{m_i}(s, 1)\}$$

where the Q-function $Q^{m_i}(s, a)$ and value function $V^{m_i}(s)$ are the solutions to the Bellman equation with penalty m_i for pulling action $a = 1$:

$$\begin{aligned} Q^{m_i}(s, a) &= -m_i a + R(s, a) + \gamma \sum_{s' \in \mathcal{S}} P_i(s' \mid s, a) V^{m_i}(s') \\ V^{m_i}(s) &= \max_{a \in \{0,1\}} Q^{m_i}(s, a) \; . \end{aligned} \tag{5}$$

The Whittle index is computable via value iteration. We utilize the implementation of the Whittle index in Wang et al. (2023) for our simulations.

A.2. General Implementation Details

The following implementation details apply to all experimental results presented in this paper, including those in Sections 5.2 and 5.3, and all related experimental results in the following Appendix sections.

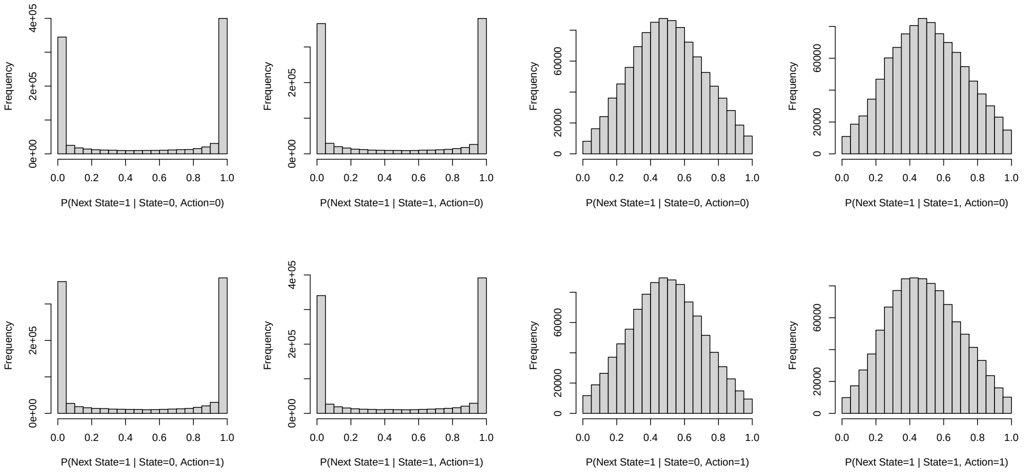
For all experimental results in this paper, we use the following prior specification for the BCoR model:

$$\begin{aligned} b_0 &\sim \mathcal{N}\left(0, 0.1^2\right) \\ b_1 &\sim \mathcal{N}\left(0, 0.1^2\right) \\ \boldsymbol{\mu_\beta} &\sim \mathcal{N}\left(\mathbf{0}_k, 0.3^2 I_{k\times k}\right) \\ \tau^2_{\alpha^{(s,a)}} &\sim \text{Inv-Gamma}(100, 1) \\ \alpha_i^{(s,a)} &\sim \mathcal{N}\left(0, \tau^2_{\alpha^{(s,a)}}\right) \\ \boldsymbol{\beta}^{(s,a)} &\sim \mathcal{N}\left(\boldsymbol{\mu_\beta}, 0.1^2 I_{k\times k}\right) \\ \boldsymbol{\eta}^{(s,a)} &\sim \mathcal{N}\left(\mathbf{0}_d, 0.3^2 I_{d\times d}\right) \end{aligned} \tag{6}$$

It may seem more natural to set wide priors on the parameter to encompass a larger possible range of RMAB instances, but setting wide priors on the parameters generate implied transition dynamics which concentrate around 0 and 1, as shown in Figure 3a.

From discussions with ARMMAN representatives, and as reflected in the transitions generated from our real data simulator in Figure 18, it is standard to assume that most beneficiaries to have some probability of switching to an engaging state that is somewhere in the middle of the $[0, 1]$ range, e.g., $[0.1, 0.9]$, with a smaller proportion of beneficiaries with transitions closer to 0 and 1. Hence, we chose the prior specification of Model (6) because it generates implied priors on the transition probabilities that reflect this default expected behavior, see Figure 3b. This prior also provides a more fair comparison with our other Bayesian approach, *TS*, which we initialize with a uniform prior on $[0, 1]$ to take advantage of Beta-Binomial conjugacy for the posterior updates. The BCoR learning model (Model 3) was fit using the `rstan` package (Stan Development Team, 2024) in the computing language `R`.

We now provide additional discussion on the Whittle index oracle used in our experimental results. Note, though the Whittle oracle approach presented in Section 5 has access to the full true RMAB transition probabilities, non-stationarity cannot be easily incorporated in the Bellman equation (Equation (5) of Definition A.1) for computing the Whittle index. Hence, to implement a Whittle-index-based oracle, we must summarize the potential non-stationarity in different ways. We implement three versions of Whittle index oracles. The *Current Time Whittle Oracle* computes the Whittle index using each arm's state-action transitions at the current time point. The *Time Average Whittle Oracle* uses the average of the transition



(a) (b)

Figure 3. Implied priors on transition probabilities using (a) a wide prior on the model parameters, $b_0 \sim \mathcal{N}\left(0, 2^2\right), b_1 \sim \mathcal{N}\left(0, 2^2\right), \boldsymbol{\mu_\beta} \sim \mathcal{N}\left(\mathbf{0}_k, 2^2 I_{k\times k}\right), \tau^2_{\alpha^{(s,a)}} \sim \text{Inv-Gamma}(100, 1), \boldsymbol{\beta}^{(s,a)} \sim \mathcal{N}\left(\boldsymbol{\mu_\beta}, 2^2 I_{k\times k}\right)$, and $\boldsymbol{\eta}^{(s,a)} \sim \mathcal{N}\left(\mathbf{0}_d, 2^2 I_{d\times d}\right)$, and (b) the prior specified in Model (6). Hence, in (b), the prior variances are set much wider than what was used for the experimental results in this paper (represented by (a)). Histograms show transitions probabilities for RMAB instances with $N = 400, T = 50, B = 10$ generated using Model (3) across 50 random seeds. The covariate matrix $\boldsymbol{X}$ and the spline matrix $\boldsymbol{B}$ are generated as described in Section A.3. Note that when using a wide prior, the transition probabilities tend to concentrate around 0 and 1, which is not representative of most realistic examples.

