Optimistic Whittle Index Policy: Online Learning for Restless Bandits

Kai Wang[*†],[1] **Lily Xu**[*†],[1] **Aparna Taneja,**[2] **Milind Tambe**[1,2]

[1]Harvard University
[2]Google Research
{kaiwang, lily_xu}@g.harvard.edu, {aparnataneja, milindtambe}@google.com

Abstract

Restless multi-armed bandits (RMABs) extend multi-armed bandits to allow for stateful arms, where the state of each arm evolves restlessly with different transitions depending on whether that arm is pulled. Solving RMABs requires information on transition dynamics, which are often unknown upfront. To plan in RMAB settings with unknown transitions, we propose the first online learning algorithm based on the Whittle index policy, using an upper confidence bound (UCB) approach to learn transition dynamics. Specifically, we estimate confidence bounds of the transition probabilities and formulate a bilinear program to compute optimistic Whittle indices using these estimates. Our algorithm, *UCWhittle*, achieves sublinear $O(H\sqrt{T\log T})$ frequentist regret to solve RMABs with unknown transitions in T episodes with a constant horizon H. Empirically, we demonstrate that UCWhittle leverages the structure of RMABs and the Whittle index policy solution to achieve better performance than existing online learning baselines across three domains, including one constructed via sampling from a real-world maternal and childcare dataset.

1 Introduction

Restless multi-armed bandits (RMABs) (Whittle 1988) generalize multi-armed bandits (MABs) by introducing states for each arm. RMABs are commonly used to model sequential scheduling problems with limited resources such as clinical health (Villar, Bowden, and Wason 2015), online advertising (Meshram, Gopalan, and Manjunath 2016), and energy-efficient scheduling (Borkar et al. 2017). As with stochastic combinatorial MABs (Chen, Wang, and Yuan 2013), the RMAB learner must repeatedly pull K out of N arms at each timestep. Unlike stochastic MABs, the reward distribution of each arm in an RMAB depends on that arm's state, which transitions based on a Markov decision process (MDP) dependent on whether the arm is pulled. These problems are called "restless" as arms may change state regardless of whether they are pulled. The reward at each timestep is the sum of reward across all arms, including arms not acted upon.

Even when the transition dynamics are given, planning an optimal policy for RMABs is PSPACE-hard (Papadimitriou and Tsitsiklis 1994) due to the state-dependent reward and combinatorial action space. To compute an approximate planning solution to RMABs, the *Whittle index policy* (Whittle 1988) defines a "Whittle index" of each arm as an estimate of the future value if acted upon, then acts on the arms with the K largest indices. The Whittle index policy is shown to be asymptotically optimal (Weber and Weiss 1990) and is commonly adopted as a scalable solution to RMAB problems (Hsu 2018; Kadota et al. 2016).

However, in many real-world applications of RMABs, transition dynamics are often unknown in advance. The learner must take actions that achieve high reward while simultaneously strategically querying to learn the underlying transition probabilities. Accordingly, in this paper we focus on the challenge of online learning in fixed-length episodic RMABs with unknown transitions, studying the Whittle index policy for its scalability.

Main contributions Here, we present *UCWhittle*, the first upper confidence bound (UCB) algorithm for RMABs that uses the Whittle index policy to achieve sublinear frequentist regret. Our algorithm maintains confidence bounds for each transition probability across all arms based on prior observations. Using these bounds, we define a bilinear program to solve for optimistic transition probabilities — the transition probabilities that yield the highest future reward. These optimistic transition probabilities enable us to compute an *optimistic Whittle index* for each arm to inform a Whittle index policy. Our UCWhittle algorithm leverages the structure of RMABs and the Whittle index solution to decompose the policy across individual arms, which greatly reduces the computation cost of finding an optimistic solution compared to other UCB-based solutions (Auer and Ortner 2006; Jaksch, Auer, and Ortner 2010).

Theoretically, we analyze the frequentist regret of UCWhittle. The *frequentist regret* is defined as the regret incurred from unknown transition dynamics; in contrast, the *Bayesian regret* is the regret averaged over all possible transitions from a prior distribution. In this paper, we define *regret* in terms of the relaxed Lagrangian of the RMAB — to make the objective tractable — which upper bounds the primal RMAB problem. We show that UCWhittle achieves

[*]These authors contributed equally.
[†]Work done during an internship at Google Research.

sublinear frequentist regret $O(H\sqrt{T \log T})$ where T is the number of episodes of interaction with the RMAB instance and H is a sufficiently large horizon that the learner can receive information from each episode. Our result extends the analysis of Bayesian regret in RMABs using Thompson sampling (Jung and Tewari 2019) to frequentist regret by removing the assumption of a prior distribution. Finally, we evaluate UCWhittle against other online RMAB approaches on real data from maternal and child healthcare (Mate et al. 2022b) and two synthetic settings, showing that UCWhittle achieves lower frequentist regret empirically as well.

2 Related Work

Offline planning for RMABs When the transition dynamics are given, an RMAB is a sequential decision-making problem. Computing the optimal policy in RMABs is PSPACE-hard (Papadimitriou and Tsitsiklis 1994) due to the state-dependent reward distribution and combinatorial action space. The Whittle index policy (Whittle 1988) approximately solves the planning problem by estimating the value of each arm state. The indexability condition (Akbarzadeh and Mahajan 2019; Wang et al. 2019) guarantees asymptotical optimality (Weber and Weiss 1990) of the Whittle index policy at an infinite time horizon. Nakhleh et al. (2021) use deep reinforcement learning to estimate Whittle indices for episodic finite-horizon RMABs, which requires the environment to be differentiable. In contrast, we focus on settings where the transition dynamics are unknown, which necessitates online learning.

Online learning for RMABs When the transition dynamics are unknown, an RMAB becomes an online learning problem to simultaneously learn the transition probabilities (exploration) while executing high-reward actions (exploitation) to minimize regret with respect to a chosen benchmark. Dai et al. (2011) achieve a regret bound of $O(\log T)$, with T timesteps of interaction with the RMAB instance, benchmarked against an optimal policy from a finite number of potential policies. Xiong, Li, and Singh (2022) use a Lagrangian relaxation and optimistic index-based algorithm, but require access to a powerful offline simulator to generate samples for any given (state, action) pair. Tekin and Liu (2012) define a weaker benchmark of the best single-action policy — the optimal policy that continues to play the same arm — and use an upper confidence bound (UCB) based algorithm to achieve $O(\log T)$ frequentist regret.

Recent works introduce oracle-based policies for the non-combinatorial setting that pulls only a single arm in each round with bandit feedback, where only the state of the pulled arm is observed. Jung and Tewari (2019) use a Thompson sampling–based algorithm to show a Bayesian regret bound $O(\sqrt{T \log T})$ under a given prior distribution. Wang, Huang, and Lui (2020) split into an exploration and an exploitation phase to achieve frequentist regret $O(T^{2/3})$. These works assume some policy oracle is given, benchmarking regret with the policy induced by the oracle using the true transition. In contrast to the meta-algorithms they propose, *we design an optimal approach custom-tailored to one specific oracle — based on the Whittle index policy — which enables us to achieve a tighter frequentist regret bound of* $O(H\sqrt{T \log T})$ with a constant horizon H.

Online reinforcement learning RMABs are a special case of Markov decision processes (MDPs) with combinatorial state and action spaces. Q-learning algorithms are popular for solving large MDPs and have been applied to standard binary-action RMABs (Avrachenkov and Borkar 2022; Fu et al. 2019; Biswas et al. 2021) and extended to the multi-action setting (Killian et al. 2021). However, these works do not provide any regret guarantee. Significant work has explored online learning for stochastic multi-armed bandits (Neu and Bartók 2013; Immorlica et al. 2019; Foster and Rakhlin 2020; Baek and Farias 2020; Xu et al. 2021), but these do not allow arms to change state.

Some papers study online reinforcement learning by using the optimal policy as the benchmark to bound regret in MDPs (Auer and Ortner 2006; Jaksch, Auer, and Ortner 2010) and RMABs (Ortner et al. 2012). These works use UCB-based algorithms (UCRL and UCRL2) to obtain a regret of $O(\sqrt{T \log T})$. However, calculating regret with respect to the optimal policy requires computing the optimal solution to the RMAB problem, which is intractable due to the combinatorial space and action spaces. To overcome this difficulty, we restrict the benchmark for computing regret to the class of Whittle index threshold policies, and leverage the weak-decomposability of the Whittle index threshold policy to establish a new regret bound.

3 Restless Bandits and Whittle Index Policy

An instance of a restless multi-armed bandit (RMAB) problem is composed of a set of N arms. Each arm $i \in [N]$ is modeled as an independent Markov decision process (MDP) defined by a tuple $(\mathcal{S}, \mathcal{A}, R, P_i)$. The state space $\mathcal{S}$, action space $\mathcal{A}$, and reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are shared across arms; the transition probability $P_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ may be unique per arm i.

We denote the state of the RMAB instance at timestep $h \in \mathbb{N}$ by $\boldsymbol{s}_h \in \mathcal{S}^N$, where $s_{h,i}$ denotes the state of arm $i \in [N]$. We assume the state is fully observable. The initial state is given by $\boldsymbol{s}_1 = \boldsymbol{s}_{\text{init}} \in \mathcal{S}^N$. The action (an "arm pull") at time h is denoted by a binary vector $\boldsymbol{a}_h \in \mathcal{A}^N = \{0, 1\}^N$ and is constrained by budget K such that $\sum_{i \in [N]} s_{h,i} \leq K$.

After taking action $a_{h,i}$ on arm i, the state $s_{h,i}$ transitions to the next state $s_{h+1,i}$ with transition probability $P_i(s_{h,i}, a_{h,i}, s_{h+1,i}) \in [0, 1]$. We denote the set of all transition probabilities by $\boldsymbol{P} = [P_i]_{i \in [N]}$. The learner receives reward $R(s_{h,i}, a_{h,i})$ from each arm i (including those not acted upon) at every timestep h; we assume the reward function R is known to the learner.