dynamics for a given arm's state-action pair across all time. Finally, recall that the *Cumulative Average Whittle Oracle* uses the average of the transition dynamics *up to the current time* for a given arm's state-action pair. Intuitively, all three Whittle oracles perform a version of the Whittle index policy, but handle the (possible) non-stationarity in different ways. Note, in a stationary setting, all three Whittle oracles reduce to the same method, which is the standard Whittle index policy for stationary RMABs. We found that all three oracles performed comparably in all simulation settings as well as our real data setting. Hence, for clarity of presentation, we only show one of the Whittle oracles, the *Cumulative Average Whittle Oracle*, in the results of Figures 1 and 2 of Sections 5.2 and 5.3.

For all methods under comparison, including the Whittle oracles, we compute Whittle indices using a discount of $\gamma = 0.9$, as implemented in (Wang et al., 2023). For each each random seed, we randomly initialize the starting state vector for all methods under comparison. The initial action vector for BCoR is randomly initialized, i.e., at $t = 1$, before any data has been observed, we randomly select B arms to assign $a = 1$.

Additionally, we implement versions of the TS and BCoR methods with a Greedy policy instead of a Whittle policy, where the greedy policy calculate the estimated treatment effect of action on arm i at time t as:

$$\text{TE}_i^{(t)}\left(s_{i,t}\right) \coloneqq \tilde{P}_i^{(t)}\left(1 \mid s_{i,t}, 1\right) - \tilde{P}_i^{(t)}\left(1 \mid s_{i,t}, 0\right),$$

and pull the top B arms with the highest $\text{TE}_i^{(t)}(s_{i,t})$. We also implement the analogous *Greedy Oracle*, which pulls the top B arms with the highest $\text{TE}_i^{(t)}(s_{i,t})$ using the true transition probabilities.

A.3. Additional Details on Section 5.2

In the simulations of Section 5.2, the covariate matrix $\boldsymbol{X} \in \mathbb{R}^{400\times 4}$ is generated with $k = 4$ simulated covariates, so that for each arm $i = 1, ..., 400$, we had:

$$\begin{aligned} X_{i,1} &\sim \text{round}(\mathcal{N}(22, 2^2)) \\ X_{i,2} &\sim \mathcal{N}(0, 1^2) \\ X_{i,3} &\sim \mathcal{N}(0, 1^2) \\ X_{i,4} &\sim \text{Bern}(0.5) \end{aligned}$$

The $X_{i,1}$'s represent a simulated age, the $X_{i,2}$'s and $X_{i,3}$'s represents some mean-centered and normalized continuous covariate, and the $X_{i,4}$'s represent a binary categorical covariate. The $X_{i,1}$'s were mean centered and standardized before being used for data generation. A P-spline matrix of degree three, implemented using the `ps` function of the `dlnm` package in the computing language `R` was used to generate non-stationarity (Gasparrini, 2011). Knots were automatically selected as described and recommended in the documentation of the `ps` function in `dlnm`.

See Figure 4 for a version of Figure 1 with all Oracles and Greedy policy versions of TS and BCoR present. We find that all three oracles performed comparably in all simulation settings. Additionally, the Greedy versions of BCoR and TS perform similarly to their Whittle counterparts across all experimental settings. This is sensible, since the Greedy oracle often performs comparably to the Whittle oracles as well.

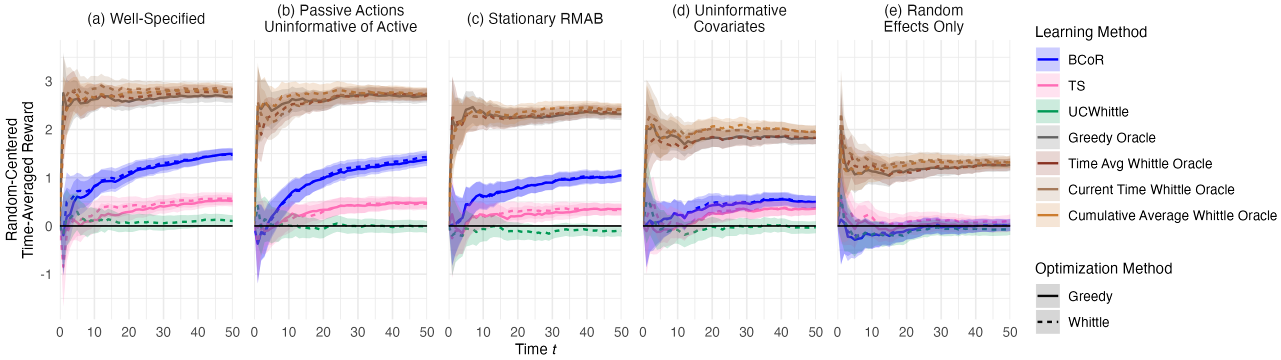


Figure 4. We generate RMAB instances using $N = 400, T = 50$, and $B = 10$, i.e., B is 2.5% of N, across 1,000 random seeds. The covariate matrix $\boldsymbol{X}$ was randomly generated with $k = 4$ (two continuous covariates and two that are categorical) across the random seeds. The various RMAB simulation settings are detailed in Section 5.2 and can be summarized as (a) a well-specified setting (no components of Model (3) 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 ($\boldsymbol{\mu}_{\boldsymbol{\beta}} = 0$, all $\boldsymbol{\beta}^{(s,a)} = 0, \forall s, a$), 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 1,000 independent instances with the Random baseline subtracted out. Error bars depict ±2 SEs.

A.3.1. Additional Simulation Results

We also varied the budget and number of covariates to assess performance. See results below.