The learner's actions are described by a deterministic policy $\pi : \mathcal{S}^N \to \mathcal{A}^N$, which maps a given state $\boldsymbol{s} \in \mathcal{S}^N$ to an action $\boldsymbol{a} \in \mathcal{A}^N$. The learner's goal is to optimize the total discounted reward, with discount factor $\gamma \in (0, 1)$:

$$\max_{\pi} \quad \mathop{\mathbb{E}}_{(\boldsymbol{s},\boldsymbol{a}) \sim (\boldsymbol{P}, \pi)} \sum\nolimits_{h \in \mathbb{N}} \gamma^{h-1} \sum\nolimits_{i \in [N]} R(s_{h,i}, a_{h,i})$$
$$\text{s.t.} \quad \sum\nolimits_{i \in [N]} (\pi(\boldsymbol{s}))_i = K \quad \forall \boldsymbol{s} \in \mathcal{S}^N \tag{1}$$

3.1 Lagrangian Relaxation

Equation 1 is intractable to optimize and evaluate over all possible policies, thus a poor candidate objective for evaluating online learning performance. Instead, we relax the constraints to use the Lagrangian as the evaluation metric:

$$U_{\pi}^{\boldsymbol{P},\lambda}(\boldsymbol{s}_1) \coloneqq \mathop{\mathbb{E}}_{(\boldsymbol{s},\boldsymbol{a})\sim(\boldsymbol{P},\pi)} \sum\nolimits_{h\in\mathbb{N}} \gamma^{h-1}\left(\sum_{i\in[N]} R(s_{h,i},a_{h,i}) - \lambda\left(\sum_{i\in[N]}(\pi(\boldsymbol{s}_h))_i - K\right)\right) \tag{2}$$

which permits actions that exceed the budget constraint, subject to a given penalty λ. The optimal value of Equation 2, which we denote $U_{\star}^{\boldsymbol{P},\lambda}$, is always an upper bound to Equation 1. Therefore, we solve Equation 2 for every λ and find the infimum $\lambda^{\star} = \arg\min_{\lambda} U_{\star}^{\boldsymbol{P},\lambda}$ afterward.

3.2 Whittle Index and Threshold Policy

Relaxing the budget constraint enables us to decompose the combinatorial policy into a set of N independent policies for each arm. The decoupled policy yields $\pi(\boldsymbol{s}) = [\pi_i(s_i)]_{i\in[N]}$, where each arm policy $\pi_i : \mathcal{S} \to \mathcal{A}$ specifies the action for arm i given state s_i. The value function is then:

$$V_{\pi_i}^{P_i,\lambda}(s_{1,i}) \coloneqq \mathop{\mathbb{E}}_{(s_{h,i},a_{h,i})\sim(P_i,\pi_i)\ \forall h} \sum\nolimits_{h\in\mathbb{N}} \gamma^{h-1}\left(R(s_{h,i},a_{h,i}) - \lambda\left(\pi_i(s_{h,i}) - K\right)\right). \tag{3}$$

Equation 3 can be interpreted as adding a penalty λ to the pulling action $a = 1$, which motivates the definition of Whittle index (Whittle 1988) as the smallest penalty for an arm such that pulling that arm is equally good as not pulling:

Definition 3.1. Given transition probabilities P_i and state s_i, the *Whittle index* of arm i is defined as:

$$W_i(P_i,s_i) = \inf_{m_i}\{m_i : Q^{m_i}(s_i,0) = Q^{m_i}(s_i,1)\} \tag{4}$$

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

$$Q^{m_i}(s,a) = -m_i a + R(s,a) + \gamma\sum_{s'\in\mathcal{S}} P_i(s,a,s')V^{m_i}(s')$$
$$V^{m_i}(s) = \max_{a\in\mathcal{A}} Q^{m_i}(s,a)\ .$$

When the Whittle index $W_i(P_i,s_i)$ for an arm is higher than the chosen global penalty λ — that is, $m_i > \lambda$ — the optimal policy in Equation 3 is to pull the arm, i.e., $\pi_i(s_i) = 1$. We denote the Whittle indices of all arms and all states by $W(\boldsymbol{P}) = [W_i(P_i,s_i)]_{i\in[N],s_i\in\mathcal{S}} \in \mathbb{R}^{N\times|\mathcal{S}|}$.

Definition 3.2 (Whittle index threshold policy)**.** Given a chosen global penalty λ and the Whittle indices $W(\boldsymbol{P})$ computed from transitions $\boldsymbol{P}$, the threshold policy is defined by:

$$\pi_{W(\boldsymbol{P}),\lambda}(\boldsymbol{s}) = [\mathbb{1}_{W_i(P_i,s_i)\geq\lambda}]_{i\in[N]} \in \mathcal{A}^N\ . \tag{5}$$

which pulls all the arms with Whittle indices larger than λ.

The Whittle index threshold policy may violate the budget constraints in Equation 1, but it maximizes the relaxed Lagrangian in Equation 2 under penalty λ.

4 Problem Statement: Online Learning in RMABs

In this paper, we consider the online setting where the true transition probabilities $\boldsymbol{P}^{\star}$ are unknown to the learner. The learner interacts with an RMAB instance across multiple episodes, and only requires observations for the first H timesteps of each episode to estimate transition probabilities. The observations of the transitions are maintained in a historical dataset which we denote $D = \{(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\}$.

At the beginning of each episode $t \geq 1$, the learner starts the RMAB instance ($h = 1$) from $\boldsymbol{s}_1 = \boldsymbol{s}_{\text{init}}$ and selects a new policy $\pi^{(t)}$. We make the following assumptions:

- Each episode is infinite horizon with discount factor γ.
- In each episode t, the learner proposes a policy $\pi^{(t)}$. The learner observes the first H timesteps[1], but receives the infinite discounted reward $U_{\pi^{(t)}}^{\boldsymbol{P},\lambda}(\boldsymbol{s}_1)$ to account for the long-term effect of $\pi^{(t)}$.
- We assume the MDP associated with each arm is *ergodic*. That is, starting from the given initial state, we assume H is large enough such that after H timesteps, there is at least $\varepsilon > 0$ probability of reaching any state $s \in \mathcal{S}$.

4.1 Offline Benchmark and Regret

To evaluate performance of our policy $\pi^{(t)}$, we compare against a Whittle index threshold policy $\pi_{W(\boldsymbol{P}^{\star}),\lambda}$ trained with knowledge of the true transitions $\boldsymbol{P}^{\star}$ and a given penalty λ. This offline benchmark measures the advantage gained from knowing the true transitions $\boldsymbol{P}^{\star}$. We measure *regret* against this full-information benchmark.

Definition 4.1 (Frequentist regret of the Lagrangian objective)**.** Given a penalty λ and the true transitions $\boldsymbol{P}^{\star}$, we define the *regret* of the policy $\pi^{(t)}$ in episode t relative to the optimal policy $\pi^{\star} = \pi_{W(\boldsymbol{P}^{\star}),\lambda}$:

$$\text{Reg}_{\lambda}^{(t)} \coloneqq U_{\pi^{\star}}^{\boldsymbol{P}^{\star},\lambda}(\boldsymbol{s}_1) - U_{\pi^{(t)}}^{\boldsymbol{P}^{\star},\lambda}(\boldsymbol{s}_1)\ ,$$
$$\text{Reg}_{\lambda}(T) \coloneqq \sum\nolimits_{t\in[T]} \text{Reg}_{\lambda}^{(t)}\ . \tag{6}$$

However, the relaxed Lagrangian in Equation 2 with a randomly chosen penalty λ may not be a good proxy to the primal RMAB problem in Equation 1. Therefore, we define the Lagrangian using the optimal Lagrangian multiplier $\lambda^{\star}$ as the tightest upper bound of Equation 1.

Definition 4.2 (Frequentist regret of the optimal Lagrangian objective)**.** Given $\boldsymbol{P}^{\star}$, we denote the optimal penalty by $\lambda^{\star} = \arg\min_{\lambda} U_{\pi^{\star}}^{\boldsymbol{P}^{\star},\lambda}(\boldsymbol{s}_1)$. The regret of the *optimal Lagrangian objective* is defined by:

$$\text{Reg}_{\lambda^{\star}}^{(t)} \coloneqq U_{\pi^{\star}}^{\boldsymbol{P}^{\star},\lambda^{\star}}(\boldsymbol{s}_1) - U_{\pi^{(t)}}^{\boldsymbol{P}^{\star},\lambda^{\star}}(\boldsymbol{s}_1)\ ,$$
$$\text{Reg}_{\lambda^{\star}}(T) \coloneqq \sum\nolimits_{t\in[T]} \text{Reg}_{\lambda^{\star}}^{(t)}\ . \tag{7}$$

While our algorithm requires satisfying the strict budget constraint, the expected regret is approximated by the regret from the relaxed Lagrangian in Equation 2 as defined in Definition 4.1 and Definition 4.2.

[1] In practice, infinite time horizon means a large horizon that is much larger than H.

4.2 Frequentist Versus Bayesian Regret

Note that the definitions of regret that we consider are *frequentist* regret, which measures worst-case regret under unknown transition probabilities. The other regret notion is Bayesian regret, calculated as the expected value over a prior distribution of possible transitions. We bound the frequentist regret of each instance of the RMAB problem with a constant dependence on the ergodicity of the problem instance; approaches to calculate Bayesian regret rely on a prior, such as Thompson sampling–based methods (Jung and Tewari 2019; Jung, Abeille, and Tewari 2019).

5 UCWhittle: Optimistic Whittle Index Threshold Policy

A key challenge to online learning in RMABs is that the confidence bounds of the estimated transitions indirectly impact future reward, affecting the future state and reward distribution. We introduce a method, UCWhittle, to compute optimistic Whittle indices that account for highest future value.

5.1 Confidence Bounds of Transition Probabilities

We maintain confidence bounds for every unknown transition probability in the RMAB instance. Specifically we maintain counts $N_i^{(t)}(s,a,s')$ for every state, action, and next state transition observed by episode t. Let $N_i^{(t)}(s,a) = \sum_{s'\in\mathcal{S}} N_i^{(t)}(s,a,s')$.

Given a chosen small constant $\delta > 0$, we estimate each transition probability $P_i(s,a,s')$ with empirical mean

$$\hat{P}_i^{(t)}(s,a,s') \coloneqq \frac{N_i^{(t)}(s,a,s')}{N_i^{(t)}(s,a)} \tag{8}$$

and confidence radius

$$d_i^{(t)}(s,a) \coloneqq \sqrt{\frac{2|\mathcal{S}|\log(2|\mathcal{S}||\mathcal{A}|N\frac{t^4}{\delta})}{\max\{1, N_i^{(t)}(s,a)\}}}\,. \tag{9}$$

With these confidence bounds, we specify the open ball $\boldsymbol{B}$ of possible values for transition probabilities $\boldsymbol{P}$ as

$$\boldsymbol{B}^{(t)} = \left\{ \boldsymbol{P} \,|\, \left\| P_i(s,a,\cdot) - \hat{P}_i^{(t)}(s,a,\cdot) \right\|_1 \le d_i^{(t)}(s,a)\ \forall i,s,a \right\}.$$

5.2 Optimistic Transitions and Whittle Indices

To translate confidence bounds in transition probabilities to the actual reward, we define an optimization problem ($\mathcal{P}_V$) to find for each arm i the *optimistic* transition probability $P_i^\dagger$, the value within the confidence bound that yields the *highest future value* from the starting state s_i:

$$\max_{V,Q,P_i\in\boldsymbol{B}_i^{(t)}} V(s_i) \qquad \text{s.t. } V(s) = \max_{a\in\mathcal{A}} Q(s,a) \qquad (\mathcal{P}_V)$$

$$Q(s,a) = -\lambda a + R(s,a) + \gamma \sum\nolimits_{s'\in\mathcal{S}} P_i(s,a,s')V(s')$$

We prove Equation ($\mathcal{P}_V$) to be optimal in Section 6.

We use the optimistic transition $P_i^\dagger$ to compute the corresponding *optimistic Whittle index* $W_i^\dagger = [W(P_i^\dagger)]_{s\in\mathcal{S}}$. The Whittle index threshold policy $\pi_i^\dagger = \pi_{W_i^\dagger,\lambda}$ achieves the same value function derived from the transition $P_i^\dagger$, which maximizes Equation ($\mathcal{P}_V$). Aggregating all the arms together, optimistic policy $\pi^\dagger$ with optimistic transitions $\boldsymbol{P}^\dagger$ maximizes the future value of the current state $\boldsymbol{s}$.

5.3 UCWhittle Algorithm

Having computed optimistic transitions and the corresponding optimistic Whittle indices ($\mathcal{P}_m$), we construct an optimistic Whittle index threshold policy to execute. The full algorithm is outlined in Algorithm 1, and implementation details — including novel techniques for speeding up computation of Whittle index — are given in Appendix F.1.

Algorithm 1: UCWhittle

Input: N arms, K budget, episode horizon H
Initialize counts $N_i^{(t)} = 0$ for all s, a, s'
Randomly initialize penalty $\lambda^{(1)}$
for episode $t \in \{1, 2, \ldots\}$ **do**
  Reset $h = 1$ and $\boldsymbol{s} = \boldsymbol{s}_{\text{init}}$ ▷ *Reset RMAB instance*
  $P_i^\dagger = \mathcal{P}_V(s_i, N_i^{(t)}, \lambda^{(t)})$ for all $i \in [N]$ ▷ *Compute an optimistic transition for each arm*
  $W_i = \textsc{ComputeWI}(P_i^\dagger, s_i)$ for all $i \in [N]$ ▷ *Compute Whittle indices using Def. 3.1*
  Execute $\pi^{(t)}$ for H steps by pulling arms with the top K Whittle indices. Observe transitions $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$.
  Update counts $N_i^{(t)}$, empirical means $\hat{\boldsymbol{P}}^{(t)}$, and confidence regions $\boldsymbol{B}^{(t)}$
  $\lambda^{(t+1)} = K$th highest Whittle index ▷ *Update penalty*
end for

5.4 Alternative Formulation for Whittle Index Upper Bound

Equation ($\mathcal{P}_V$) is computationally expensive, so we formulate a heuristic optimization to compute an upper bound on the Whittle index. We solve for the transition probability and the value function that yield the highest *Whittle index* (instead of highest *future value*) at the current state $s_{h,i}$:

$$\max_{m_i,V,Q,P_i,\in\boldsymbol{B}_i^{(t)}} m_i \qquad (\mathcal{P}_m)$$

$$\text{s.t. } V(s) = \max_{a\in\mathcal{A}} Q(s,a),\ \ Q(s_{h,i}, a{=}0) = Q(s_{h,i}, a{=}1)$$

$$Q(s,a) = -m_i a + R(s,a) + \gamma \sum\nolimits_{s'\in\mathcal{S}} P_i(s,a,s')V(s')$$

This optimization problem differs slightly from the optimization in Equation ($\mathcal{P}_V$) with a different objective, additional variables m_i representing the penalty for each arm, and added Whittle index constraint.

Solving Equation ($\mathcal{P}_m$) directly gives us the maximal Whittle index estimate within the confidence bound. We thus save computation cost while maintaining a valid upper bound to the optimistic Whittle index from Equation ($\mathcal{P}_V$), which requires an intermediate step to compute the optimistic Whittle index. However, the theoretical analysis

does not hold for the heuristic. Empirically, we show that this heuristic achieves comparable performance with significantly lower computation.

6 Regret Analysis

We provide regret guarantees of our UCWhittle algorithm. In the following section, we use the Lagrangian objective as a proxy to the reward received from the proposed policy. Section 6.1 first assumes an arbitrary penalty λ is given to define the regret (Definition 4.1). Section 6.2 generalizes to define the regret of optimal Lagrangian objective based on the unknown optimal penalty $\lambda^\star$ (Definition 4.2). Section 6.3 provides an update rule for updating the penalty $\lambda^{(t)}$ after each episode. Full proofs are given in Appendix D.

6.1 Regret Bound with Known Penalty

By the Chernoff bound, we know that with high probability the true transition $\boldsymbol{P}^\star$ lies within $\boldsymbol{B}^{(t)}$:

Proposition 6.1. *Given $\delta > 0$ and $t \geq 0$, we have:* $\Pr\left(\boldsymbol{P}^\star \in \boldsymbol{B}^{(t)}\right) \geq 1 - \frac{\delta}{t^4}$.

This bound can be used to bound the regret incurred when the confidence bound fails. In the following theorem, we bound the regret in the case where the confidence bound holds and when the penalty λ is given.

Theorem 6.2 (Regret decomposition)**.** *Given the penalty λ and $\boldsymbol{P}^\star \in \boldsymbol{B}^{(t)}$ for all t, we have:*

$$\begin{aligned} \mathit{Reg}_\lambda(T) &= \sum\nolimits_{t\in[T]} U_{\pi^\star}^{\boldsymbol{P}^\star,\lambda}(\boldsymbol{s}_1) - U_{\pi^{(t)}}^{\boldsymbol{P}^\star,\lambda}(\boldsymbol{s}_1) \\ &\leq \sum\nolimits_{t\in[T]} U_{\pi^{(t)}}^{\boldsymbol{P}^{(t)},\lambda}(\boldsymbol{s}_1) - U_{\pi^{(t)}}^{\boldsymbol{P}^\star,\lambda}(\boldsymbol{s}_1)\,. \end{aligned} \tag{10}$$

Proof. By optimality of Equation ($\mathcal{P}_V$) to enable $(P_i^{(t)}, \pi_i^{(t)}) = \arg\max_{P_i \in B_i^{(t)}, \pi_i} V_{\pi_i}^{P_i,\lambda}(s_{1,i})$ and the assumption that the true transition lies within the confidence region $P_i^\star \in B_i^{(t)}$, we show that:

$$\begin{aligned} U_{\pi^\star}^{\boldsymbol{P}^\star,\lambda}(\boldsymbol{s}_1) &= \sum\nolimits_{i\in[N]} V_{\pi_i^\star}^{P_i^\star,\lambda}(s_{1,i}) \\ &\leq \sum\nolimits_{i\in[N]} V_{\pi_i^{(t)}}^{P_i^{(t)},\lambda}(s_{1,i}) = U_{\pi^{(t)}}^{\boldsymbol{P}^{(t)},\lambda}(\boldsymbol{s}_1)\,. \end{aligned}$$ □

Theorem 6.2 enables us to bound our regret by the difference between two future values under the same policy $\pi^{(t)}$.

Definition 6.3 (Bellman operator)**.** Define the *Bellman operator* as:

$$\mathcal{T}_{\pi_i}^{P_i} V(s) = \mathop{\mathbb{E}}_{a\sim\pi_i}\left[-\lambda a + R(s,a) + \gamma \sum_{s'\in S} P_i(s,a,s')V(s')\right]$$

Using Theorem 6.2 and the Bellman operator, we can further decompose the regret by:

Theorem 6.4 (Per-episode regret decomposition in the fully observable setting)**.** *For an arm i, fix $P_i^{(t)}$, $P_i^\star$, λ, and the initial state $s_{1,i}$. We have:*

$$\begin{aligned} &V_{\pi_i^{(t)}}^{P_i^{(t)},\lambda}(s_{1,i}) - V_{\pi_i^{(t)}}^{P_i^\star,\lambda}(s_{1,i}) = \\ &\mathop{\mathbb{E}}_{P_i^\star,\pi_i^{(t)}}\left[\sum_{h=1}^{\infty} \gamma^{h-1}\left(\mathcal{T}_{\pi_i^{(t)}}^{P_i^{(t)}} - \mathcal{T}_{\pi_i^{(t)}}^{P_i^\star}\right) V_{\pi_i^{(t)}}^{P_i^{(t)},\lambda}(s_{h,i})\right]. \end{aligned} \tag{11}$$

Theorem 6.4 further decomposes the regret in Equation 10 into individual differences in Bellman operators. The next theorem bounds the differences in Bellman operators by differences in transition probabilities.