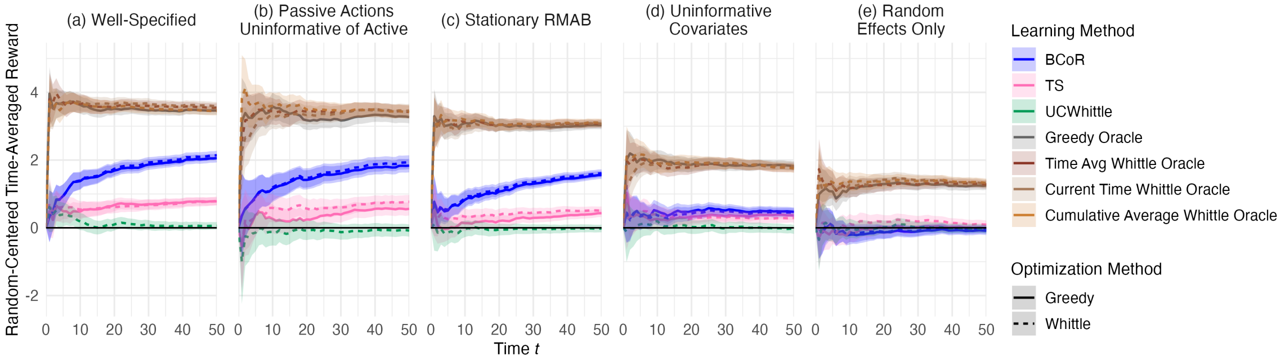


Figure 5. **Changing the number of covariates:** We generate RMAB instances using $N = 400, T = 50$, and $B = 10$, i.e., B is 2.5% of N, across 1,000 random seeds. The covariate matrix $\boldsymbol{X}$ is randomly generated with $k = 8$ (five continuous covariates and three categorical generated as described in Section A.3, adding another Bern(0.5) covariate and three additional $\mathcal{N}(0, 1)$ distributed continuous covariates) 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 (3) 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 ($\boldsymbol{\mu}_{\boldsymbol{\beta}} = 0$, all $\boldsymbol{\beta}^{(s,a)} = 0, \forall s, a$), 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.

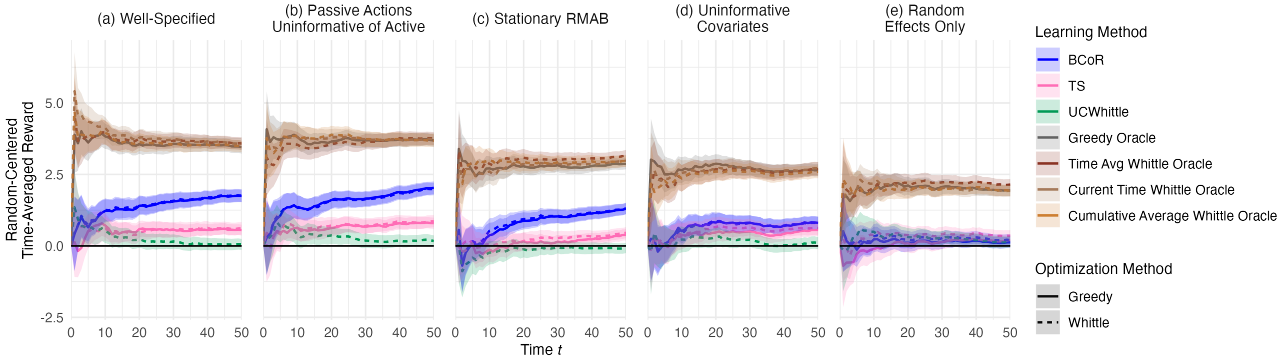


Figure 6. **Increasing the budget:** We generate RMAB instances using $N = 400, T = 50$, and $B = 15$, i.e., B is 3.75% of N, across 1,000 random seeds. The covariate matrix $\boldsymbol{X}$ was randomly generated with $k = 4$ (two continuous covariates and two that are categorical) across the random 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 (3) 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 ($\boldsymbol{\mu}_{\boldsymbol{\beta}} = 0$, all $\boldsymbol{\beta}^{(s,a)} = 0, \forall s, a$), 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.

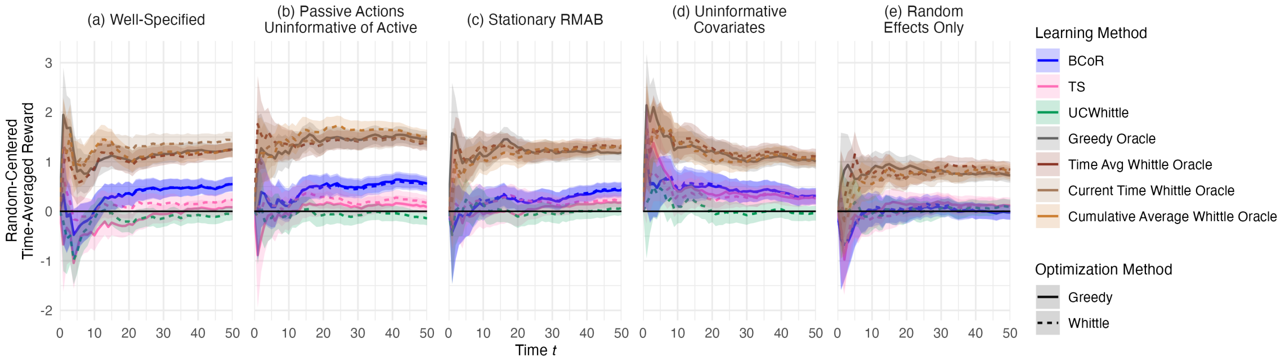


Figure 7. **Decreasing the budget:** We generate RMAB instances using $N = 400, T = 50$, and $B = 5$, i.e., B is 1.25% of N, across 1,000 random seeds. The covariate matrix $\boldsymbol{X}$ was randomly generated with $k = 4$ (two continuous covariates and two that are categorical) across the random 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 (3) 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 ($\boldsymbol{\mu}_{\boldsymbol{\beta}} = 0$, all $\boldsymbol{\beta}^{(s,a)} = 0, \forall s, a$), 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.

We also repeated the same experiment with a smaller N and T and varied the budget B. We see similar trends as in the previous simulations, exhibiting BCoR's robustness to these different experimental settings. See results below.

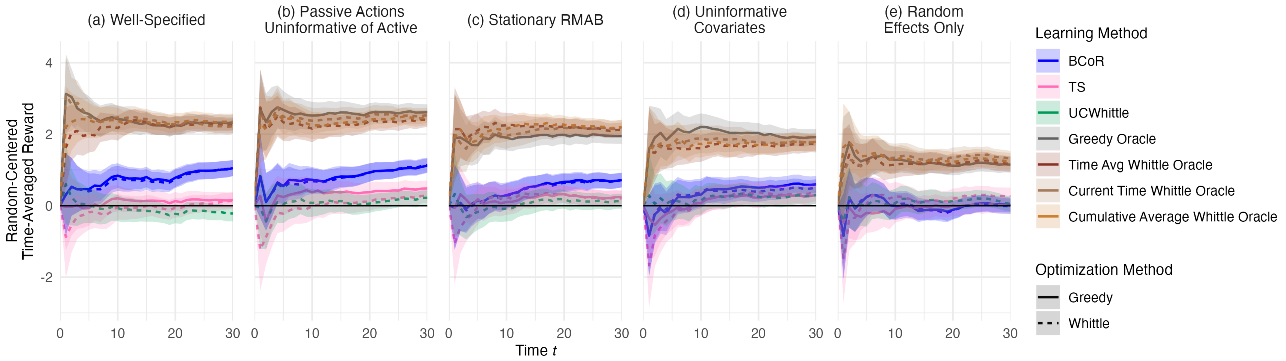


Figure 8. We all generate RMAB instances using $N = 300, T = 30$, and $B = 10$, i.e., B is 3.33% of N. The covariate matrix $\boldsymbol{X}$ was randomly generated with $k = 4$ (two continuous covariates and two that are categorical) across the random seeds. The various RMAB simulation settings are detailed in Section 5.2 and can be summarized as (a) a well-specified setting (no components of Model (3) 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 ($\boldsymbol{\mu_\beta} = 0$, all $\boldsymbol{\beta}^{(s,a)} = 0, \forall s, a$), 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 400 random seeds with the Random baseline subtracted out. Error bars depict ±2 SEs.

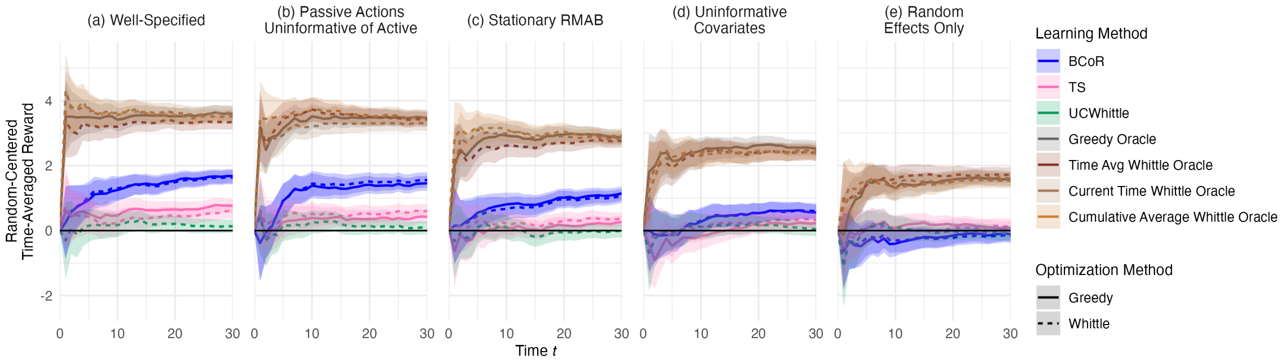


Figure 9. We all generate RMAB instances using $N = 300, T = 30$, and $B = 15$, i.e., B is 5% of N. The covariate matrix $\boldsymbol{X}$ was randomly generated with $k = 4$ (two continuous covariates and two that are categorical) across the random seeds. The various RMAB simulation settings are detailed in Section 5.2 and can be summarized as (a) a well-specified setting (no components of Model (3) 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 ($\boldsymbol{\mu_\beta} = 0$, all $\boldsymbol{\beta}^{(s,a)} = 0, \forall s, a$), 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 400 random seeds with the Random baseline subtracted out. Error bars depict ±2 SEs.

A.3.2. Transition Dynamic Visualizations for Section 5.2

Here, we provide visualizations of some of the transition dynamics used in Section 5.2. Note, for the results of Section 5.2, we draw a new RMAB instance from the data generating model for each random seed, so the provided plots are examples of a *single* draw from the data generating model; results in Section 5.2 are averaged over all 1,000 random seeds, hence representing the average reward over 1,000 unique transition dynamics from the data generating model.

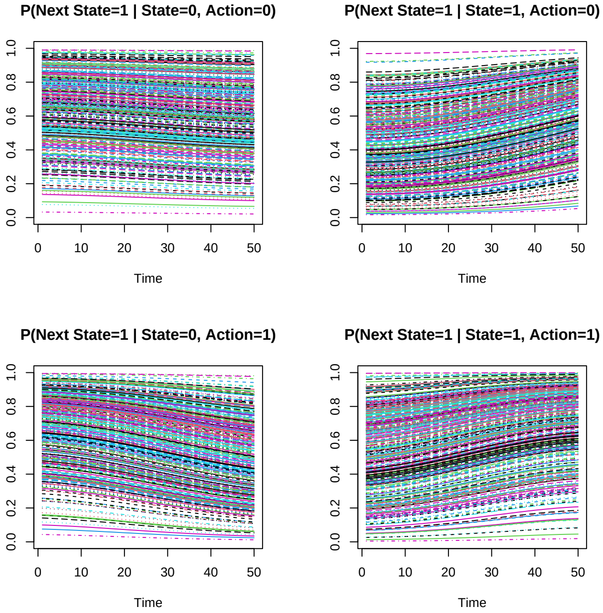


Figure 10. Example of well-specified RMAB instance from Figure 1a.