Theorem 6.5. *Assume the penalty term $\lambda^{(t)} = \lambda$ is given and the RMAB instance is ε-ergodicity after H timesteps. The following bound on the cumulative regret in T episodes holds with probability $1-\delta$:*

$$\mathit{Reg}_\lambda(t) \leq O\left(\frac{1}{\varepsilon}|S||A|^{\frac{1}{2}} NH\sqrt{T\log T}\right). \tag{12}$$

Proof sketch. We focus on bounding the regret when the confidence bounds hold. By Theorem 6.2 and Theorem 6.4, we estimate the right-hand side of Equation 11 to bound the total regret by the L^1-difference in the transition probability:

$$\begin{aligned} &\sum\nolimits_{h=1}^{\infty} \gamma^{h-1}\left(\mathcal{T}_{\pi_i^{(t)}}^{P_i^{(t)}} - \mathcal{T}_{\pi_i^{(t)}}^{P_i^\star}\right) V_{\pi_i^{(t)}}^{P_i^{(t)}}(s_{h,i}) \\ &\leq \sum\nolimits_{h=1}^{\infty} \gamma^{h-1}\left\| P_i^{(t)}(s_{h,i},a_{h,i},\cdot) - P_i^\star(s_{h,i},a_{h,i},\cdot)\right\|_1 V_{\max}. \end{aligned} \tag{13}$$

We bound the regret outside of the horizon H by the ergodic assumption of the MDPs. For the regret inside the horizon H, we use the confidence radius to bound the L^1-norm of transition probability differences and count the number of observations for each state-action pair to express the regret as a sequence of random variables, whose sum can be bounded by Lemma D.3 to conclude the proof. □

When the penalty term λ is given, Theorem 6.5 bounds the frequentist regret with a constant term depending on the ergodicity ε of the underlying true MDPs.

6.2 Regret Bound with Unknown Optimal Penalty

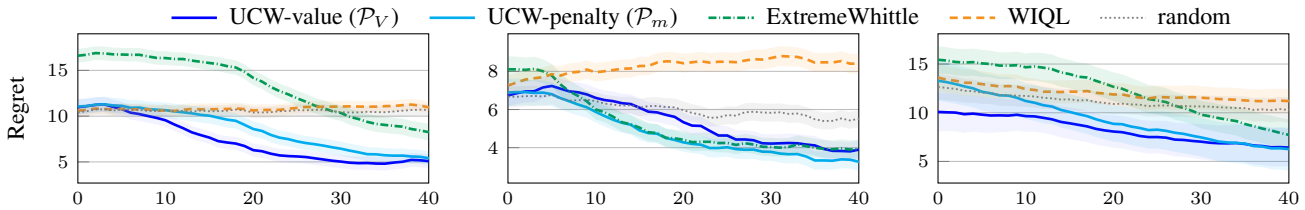
The analysis in Theorem 6.2 assumes a fixed and given penalty λ. In this section, we generalize to regret defined in terms of the optimal but unknown penalty $\lambda^\star$ (Definition 4.2). We show that updating penalty $\lambda^{(t)}$ in Algorithm 1 achieves the same regret bound without requiring knowledge of the true transitions $\boldsymbol{P}^\star$ or optimal penalty $\lambda^\star$:

Theorem 6.6 (Regret bound with optimal penalty)**.** *Assume the penalty $\lambda^{(t)}$ in Algorithm 1 is updated by a saddle point $(\lambda^{(t)}, \boldsymbol{P}^{(t)}, \pi^{(t)}) = \arg\min_\lambda \max_{\boldsymbol{P},\pi} U_\pi^{\boldsymbol{P},\lambda}(\boldsymbol{s}_1)$ subject to constraints in Equation ($\mathcal{P}_V$). The cumulative regret of the optimal Lagrangian objective is bounded with probability $1-\delta$:*

$$\mathit{Reg}_{\lambda^\star}(t) \leq O\left(\frac{1}{\varepsilon}|S||A|^{\frac{1}{2}} NH\sqrt{T\log T}\right). \tag{14}$$

Proof sketch. The main challenge of an unknown penalty term $\lambda^\star$ is that the optimality of the chosen transition $\boldsymbol{P}^{(t)}$ and policy $\pi^{(t)}$ does not hold in Theorem 6.2 due to the misalignment of the penalty $\lambda^{(t)}$ used in solving Equation ($\mathcal{P}_V$) and the penalty $\lambda^\star$ used in the regret.

Surprisingly, the optimality of $(\lambda^{(t)}, \boldsymbol{P}^{(t)}, \pi^{(t)}) = \arg\min_\lambda \max_{\boldsymbol{P},\pi} U_\pi^{\boldsymbol{P},\lambda}(\boldsymbol{s}_1)$ and $\lambda^\star = \inf_\lambda U_{\pi^\star}^{\boldsymbol{P}^\star,\lambda}(\boldsymbol{s}_1)$ is



(a) ARMMAN (b) Wide Margin (c) Thin Margin

Figure 1: Cumulative discounted regret (lower is better) in each episode (x-axis) incurred by our UCWhittle approaches compared to baselines across the three domains with $N = 8$ arms, budget $B = 3$, episode length $H = 20$, and $T = 40$ episodes.

sufficient to show Theorem 6.2 by:

$$\underbrace{U^{\boldsymbol{P}^\star,\lambda^\star}_{\pi^\star}}_{\lambda^\star \text{ minimizes } U^{\boldsymbol{P}^\star,\lambda}_{\pi^\star}} \leq \underbrace{U^{\boldsymbol{P}^\star,\lambda^{(t)}}_{\pi^\star} \leq U^{\boldsymbol{P}^{(t)},\lambda^{(t)}}_{\pi^{(t)}}}_{\boldsymbol{P}^{(t)},\,\pi^{(t)} \text{ maximizes } U^{\boldsymbol{P},\lambda^{(t)}}_{\pi}} \leq \underbrace{U^{\boldsymbol{P}^{(t)},\lambda^\star}_{\pi^{(t)}}}_{\lambda^{(t)} \text{ minimizes } U^{\boldsymbol{P}^{(t)},\lambda}_{\pi^{(t)}}}$$

$$\implies \text{Reg}^{(t)}_{\lambda^\star} = U^{\boldsymbol{P}^\star,\lambda^\star}_{\pi^\star} - U^{\boldsymbol{P}^\star,\lambda^\star}_{\pi^{(t)}} \leq U^{\boldsymbol{P}^{(t)},\lambda^\star}_{\pi^{(t)}} - U^{\boldsymbol{P}^\star,\lambda^\star}_{\pi^{(t)}}. \tag{15}$$

where we omit the dependency on $\boldsymbol{s}_1$.

After taking summation over $t \in [T]$, Equation 15 leads to the same result as shown in Theorem 6.2 but the penalty update rule does not use the knowledge of the optimal penalty $\lambda^\star$. The rest of the proof follows the same argument in Theorem 6.4 and Theorem 6.5 with the same regret bound. □

6.3 Penalty Update Rule

Theorem 6.6 suggests that the penalty term $\lambda^{(t)}$ should be defined by solving a minimax problem $(\lambda^{(t)}, \boldsymbol{P}^{(t)}, \pi^{(t)}) = \arg\min_\lambda \max_{\boldsymbol{P},\pi} U^{\boldsymbol{P},\lambda}_\pi(\boldsymbol{s}_1)$. However, solving a minimax problem with a bilinear objective is difficult. A heuristic solution is to solve the maximization problem using the previous penalty $\lambda^{(t-1)}$ to determine $\boldsymbol{P}^{(t)}$ and $\pi^{(t)}$ (Equation ($\mathcal{P}_V$)). We update $\lambda^{(t)}$ based on the current policy, set equal to the Kth largest Whittle index pulled at time t to minimize the Lagrangian. This update rule mimics the minimax update rule required by Theorem 6.6.

7 Experiments

We show that UCWhittle achieves consistently low regret across three domains, including one generated from real-world data on maternal health. Additional details about the dataset and data usage are shown in Appendix E, and details about implementation (including novel techniques to speed up computation) and experiments are shown in Appendix F.

7.1 Preliminaries

Domains We consider three binary-action, binary-state settings. Across all domains, the binary states are *good* or *bad*, with reward 1 and 0 respectively. We impose two assumptions: that acting is always beneficial (more likely to transition to the good state), and that it is always better to start from the good state (more likely to stay in good state). The first domain is an online learning environment constructed from the ARMMAN dataset, and the other two are synthetic environments sampling from specific ranges of transition probabilities.

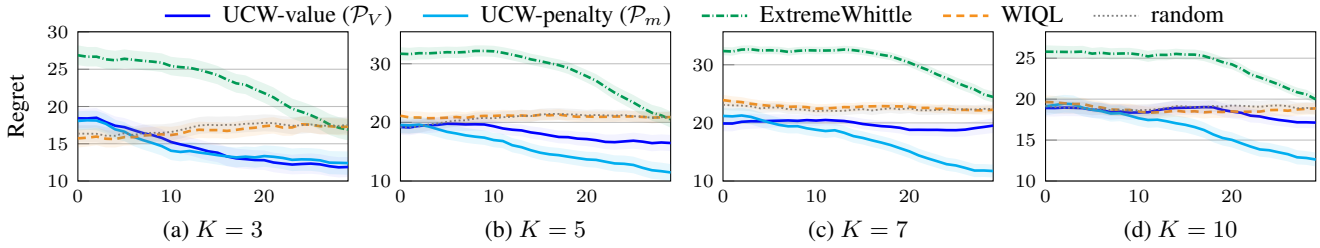
ARMMAN (2022) is a non-profit based in India that disseminates health information to pregnant women and mothers to reduce mortality and morbidity for mothers and their children. Twice a week, ARMMAN sends automated voice messages to enrolled beneficiaries relaying critical preventative health information. To improve listenership, the organization provides service calls to a subset of beneficiaries; the challenge is selecting which subset to call to maximize engagement. We use a real-world anonymized dataset based on the engagement behavior of 7,656 mothers from a previous RMAB field study (Mate et al. 2022b). We construct instances of RMAB problem with transition probabilities randomly sampled from the real dataset.

Wide Margin We randomly generate transition probabilities with high variance, while respecting the constraints specified above.

Thin Margin For a more challenging setting, we consider a synthetic domain with probabilities of transitioning to the good state constrained to the interval $[0.2, 0.4]$ to test the ability of each approach to discern smaller differences in transition probabilities.