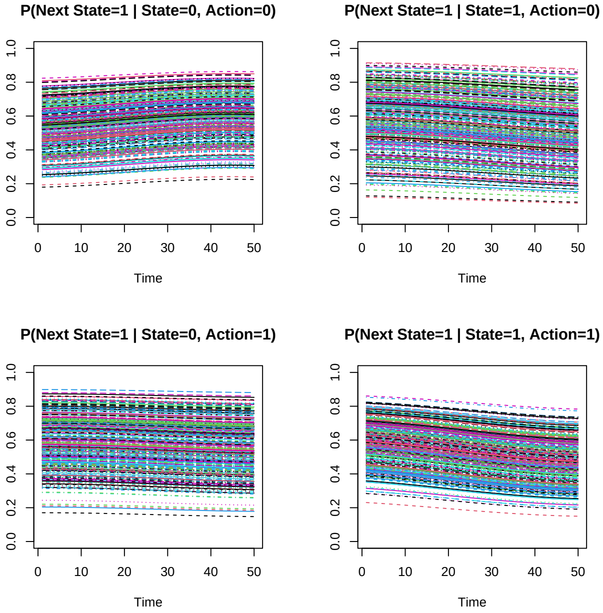


Figure 11. Example of RMAB instance with no information sharing between passive and active actions within an arm from Figure 1b.

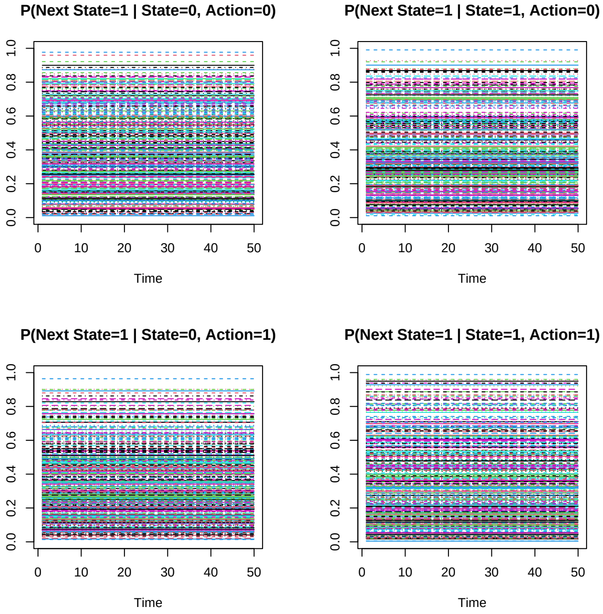


Figure 12. Example of stationary RMAB instance from Figure 1c.

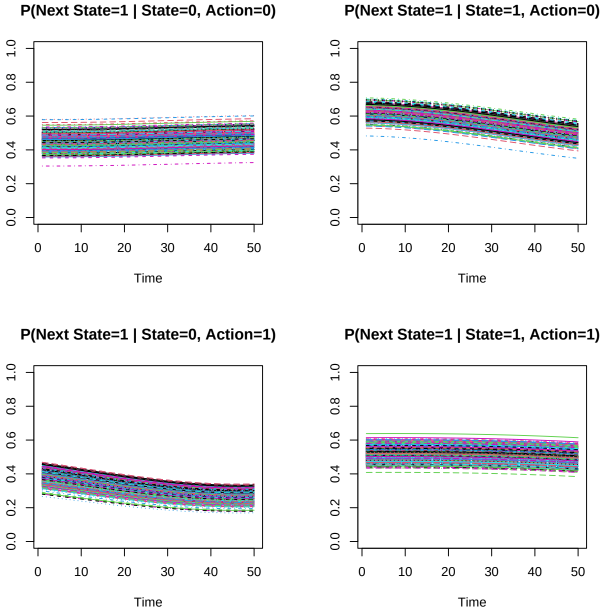


Figure 13. Example of RMAB instance with uninformative covariates from Figure 1d.

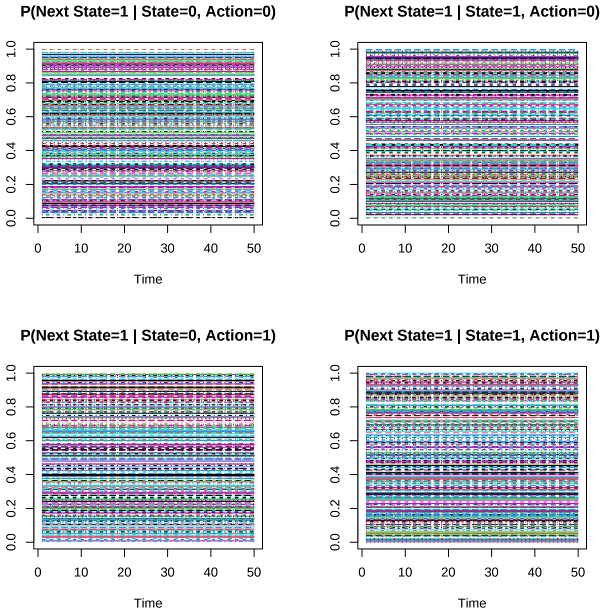


Figure 14. Example of highly misspecified RMAB instance from Figure 1e.

A.4. Further Details on Section 5.3

A.4.1. Responsible Data Usage

We provide additional details about data collection, usage and sharing. All assets used in the ARMMAN real data example of Section 5.3 and Appendix A.4.2, such as the covariate information, the risk scores, and the distributional estimates generated from the risk scores, is owned by ARMMAN and only ARMMAN is allowed to share this information. Beneficiaries for which their data was used to develop the data-driven simulator granted consent to have their anonymized data collected and used for research purposes, with the understanding that such data would not be released publicly. The data collection process, anonymization procedures, and potential use cases were carefully explained to the beneficiaries prior to soliciting their consent and collecting their data. All personally identifiable information about the beneficiaries was removed before sharing with the authors of this paper. We complied with all ARMMAN data exchange protocols, such as having read-access only to the anonymized data, restricted usage of the data for only our stated purpose (the analyses of Section 5.3 and Appendix A.4), and approval from ARMMAN's ethics review committee.