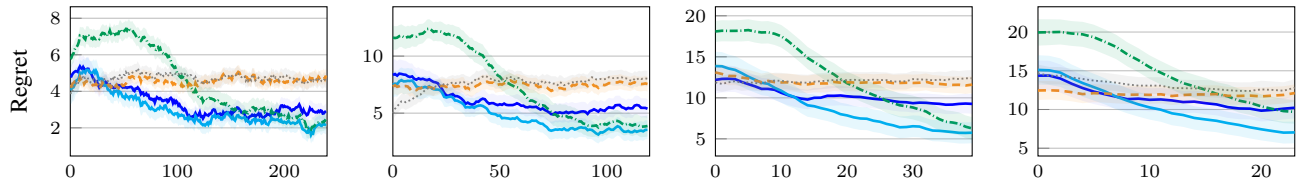
Algorithms We evaluate both variants of UCWhittle (Algorithm 1) introduced in this paper. *UCWhittle-value* uses the value-maximizing bilinear program ($\mathcal{P}_V$) while *UCWhittle-penalty* uses the penalty-maximizing bilinear program ($\mathcal{P}_m$).

In this paper, we focus on frequentist regret, thus we exclude the Bayesian regret baselines, e.g., Thompson sampling (Jung and Tewari 2019), because their regret bounds are averaged over a prior. We consider the following three regret baselines: *ExtremeWhittle* is similar to the the approach by Wang et al. (2019): estimate Whittle indices from the extreme points, using UCBs of active transition probabilities and lower confidence bounds (LCB) for passive transition probabilities to estimate the gap between the value of acting versus not acting. We then solve a Whittle index policy using these estimates. *WIQL* (Biswas et al. 2021) uses Q-learning to learn the value function of each arm at each state by interacting with the RMAB instance. *Random* takes a random action in each step, serving as a baseline for expected reward without using any strategic learning algorithm. Lastly, we evaluate an *optimal* policy which computes a Whittle index policy with access to the true transition probabilities.



(a) $K = 3$ (b) $K = 5$ (c) $K = 7$ (d) $K = 10$

Figure 2: Varying budget ratio K/N, with $N = 15$ arms, on the ARMMAN domain. Our UCWhittle approaches perform stronger than baselines, particularly in the challenging low-budget scenarios.



(a) $H = 5$ (b) $H = 10$ (c) $H = 30$ (d) $H = 50$

Figure 3: Changing episode length H on the ARMMAN domain. We run each setting for $1,200$ total timesteps. *UCW-penalty* performs best with longer horizons. At shorter horizons, *UCW-value* converges in fewer timesteps, but more episodes are necessary: around episode $t = 100$ with a horizon $H = 5$ compared to episode $t = 16$ with horizon $H = 50$.

Experiment setup We evaluate the performance of each algorithm across T episodes of length H. The per-episode reward is the cumulative discounted reward with discount rate $\gamma = 0.9$. We then compute regret by subtracting the reward earned by each algorithm from the reward of the *optimal* policy. Results are averaged over 30 random seeds and smoothed using exponential smoothing with a weight of 0.9. We ensure consistency by enforcing, across all algorithms, identical populations (transition probabilities for each arm) and initial state for each episode.

7.2 Results

The performance results across all three domains are shown in Figure 1. Our UCWhittle algorithm using the value-maximizing bilinear program (*UCW-value*) achieves consistently strong performance and generally converges by 600 timesteps (across varying episode lengths). In Figures 2 and 3 we evaluate performance while varying the budget K and episode length H, as the regret of UCWhittle (Theorem 6.5) has dependency on both the budget as a ratio of total number of arms (K/N) and episode length H. We see that UCW-value performs comparatively stronger than the baselines in the challenging low-budget settings, in which each arm pull has greater impact.

Our heuristic approach *UCW-penalty* — the penalty-maximizing bilinear program we present in Eq. ($\mathcal{P}_m$) — shows strong performance. UCW-penalty performs even better than UCW-value in some settings, particularly in the ARMMAN domain with $N = 15$ arms (Figure 2). Notably in Table 1 we see this heuristic approach performs dramatically faster than UCW-value — a $6.1\times$ speedup. Therefore while are able to establish regret guarantees only for UCW-value, we also propose UCW-penalty as a strong candidate for its strong performance and quick execution.

Method	**Time (s)**
UCWhittle-value	1090.92
UCWhittle-penalty	177.57
ExtremeWhittle	109.44
WIQL	3.39
random	1.32

Table 1: Average runtime of the different approaches across 500 timesteps with $N = 30$ arms and budget $B = 6$

In Figures 2 and 3 we see *ExtremeWhittle* has poor performance particularly in the early episodes, consistently achieving higher regret than the random policy. Additionally, *WIQL* is slow to converge, performing similarly to the random baseline across the time horizons that we consider.

8 Conclusion

We propose the first online learning algorithm for RMABs based on the Whittle index policy, using an upper confidence bound–approach to learn transition dynamics. We formulate a bilinear program to compute optimistic Whittle indices from the confidence bounds of transition dynamics, enabling an online learning algorithm using an optimistic Whittle index threshold policy. Theoretically, our work pushes the boundary of existing frequentist regret bounds in RMABs while maintaining scalability by leveraging decomposition in the Whittle index threshold policy.

References

Akbarzadeh, N.; and Mahajan, A. 2019. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 7294–7300. IEEE.

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

Auer, P.; and Ortner, R. 2006. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 19.

Avrachenkov, K. E.; and Borkar, V. S. 2022. Whittle index based Q-learning for restless bandits with average reward. *Automatica*, 139: 110186.

Baek, J.; and Farias, V. F. 2020. TS-UCB: Improving on Thompson Sampling With Little to No Additional Computation. *arXiv preprint arXiv:2006.06372*.

Biswas, A.; Aggarwal, G.; Varakantham, P.; and Tambe, M. 2021. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*.

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

Chen, W.; Wang, Y.; and Yuan, Y. 2013. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, 151–159. PMLR.

Dai, W.; Gai, Y.; Krishnamachari, B.; and Zhao, Q. 2011. The non-Bayesian restless multi-armed bandit: A case of near-logarithmic regret. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2940–2943. IEEE.

Foster, D.; and Rakhlin, A. 2020. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning (ICML)*, 3199–3210. PMLR.

Fu, J.; Nazarathy, Y.; Moka, S.; and Taylor, P. G. 2019. Towards Q-learning the Whittle index for restless bandits. In *2019 Australian & New Zealand Control Conference (ANZCC)*, 249–254. IEEE.

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

Immorlica, N.; Sankararaman, K. A.; Schapire, R.; and Slivkins, A. 2019. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, 202–219. IEEE.

Jaksch, T.; Auer, P.; and Ortner, R. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11: 1563–1600.

Jung, Y. H.; Abeille, M.; and Tewari, A. 2019. Thompson sampling in non-episodic restless bandits. *arXiv preprint arXiv:1910.05654*.

Jung, Y. H.; and Tewari, A. 2019. Regret bounds for Thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Kadota, I.; Uysal-Biyikoglu, E.; Singh, R.; and Modiano, E. 2016. Minimizing the age of information in broadcast wireless networks. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 844–851. IEEE.

Killian, J. A.; Biswas, A.; Shah, S.; and Tambe, M. 2021. Q-Learning Lagrange Policies for Multi-Action Restless Bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, 871–881.

Mate, A.; Biswas, A.; Siebenbrunner, C.; and Tambe, M. 2022a. Efficient algorithms for finite horizon and streaming restless multi-armed bandit problems. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Mate, A.; Madaan, L.; Taneja, A.; Madhiwalla, N.; Verma, S.; Singh, G.; Hegde, A.; Varakantham, P.; and Tambe, M. 2022b. Field Study in Deploying Restless Multi-Armed Bandits: Assisting Non-Profits in Improving Maternal and Child Health. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*.

Meshram, R.; Gopalan, A.; and Manjunath, D. 2016. Optimal recommendation to users that react: Online learning for a class of POMDPs. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 7210–7215. IEEE.

Nakhleh, K.; Ganji, S.; Hsieh, P.-C.; Hou, I.; Shakkottai, S.; et al. 2021. NeurWIN: Neural whittle index network for restless bandits via deep RL. *Advances in Neural Information Processing Systems*, 34: 828–839.

Neu, G.; and Bartók, G. 2013. An efficient algorithm for learning with semi-bandit feedback. In *International Conference on Algorithmic Learning Theory (ALT)*, 234–248. Springer.

Ortner, R.; Ryabko, D.; Auer, P.; and Munos, R. 2012. Regret bounds for restless Markov bandits. In *International Conference on Algorithmic Learning Theory (ALT)*, 214–228. Springer.

Papadimitriou, C. H.; and Tsitsiklis, J. N. 1994. The complexity of optimal queueing network control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*, 318–322. IEEE.

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

Spielman, D. A. 2007. Spectral graph theory and its applications. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 29–38. IEEE.

Tekin, C.; and Liu, M. 2012. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8): 5588–5611.

Villar, S. S.; Bowden, J.; and Wason, J. 2015. 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.

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

Wang, S.; Huang, L.; and Lui, J. 2020. Restless-UCB, an Efficient and Low-complexity Algorithm for Online Restless Bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 11878–11889.

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

Weissman, T.; Ordentlich, E.; Seroussi, G.; Verdu, S.; and Weinberger, M. J. 2003. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*.

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

Xiong, G.; Li, J.; and Singh, R. 2022. Reinforcement Learning Augmented Asymptotically Optimal Index Policy for Finite-Horizon Restless Bandits. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Xu, L.; Bondi, E.; Fang, F.; Perrault, A.; Wang, K.; and Tambe, M. 2021. Dual-Mandate Patrols: Multi-Armed Bandits for Green Security. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*.

A Notation

All the notations used in the problem statement, restless multi-armed bandits, and regret analysis are shown in Table 2 and Table 3.

Problem instantiation	
Symbol	**Definition**
K	Budget in each timestep
N	Number of arms. Each arm indexed by $i \in [N]$
t	Episode
T	Number of episodes
h	Timestep within a single episode
H	Horizon length for a single episode
γ	Discount factor, with $\gamma \in (0, 1)$

Table 2: List of common notations in the problem statement

Restless bandit notation	
Symbol	**Definition**
$\boldsymbol{P}$	Set of transition probabilities across all arms, with P_i as transitions for a single arm
$\boldsymbol{P}^\star$	True transition probabilities
$\mathcal{S}$	Set of finitely many possible states with $\lvert\mathcal{S}\rvert = M$ possible states
$\boldsymbol{s}_h$	State of the RMAB instance at timestep h, with $\boldsymbol{s}_h \in \mathcal{S}^N$ and initial state $\boldsymbol{s}_{\text{init}}$
$s_{h,i}$	State of arm $i \in [N]$ at timestep h
$\mathcal{A}$	Set of possible actions. We consider $\{0, 1\}$
$\boldsymbol{a}_h$	Action at time h, with $\boldsymbol{a}_h \in \mathcal{A}^N$
$a_{h,i}$	Action taken on arm i at timestep h
R	Given reward function as a function of the state and action $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.
$\pi^{(t)}$	Learner's policy in episode t, where $\pi^{(t)} : \mathcal{S}^N \to \mathcal{A}^N$
$\pi^\star$	The optimal policy that maximizes the total future reward.
$\mathcal{P}_m$	The optimization problem defined to maximize the optimistic Whittle index value.
$\mathcal{P}_V$	The optimization problem defined to maximize the optimistic future value.
$Q^{m_i}(s, a)$	Q-value in Bellman equation. The Q-value is defined as the future value associated to the current state and action.
$R(s_{h,i}, a_{h,i})$	Reward from arm i at timestep h with action $a_{h,i}$
$U_\pi^{\boldsymbol{P},\lambda}(\boldsymbol{s}_1)$	Lagrangian relaxation of learner's objective, with optimal value $U_\star^{\boldsymbol{P},\lambda}$
$V_{\pi_i}^{P_i,\lambda}(s_{1,i})$	Value for being in state s_i
λ	Global penalty for taking action $a = 1$
m_i	Whittle index penalty for arm i
$W_i(P_i, s_i)$	Whittle index of arm i with transitions P_i and state s_i
$\mathcal{D}$	Dataset of historical transitions

Table 3: List of common notations in the RMAB regret analysis

B Societal Impacts

Restless bandits have been increasingly applied to socially impactful problems including healthcare and energy distribution. In these settings, we would likely not know the transition dynamics in advance, particularly if we are working with a new patient population (for healthcare) or new residential community (for energy). Even past work on streaming bandits (Mate et al. 2022a) which allow for new beneficiaries to enroll over time assume that the transition probabilities are fully known in advance, which is not realistic. Our UCWhittle approach enabling online learning for RMABs has the potential to greatly broaden the applicability of RMABs for social impact, particularly as our theoretical results guarantee limited regret.

C Limitations

One challenge with our UCWhittle approach is that online learning often converges slower than offline learning that reuses all the data to train for many epochs. In order to accommodate new data coming in, online learning approaches often take a single

update when each new data arrives. In contrast, offline learning can iterate through the same data for many times, which allows offline learning approaches to fit the data repeatedly. Therefore, online learning approaches often require more data to reach the same performance as offline learning approaches.

However, this slower learning behavior also allows online learning approaches to be less biased to the existing dataset. Online learning approaches are incentivized to explore and update data that is less queried previously, which also encourages exploring underrepresented groups. This property encourages the exploration process and reduce bias to the learned model. This is particularly important when there are features involved in the learning process. Online learning approaches are able to explore unseen features more, while offline learning approaches often rely on extrapolation and are unable to handle unseen features. Our work further extends research in online learning in RMABs, which also helps explore more possibility to accommodate new data and new features that are unseen in the existing dataset.

D Full Proofs

D.1 Confidence Bound

Proposition 6.1. *Given $\delta > 0$ and $t \geq 0$, we have:* $\Pr\left(\boldsymbol{P}^\star \in \boldsymbol{B}^{(t)}\right) \geq 1 - \frac{\delta}{t^4}$.

Proof. Generally, the L1-deviation of the true distribution and the empirical distribution over m distinct events from n samples is bounded according to (Weissman et al. 2003) by:

$$\Pr(\|\hat{p} - p\|_1 \geq \varepsilon) \leq (2^m - 2)\exp^{(-\frac{n\varepsilon^2}{2})} \tag{16}$$

This result can be applied to our case to compare $P_i^{(t)}(s,a,\cdot) \in \mathbb{R}^{|S|}$ with $P^\star(s,a,\cdot) \in \mathbb{R}^{|S|}$ for every state s and action a. We have:

$$\Pr\left(\left\|P_i^{(t)}(s,a,\cdot) - P^\star(s,a,\cdot)\right\|_1 \geq \varepsilon\right) \leq \left(2^{|S|} - 2\right)\exp^{\left(-\frac{n\varepsilon^2}{2}\right)} \tag{17}$$

By choosing $\varepsilon = \sqrt{\frac{2}{n}\log\left(2^{|S|}|S||A|N\frac{t^4}{\delta}\right)} \leq \sqrt{\frac{2|S|}{n}\log\left(2|S||A|N\frac{t^4}{\delta}\right)}$, we have:

$$\begin{aligned}\Pr\left(\left\|P_i^{(t)}(s,a,\cdot) - P^\star(s,a,\cdot)\right\|_1 \geq \sqrt{\frac{2|S|}{n}\log\left(2|S||A|N\frac{t^4}{\delta}\right)}\right) &\leq 2^{|S|}\exp^{-\log\left(2^{|S|}|S||A|N\frac{t^4}{\delta}\right)} \\ &= \frac{\delta}{|S||A|Nt^4}\end{aligned} \tag{18}$$

Set $n = \max\{1, N_i^{(t)}(s,a)\}$ for each pair of (s,a). Taking union bound over all states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, and arms $i \in [N]$ yields:

$$\Pr\left(\boldsymbol{P}^\star \notin \boldsymbol{B}^{(t)}\right) \leq \frac{\delta}{t^4} \quad \Longrightarrow \quad \Pr\left(\boldsymbol{P}^\star \in \boldsymbol{B}^{(t)}\right) \geq 1 - \frac{\delta}{t^4} \tag{19}$$

□

D.2 Regret Decomposition

Theorem 6.4 (Per-episode regret decomposition in the fully observable setting)**.** *For an arm i, fix $P_i^{(t)}$, $P_i^\star$, λ, and the initial state $s_{1,i}$. We have:*

$$\begin{aligned}&V_{\pi_i^{(t)}}^{P_i^{(t)},\lambda}(s_{1,i}) - V_{\pi_i^{(t)}}^{P_i^\star,\lambda}(s_{1,i}) = \\ &\mathop{\mathbb{E}}_{P_i^\star,\pi_i^{(t)}}\left[\sum_{h=1}^{\infty}\gamma^{h-1}\left(\mathcal{T}_{\pi_i^{(t)}}^{P_i^{(t)}} - \mathcal{T}_{\pi_i^{(t)}}^{P_i^\star}\right)V_{\pi_i^{(t)}}^{P_i^{(t)},\lambda}(s_{h,i})\right].\end{aligned} \tag{11}$$

Proof. Since the value function is a fixed point to the corresponding Bellman operator, we have:

$$\begin{aligned}V_{\pi_i^{(t)}}^{P_i^{(t)}}(s_{1,i}) - V_{\pi_i^{(t)}}^{P_i^\star}(s_{1,i}) &= \left(\mathcal{T}_{\pi_i^{(t)}}^{P_i^{(t)}}V_{\pi_i^{(t)}}^{P_i^{(t)}} - \mathcal{T}_{\pi_i^{(t)}}^{P_i^\star}V_{\pi_i^{(t)}}^{P_i^\star}\right)(s_{1,i}) \\ &= \left(\mathcal{T}_{\pi_i^{(t)}}^{P_i^{(t)}} - \mathcal{T}_{\pi_i^{(t)}}^{P_i^\star}\right)V_{\pi_i^{(t)}}^{P_i^{(t)}}(s_{1,i}) + \mathcal{T}_{\pi_i^{(t)}}^{P_i^\star}\left(V_{\pi_i^{(t)}}^{P_i^{(t)}} - V_{\pi_i^{(t)}}^{P_i^\star}\right)(s_{1,i})\end{aligned} \tag{20}$$

where the second term in Equation (20) can be further expanded by the Bellman operator:

$$\begin{aligned}\mathcal{T}_{\pi_i^{(t)}}^{P_i^\star}(V_{\pi_i^{(t)}}^{P_i^{(t)}} - V_{\pi_i^{(t)}}^{P_i^\star})(s_{1,i}) &= \mathbb{E}_{a\sim\pi_i^{(t)}}\left[\gamma\sum\nolimits_{s'\in\mathcal{S}} P_i^\star(s_{1,i},a,s')(V_{\pi_i^{(t)}}^{P_i^{(t)}}(s') - V_{\pi_i^{(t)}}^{P_i^\star}(s'))\right] \\ &= \gamma\mathbb{E}_{s_{2,i}\sim P_i^\star,\pi_i^{(t)}}\left[V_{\pi_i^{(t)}}^{P_i^{(t)}}(s_{2,i}) - V_{\pi_i^{(t)}}^{P_i^\star}(s_{2,i})\right] \end{aligned} \tag{21}$$

We can repeatedly apply the decomposition process in Equation (20) to the value function difference in Equation (21) to get Equation (11), which concludes the proof. □

D.3 Regret Bound with Given Penalty

Theorem 6.5. *Assume the penalty term $\lambda^{(t)} = \lambda$ is given and the RMAB instance is ε-ergodicity after H timesteps. The following bound on the cumulative regret in T episodes holds with probability $1-\delta$:*

$$\mathit{Reg}_\lambda(t) \le O\left(\frac{1}{\varepsilon}|S||A|^{\frac{1}{2}}NH\sqrt{T\log T}\right) . \tag{12}$$

Proof. We can write

$$\begin{aligned}\mathrm{Reg}(T) &= \sum\nolimits_{t=1}^{T}\mathrm{Reg}^{(t)} = \sum\nolimits_{t=1}^{T}\left(\mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\notin\boldsymbol{B}^{(t)}} + \mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\in\boldsymbol{B}^{(t)}}\right) \\ &= \sum\nolimits_{t=1}^{T}\mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\notin\boldsymbol{B}^{(t)}} + \sum\nolimits_{t=1}^{T}\mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\in\boldsymbol{B}^{(t)}}\end{aligned} \tag{22}$$

We will analyze both terms separately and combine them together in the end.