A.4.2. Data Description and Implementation Details

ARMMAN provided anonymized covariate information from $24,011$ beneficiaries enrolled in their maternal health program in 2022. The covariate information included 9 metrics in total, which we can categorize as follows:

- Demographic Information:
  - 'age' (continuous)
  - 'income', 'education', 'phone ownership', 'language' (categorical)
- Program Information:
  - 'gestational age' (continuous)
  - 'call slot', 'enroll delivery', 'enrollment channel' (categorical)

Through previous analyses, ARMMAN has identified 3 factors that they use to define risk: education, income, and phone ownership. For instance, beneficaries who are illiterate and who do not own their own phone are less likely to engage on average. If a beneficiary falls into the following buckets for education, income, and phone ownership, their risk score increases by 1:

- Education - 'Illiterate', '1-5', '6-9'
- Income - '0-5k', '5k-10k'
- Phone Owner - 'family phone', 'husband'.

Their final risk score is a cumulative count of the number of at-risk metrics they have, so risk scores vary from $0-3$. For instance, a beneficiary who is illiterate and does not own her own phone, but has a household income over $10k$ will have a risk score of 2. ARMMAN expects patients with risk scores 2 or 3 will benefit the most from a live call, although if the risk score is 3, the beneficiary may not have the means to act on recommendations even when those are given. Figure 18 depicts the transition dynamics used in Section 5.3.BCoR was initialized with the same prior as in Model (6). TS was initialized with a Uniform prior on $[0, 1]$, and we compute Whittle indices using a discount of $\gamma = 0.9$. We use a spline basis model over 1,000 timesteps to generate the non-stationarity using the `ps` function of `dlnm`, where the degrees of freedom were set to three and the knots were automatically selected as described and recommended in the `ps` function documentation of `dlnm` (Gasparrini, 2011). To emulate our application area, we run all methods under comparison for only the first $T = 40$ timesteps. Hence, the knots used in the spline model to generate the ARMMAN data-driven RMAB instance was set over 1,000 timesteps, but the time horizon provided to BCoR was only 40. Hence, the resulting RMAB instance has different implied knots and degrees of freedom than the spline model provided to BCoR, meaning the time model is still misspecified for our ARMMAN data-driven example. See Figure 18 for a visualization of the resulting transition dynamics.

See Figure 15 for a version of Figure 2 with all Oracles and Greedy policy versions of TS and BCoR present. We find that all three oracles performed comparably. Additionally, the Greedy versions of BCoR and TS perform similarly to their Whittle counterparts. This is sensible, since the Greedy oracle performed comparably to the Whittle oracles.

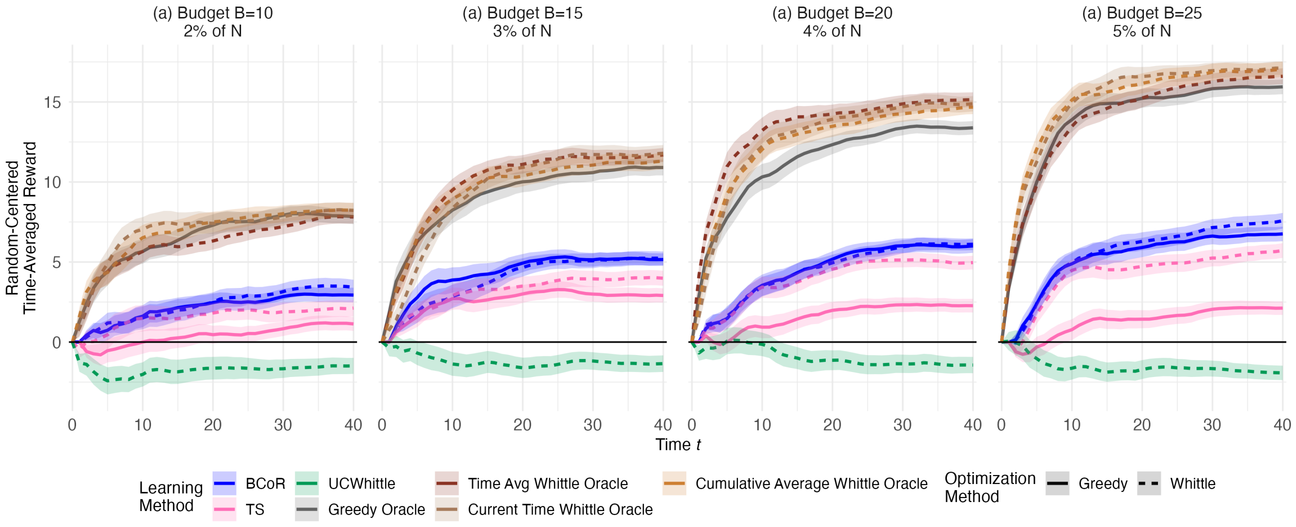


Figure 15. Performance of various methods on the ARMMAN real-data-driven example described in Section 5.3 with $N = 500, T = 40$, with varying budget B, where all $B \leq 5\%$ of N to reflect ARMMAN's true budget constraints. Lines represent the time-averaged reward of each method with the Random baseline subtracted out averaged over 100 random seeds. Note the top brown and grey methods are oracle approaches with access to the true transitions. Error bars depict ± 2 SEs.