Regret when the confidence bounds do not hold

$$\begin{aligned}\sum\nolimits_{t=1}^{T}\mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\notin\boldsymbol{B}^{(t)}} &= \sum\nolimits_{t=1}^{\sqrt{T}}\mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\notin\boldsymbol{B}^{(t)}} + \sum\nolimits_{t=\sqrt{T}+1}^{T}\mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\notin\boldsymbol{B}^{(t)}} \\ &\le \frac{NR_{\max}}{1-\gamma}\sqrt{T} + \sum\nolimits_{t=\sqrt{T}+1}^{T}\mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\notin\boldsymbol{B}^{(t)}}\end{aligned} \tag{23}$$

where we use the trivial upper bound of the individual regret $\mathrm{Reg}^{(t)} \le \frac{NR_{\max}}{1-\gamma}$ for all t, where $R_{\max}$ is the maximal reward per time step.

Notice that the second term vanishes with probability:

$$\begin{aligned}\Pr\left(\left\{\boldsymbol{P}^\star\in\boldsymbol{B}^{(t)}\ \forall\sqrt{T}\le t\le T\right\}\right) &\ge 1 - \sum\nolimits_{\sqrt{T}\le t\le T}\Pr\left(\left\{\boldsymbol{P}^\star\in\boldsymbol{B}^{(t)}\right\}\right) \\ &\ge 1 - \sum\nolimits_{\sqrt{T}\le t\le T}\frac{\delta}{t^4} \\ &\ge 1 - \sum\nolimits_{\sqrt{T}\le t\le T}\frac{3\delta}{t^4} \\ &\ge 1 - \int_{\sqrt{T}}^{\infty}\frac{3\delta}{t^4}dt \\ &= 1 - \frac{\delta}{T^{3/2}}\end{aligned} \tag{24}$$

Therefore, the regret outside of confidence bounds is upper bounded by $O(\sqrt{T})$ with probability at least $1-\frac{\delta}{T^{3/2}}$. We can apply union bound to all possible $T\in\mathbb{N}$, which holds with high probability:

$$1 - \sum\nolimits_{T=1}^{\infty}\frac{\delta}{T^{3/2}} = 1 - O(\delta) . \tag{25}$$

Regret when the confidence bounds hold Notice that

$$\begin{aligned}&\left(\mathcal{T}_{\pi_i^{(t)}}^{P_i^{(t)}} - \mathcal{T}_{\pi_i^{(t)}}^{P_i^\star}\right)V(s) \\ =& \mathop{\mathbb{E}}_{a\sim\pi_i^{(t)}}\left[\left(R(s,a) + \sum\nolimits_{s'\in\mathcal{S}}P_i^{(t)}(s,a,s')V(s')\right) - \left(R(s,a) + \sum\nolimits_{s'\in\mathcal{S}}P_i^\star(s,a,s')\right)V(s')\right] \\ =& \mathop{\mathbb{E}}_{a\sim\pi_i^{(t)}}\left[\sum\nolimits_{s'\in\mathcal{S}}(P_i^{(t)}(s,a,s') - P_i^\star(s,a,s'))V(s')\right]\end{aligned}$$

When the confidence bound holds $\boldsymbol{P}^\star \in \boldsymbol{B}^{(t)}$, we can bound the regret at round l by:

$$\begin{aligned}
\mathrm{Reg}^{(t)} &= U^{\boldsymbol{P}^{(t)}}_{\pi^{(t)}}(\boldsymbol{s}_1) - U^{\boldsymbol{P}^\star}_{\pi^{(t)}}(\boldsymbol{s}_1) \\
&= \sum_{i=1}^N V^{P_i^{(t)}}_{\pi_i^{(t)}}(s_{1,i}) - V^{P_i^\star}_{\pi_i^{(t)}}(s_{1,i}) \\
&= \sum_{i=1}^N \mathbb{E}_{P_i^\star,\pi_i^{(t)}}\left[\sum_{h=1}^\infty \gamma^{h-1}(\mathcal{T}^{P_i^{(t)}}_{\pi_i^{(t)}} - \mathcal{T}^{P_i^\star}_{\pi_i^{(t)}})V^{P_i^{(t)}}_{\pi_i^{(t)}}(s_{h,i})\right] \\
&= \sum_{i=1}^N \mathbb{E}_{P_i^\star,\pi_i^{(t)}} \sum_{h=1}^\infty \sum_{s'\in\mathcal{S}} \gamma^{h-1}(P_i^{(t)}(s_{h,i},a_{h,i},s') - P_i^\star(s_{h,i},a_{h,i},s'))V^{P_i^{(t)}}_{\pi_i^{(t)}}(s') \\
&\le \sum_{i=1}^N \mathbb{E}_{P_i^\star,\pi_i^{(t)}} \sum_{h=1}^\infty \gamma^{h-1} \left\| P_i^{(t)}(s_{h,i},a_{h,i},\cdot) - P_i^\star(s_{h,i},a_{h,i},\cdot)\right\|_1 V_{\max} \\
&\le 2\sum_{i=1}^N \mathbb{E}_{P_i^\star,\pi^{(t)}} \sum_{h=1}^\infty \gamma^{h-1} d_i^{(t)}(s_{h,i},a_{h,i}) V_{\max}
\end{aligned} \tag{26}$$

Next, we split the term into regret within H horizon and the regret outside of H horizon. By applying Theorem D.2 with the assumption (Assumption D.1) of the H-step ergodicity ε of MDP associated to arm i, we can bound the regret outside of H horizon by the regret at H time step:

$$\begin{aligned}
&\mathbb{E}_{P_i^\star,\pi^{(t)}} \sum_{h=H+1}^\infty \gamma^{h-1} d_i^{(t)}(s_{h,i},a_{h,i}) V_{\max} \\
=& \sum_{h=H+1}^\infty \gamma^{h-1} \mathbb{E}_{s_{h,i},a_{h,i}\sim P_i^\star,\pi_i^{(t)}}\left[d_i^{(t)}(s_{h,i},a_{h,i})V_{\max}\right] \\
\le& \sum_{h=H+1}^\infty \gamma^{h-1} \frac{1}{\varepsilon} \mathbb{E}_{s_{H,i},a_{H,i}\sim P_i^\star,\pi_i^{(t)}}\left[d_i^{(t)}(s_{h,i},a_{h,i})V_{\max}\right] \\
=& \frac{\gamma^H}{\varepsilon(1-\gamma)} \mathbb{E}_{s_{H,i},a_{H,i}\sim P_i^\star,\pi_i^{(t)}}\left[d_i^{(t)}(s_{h,i},a_{h,i})V_{\max}\right]
\end{aligned} \tag{27}$$

Now, we can further bound the contribution of arm i in Equation 26 by substituting the regret after H steps by Equation 27 to get:

$$\begin{aligned}
&\mathbb{E}_{P_i^\star,\pi_i^{(t)}} \sum_{h=1}^\infty \gamma^{h-1} d_i^{(t)}(s_{h,i},a_{h,i}) V_{\max} \\
\le& \mathbb{E}_{P_i^\star,\pi_i^{(t)}} \left(\sum_{h=1}^H \gamma^{h-1} d_i^{(t)}(s_{h,i},a_{h,i}) + \frac{\gamma^H}{\delta(1-\gamma)} d_i^{(t)}(s_{H,i},a_{H,i})\right) V_{\max} \\
\le& \left(1 + \frac{\gamma^H}{\varepsilon(1-\gamma)}\right) \mathbb{E}_{P_i^\star,\pi_i^{(t)}} \left(\sum_{h=1}^H d_i^{(t)}(s_{h,i},a_{h,i}) V_{\max}\right) \\
=& \left(1 + \frac{\gamma^H}{\varepsilon(1-\gamma)}\right) \sqrt{2|S|\log(2|A|Nt)} V_{\max} \mathbb{E}_{P_i^\star,\pi_i^{(t)}} \left(\sum_{h=1}^H \frac{1}{\sqrt{\max\{1, N_i^{(t)}(s,a)\}}}\right) \\
\le& \left(1 + \frac{\gamma^H}{\varepsilon(1-\gamma)}\right) \sqrt{2|S|\log(2|A|NT)} V_{\max} \mathbb{E}_{P_i^\star,\pi_i^{(t)}} \left(\sum_{s\in\mathcal{S},a\in\mathcal{A}} \frac{v_i^{(t)}(s,a)}{\sqrt{\max\{1, N_i^{(t)}(s,a)\}}}\right)
\end{aligned} \tag{28}$$

where $v_i^{(t)}(s,a)$ is a random variable denoting the number of visitations to the pair (s,a) at arm i that the policy $\pi_i^{(t)}$ visits within H steps under the transition probability $P_i^\star$.

Recall that $\sum_{j=1}^{l-1} v_i^{(j)}(s,a) = N_i^{(t)}(s,a)$. We also know that $0 \le v_i^{(j)}(s,a) \le H$. Applying Lemma D.3, we have:

$$\sum_{t=1}^T \frac{v_i^{(t)}(s,a)}{\sqrt{\max\{1, N_i^{(t)}(s,a)\}}} \le \left(\sqrt{H+1}+1\right)\sqrt{N_i^{(t)}(s,a)} \tag{29}$$

Taking summation over all the (s, a) pairs and applying Jensen inequality give us:

$$\begin{aligned}
&\left(\sqrt{H+1}+1\right)\sum\nolimits_{s\in\mathcal{S},a\in\mathcal{A}}\sqrt{N_i^{(t)}(s,a)} \\
\leq&\left(\sqrt{H+1}+1\right)|S||A|\sqrt{\frac{\sum_{s\in\mathcal{S},a\in\mathcal{A}}N_i^{(t)}(s,a)}{|S||A|}} \\
=&\left(\sqrt{H+1}+1\right)\sqrt{|S||A|TH} \qquad (30)
\end{aligned}$$

where $\sum_{s\in\mathcal{S},a\in\mathcal{A}}N_i^{(t)}(s,a)=TH$ is the total number of state-action pairs visited in T rounds.