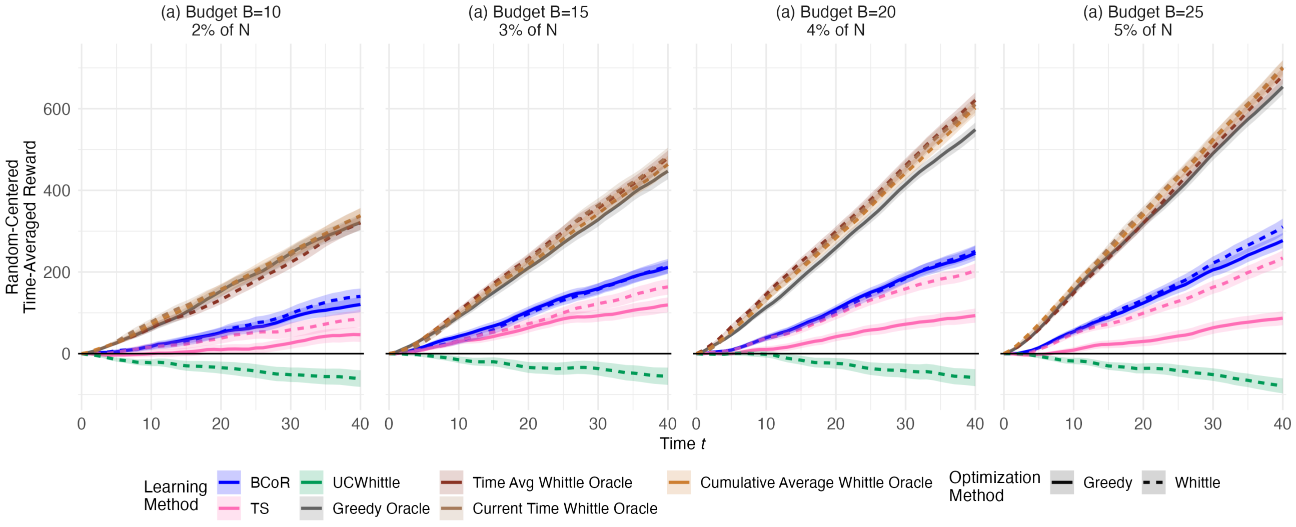


Figure 16. Performance of various methods on the ARMMAN real-data-driven example described in Section 5.3 with $N = 500, T = 40$, with varying budget B, where all $B \leq 5\%$ of N to reflect ARMMAN's true budget constraints. Lines represent the cumulative averaged reward of each method with the Random baseline subtracted out averaged over 100 random seeds. Note the top brown and grey methods are oracle approaches with access to the true transitions. BCoR consistently outperforms the other approaches. For instance, in $B = 10$ setting, BCoR-Whittle had an average random-centered cumulative reward of 140.25 by the end of the time horizon while TS-Whittle only had 87, corresponding to an increase of over 61%. Error bars depict ± 2 SEs.

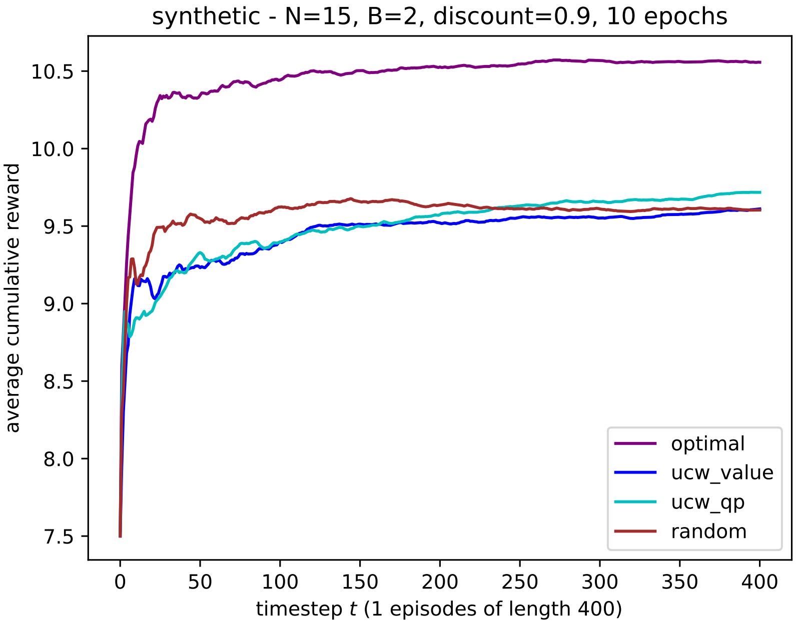


Figure 17. Plot of UCWhittle's performance on a simulated RMAB instance generated using the simulation environment in Section 7 of Wang et al. (2023), as implemented in the provided Github repository from Wang et al. (2023). "ucw_value" is the method we refer to as UCWhittle in our paper and is the approach primarily presented in Wang et al. (2023), as they establish asymptotic regret bounds for this approach. "ucw_qp" is a heuristic version of ucw_value from Wang et al. (2023), which tends to perform comparably to ucw_value in the short-term across the simulation settings in Wang et al. (2023) (Given ucw_value's theoretical guarantees, we use ucw_value as the method for comparison in our paper). The "optimal" line refers to the Whittle index policy which has access to the true transition dynamics. See Wang et al. (2023) for further implementation details. Note, "ucw_value" performs worse than random across the short term time horizons, but matches random after some time.

In our real data experiments, UCWhittle performed worse than random during the entire time horizon of 40 timesteps. We found that, even when using UCWhittle's own simulation environment from Section 7 of their paper, Wang et al. (2023), UCWhittle can perform worse than random in the short term, but eventually matches or outperforms random over time. As UCWhittle's regret bound are asymptotic, it seems that it can have poor finite-sample performance and may require longer time horizons to perform well. See Figure 17 as an example. Figure 17 can be reproduced via the repository `https://github.com/lily-x/online-rmab` and running `python main.py -N 15 -H 400 -T 1 -B 2 -D synthetic`.

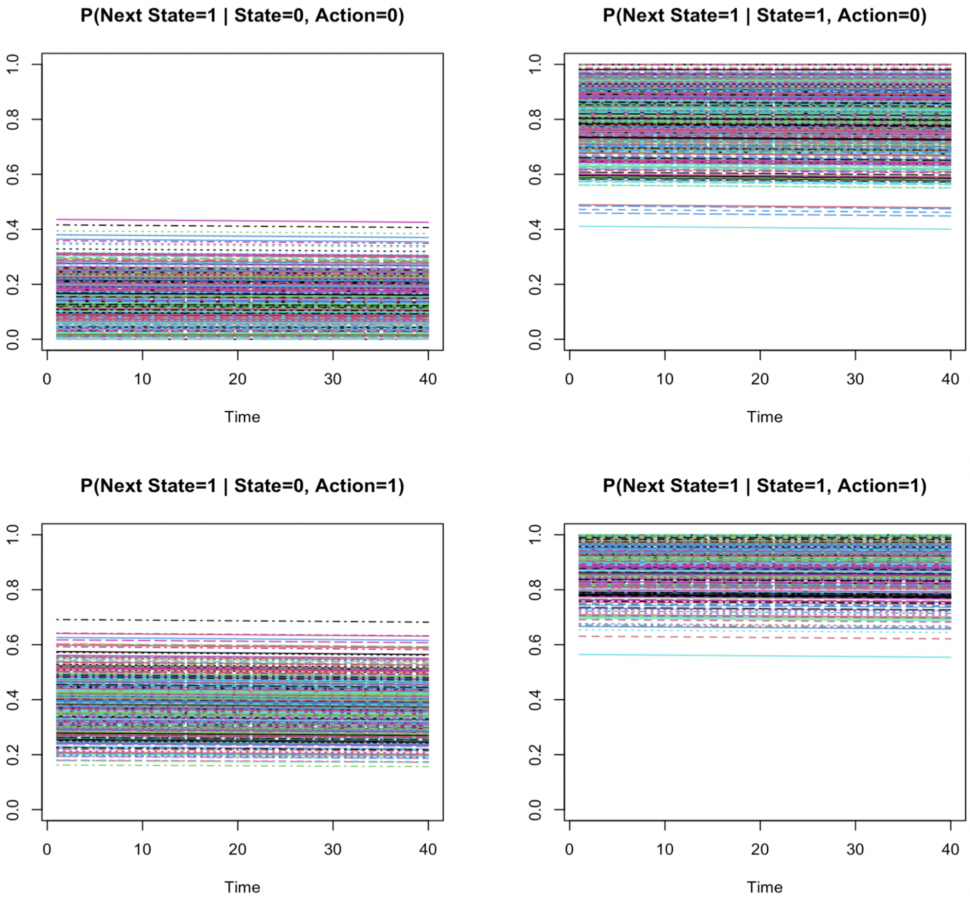


Figure 18. Underlying transition dynamics of the RMAB instance generated from our ARMMAN data-driven simulator, with $N = 500$, $T = 40$. Each line represents the transition dynamics of an individual arm. Note, the transition dynamics tend to be higher when receiving action$= 1$, and that those already in an engaging state $(s = 1)$ are more likely to stay in an engaging state than those who are in $s = 0$. We confirmed with ARMMAN representatives that these features are reasonable and reflect the dynamics we would expect among actual ARMMAN beneficiaries. Note that there is a only slight amount of non-stationarity which indicates a slight decline in adherence over time.

A.5. Accomodating Continuous Enrollment

In Algorithm 1, we choose to set a hyperprior on the random effects. We do so to help our algorithm learn efficiently even in a continuous enrollment setting, where new beneficiaries may join partway through ongoing programs. In principle we do not have any information about the new arms when they first join, but we hope to use information about previously observed arms, particularly those with similar context to the new ones, to better estimate their transition dynamics. Doing so would enable us to quickly and efficiently incorporate them into our intervention allocation. Since the parameters b_0, b_1, $\boldsymbol{\mu_\beta}$, the $\boldsymbol{\eta}^{(s,a)}$'s, and the $\boldsymbol{\beta}^{(s,a)}$'s are shared across all arms, the Bayesian model can immediately use the posterior distribution of the parameters based on previously observed arms and apply them to new arms. However, the $\alpha_j^{(s,a)}$'s are modeled *per arm*. Hence, we cannot directly use the posteriors of the $\alpha_j^{(s,a)}$'s to infer anything about the new arms. However, we can interpret the *variances* of the $\alpha_j^{(s,a)}$'s as roughly quantifying the remaining randomness that cannot be explained by the covariate and time effects. We can modify our model to learn these variances so that when a new user enters, our model knows how confident to be in its covariate- and time-based estimate of that user's transition dynamics. We accomplish this by putting a hyperprior on the variance of the random effects $\alpha_j^{(s,a)}$, which we denote by $\tau^2_{\alpha^{(s,a)}}$, effectively treating the variance of the $\alpha_j^{(s,a)}$'s as a parameter in the model.

Specifically, we model the random effects as:

$$\begin{aligned}\tau^2_{\alpha^{(s,a)}} &\sim \text{Inv-Gamma}(\tau_0, \sigma_0)\\ \alpha_i^{(s,a)} &\sim \mathcal{N}\left(0, \tau^2_{\alpha^{(s,a)}}\right).\end{aligned}$$

Since $\tau^2_{\alpha_{(s,a)}}$ is shared across all arms for each state-action pair, we can use the posterior distribution of $\tau^2_{\alpha^{(s,a)}}$ given all previously observed data for that state-action pair to infer the distribution of the random effects for the new arms when a new arm is encountered.

Hence, BCoR can use posterior distributions based on previously observed data to infer the transitions of new (unobserved) beneficiaries that join partway through an ongoing program. For instance, if a cohort of new beneficiaries joins at some intermediate time point, BCoR has already observed how previous beneficiaries behaved when they were at that point in the program. To provide a concrete example, for ARMMAN's maternal health program, if a cohort of new beneficiaries joins after an initial cohort has already passed their first trimester, BCoR will have already learned from the first cohort how the new beneficiaries are likely to behave in their first trimester. Existing online RL approaches for RMABs cannot use information about previously observed arms to infer the transition dynamics of new arms, so they must incrementally incorporate new arms, about which they initially know nothing, into their existing intervention allocation policy, making them particularly ill-suited for such a continuous enrollment setting. Hence, a continuous enrollment setting would highlight a unique advantage of BCoR, and is often realistic given our motivating application.