Lastly, using the trivial upper bound $V_{\max}\leq\frac{R_{\max}}{1-\gamma}$, we can take summation over the regret from all T rounds. This give us:

$$\begin{aligned}
&\sum\nolimits_{t=1}^{T}\mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\in\boldsymbol{B}^{(t)}} \qquad (31)\\
\leq&\sum\nolimits_{i=1}^{N}2\left(1+\frac{\gamma^H}{\varepsilon(1-\gamma)}\right)\sqrt{2|S|\log(2|A|NT)}V_{\max}\left(\sqrt{H+1}+1\right)\sqrt{|S||A|TH} \\
\leq& O\left(\frac{1}{\varepsilon}|S||A|^{\frac{1}{2}}NH\sqrt{T\log T}\right) \qquad (32)
\end{aligned}$$

Combining everything together In the first part, we show that $\sum_{t=1}^{T}\mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\notin\boldsymbol{B}^{(t)}}$ is upper bounded by $O(\sqrt{T})$ for all $T\in\mathbb{N}$ with probability $1-O(\delta)$. In the second part, we show that $\sum_{t=1}^{T}\mathrm{Reg}^{(t)}\mathbb{1}_{\boldsymbol{P}^\star\in\boldsymbol{B}^{(t)}}=O(|S||A|^{\frac{1}{2}}N\sqrt{T\log T})$. Therefore, we can conclude that the total regret $\mathrm{Reg}(T)$ is upper bounded by $O(|S||A|^{\frac{1}{2}}N\sqrt{T\log T})$ for all $T\in\mathbb{N}$ with probability $1-O(\delta)$. □

D.4 Supplementary Lemma and Theorem

Assumption D.1 (Ergodic Markov chain)**.** We denote $u_h^{P_i^\star,\pi_i}$ to be the state distribution of Markov chain induced by the MDP with transition probability $P_i^\star$ and policy π_i after h time steps. We assume $u_h^{P_i^\star,\pi_i}(s)>\varepsilon>0$ for all entry $s\in\mathcal{S}$, all arm $i\in[N]$, $h\geq H$, and all policy π_i. In other words, the state distribution after H steps is universally lower-bounded by $\varepsilon>0$, which we say that the MDP is H-step ε-ergodic.

Assumption D.1 can be achieved when both the MDP is ergodic and the horizon H is large enough.

Theorem D.2 (Regret outside of H steps)**.** *When the Markov chain induced by transition $P_i^\star$ and policy π is H-step ε ergodic, we have:*

$$\mathbb{E}_{s_{h,i},a_{h,i}\sim P_i^\star,\pi}f(s_{h,i},a_{h,i})\leq\frac{1}{\varepsilon}\mathbb{E}_{s_{H,i},a_{H,i}\sim P_i^\star,\pi}f(s_{H,i},a_{H,i}) \qquad (33)$$

for all non-negative function f and $h\geq H$.

Proof.

$$\begin{aligned}
\mathbb{E}_{s_{h,i},a_{h,i}\sim P_i^\star,\pi}f(s_{h,i},a_{h,i})=&\sum\nolimits_{s\sim S,a\sim A}\Pr(\pi_i(s)=a)u_h(s)f(s,a) \\
\leq&\sum\nolimits_{s\sim S,a\sim A}\Pr(\pi_i(s)=a)f(s,a) \\
\leq&\sum\nolimits_{s\sim S,a\sim A}\Pr(\pi_i(s)=a)\frac{u_H(s)}{\varepsilon}f(s,a) \\
=&\frac{1}{\varepsilon}\sum\nolimits_{s\sim S,a\sim A}\Pr(\pi_i(s)=a)u_H(s)f(s,a) \\
=&\frac{1}{\varepsilon}E_{s_{H,i},a_{H,i}\sim P_i^\star,\pi}f(s_{H,i},a_{H,i}) \qquad (34)
\end{aligned}$$

□

Lemma D.3. *For any sequence of numbers $z_1,\cdots,z_T$ with $0\leq z_j\leq H$ and $Z_t=\max\{1,\sum_{j=1}^{t}z_j\}$, we have:*

$$\sum\nolimits_{t=1}^{T}\frac{z_t}{\sqrt{Z_{t-1}}}\leq\left(\sqrt{H+1}+1\right)\sqrt{Z_T} \qquad (35)$$

Proof. Proof by induction. Assume that Equation 35 holds for $T-1$. We have:

$$\sum_{t=1}^{T-1} \frac{z_t}{\sqrt{Z_{t-1}}} \leq \left(\sqrt{H+1}+1\right)\sqrt{Z_{T-1}} \tag{36}$$

Adding an additional term $\frac{z_T}{\sqrt{Z_{T-1}}}$, we get:

$$\begin{aligned}
\sum_{t=1}^{T-1} \frac{z_t}{\sqrt{Z_{t-1}}} + \frac{z_T}{\sqrt{Z_{T-1}}} &\leq \left(\sqrt{H+1}+1\right)\sqrt{Z_{T-1}} + \frac{z_T}{\sqrt{Z_{T-1}}} \\
&= \sqrt{\left(\sqrt{H+1}+1\right)^2 Z_{T-1} + 2\left(\sqrt{H+1}+1\right) z_T + \frac{z_T^2}{Z_{T-1}}} \\
&\leq \sqrt{\left(\sqrt{H+1}+1\right)^2 Z_{T-1} + 2\left(\sqrt{H+1}+1\right) z_T + H z_T} \\
&\leq \sqrt{\left(\sqrt{H+1}+1\right)^2 Z_{T-1} + \left(\sqrt{H+1}+1\right)^2 z_T} \\
&\leq \left(\sqrt{H+1}+1\right)\sqrt{Z_{T-1}+z_T} \\
&= \left(\sqrt{H+1}+1\right)\sqrt{Z_T}
\end{aligned} \tag{37}$$

which implies the Equation 35 also holds for T.

The initial case with $T=1$ holds trivially. Therefore, by induction, we conclude the proof. □

D.5 Regret Bound with Unknown Optimal Penalty

Theorem 6.6 (Regret bound with optimal penalty)**.** *Assume the penalty $\lambda^{(t)}$ in Algorithm 1 is updated by a saddle point $(\lambda^{(t)}, \boldsymbol{P}^{(t)}, \pi^{(t)}) = \arg\min_\lambda \max_{\boldsymbol{P},\pi} U_\pi^{\boldsymbol{P},\lambda}(\boldsymbol{s}_1)$ subject to constraints in Equation ($\mathcal{P}_V$). The cumulative regret of the optimal Lagrangian objective is bounded with probability $1-\delta$:*

$$\mathit{Reg}_{\lambda^\star}(t) \leq O\left(\frac{1}{\varepsilon}|S||A|^{\frac{1}{2}} N H \sqrt{T\log T}\right) . \tag{14}$$

Proof. The main challenge of an unknown penalty term $\lambda^\star$ is that the optimality of the chosen transition $\boldsymbol{P}^{(t)}$ and policy $\pi^{(t)}$ does not hold in Theorem 6.2 due to the misalignment of the penalty $\lambda^{(t)}$ used in solving the optimization in Equation ($\mathcal{P}_V$) and the penalty $\lambda^\star$ used in computing the regret.

The optimality of $\lambda^{(t)}$ (minimizing $U_\pi^{\boldsymbol{P},\lambda}$) and the optimality of $\boldsymbol{P}^{(t)}, \pi^{(t)}$ (maximizing $U_\pi^{\boldsymbol{P},\lambda}$) are given by:

$$\lambda^{(t)}, \boldsymbol{P}^{(t)}, \pi^{(t)} = \arg\min_\lambda \max_{\boldsymbol{P},\pi} U_\pi^{\boldsymbol{P},\lambda}$$

which give us, respectively:

$$U_{\pi^{(t)}}^{\boldsymbol{P}^{(t)},\lambda^{(t)}} \leq U_{\pi^{(t)}}^{\boldsymbol{P}^{(t)},\lambda^\star}, \quad U_{\pi^\star}^{\boldsymbol{P}^\star,\lambda^{(t)}} \leq U_{\pi^{(t)}}^{\boldsymbol{P}^{(t)},\lambda^{(t)}} \tag{38}$$

Similarly, the optimality of $\lambda^\star$ can be written as:

$$\lambda^\star = \arg\min_\lambda U_{\pi^\star}^{\boldsymbol{P}^\star,\lambda}$$

which gives us

$$U_{\pi^\star}^{\boldsymbol{P}^\star,\lambda^\star} \leq U_{\pi^\star}^{\boldsymbol{P}^\star,\lambda^{(t)}} \tag{39}$$

Combining Inequality 38 and Inequality 39, we can bound:

$$U_{\pi^\star}^{\boldsymbol{P}^\star,\lambda^\star} \leq U_{\pi^\star}^{\boldsymbol{P}^\star,\lambda^{(t)}} \leq U_{\pi^{(t)}}^{\boldsymbol{P}^{(t)},\lambda^{(t)}} \leq U_{\pi^{(t)}}^{\boldsymbol{P}^{(t)},\lambda^\star}$$

This implies that:

$$\mathrm{Reg}_{\lambda^\star}^{(t)} = U_{\pi^\star}^{\boldsymbol{P}^\star,\lambda^\star} - U_{\pi^{(t)}}^{\boldsymbol{P}^\star,\lambda^\star} \leq U_{\pi^{(t)}}^{\boldsymbol{P}^{(t)},\lambda^\star} - U_{\pi^{(t)}}^{\boldsymbol{P}^\star,\lambda^\star} \tag{40}$$

which is exactly the same result as shown in Equation 10. The rest of the proof follows the same argument of Theorem 6.4 and Theorem 6.5, which concludes the proof. □

D.6 Choice of Horizon and Ergodicity Constant ε

For a given Markov chain, we need H to be sufficiently large to ensure the probability of visiting any state after H steps is at least a positive constant $\varepsilon > 0$. The choice of H depends on the MDP; we elaborate below how to select H and ε.

We follow a similar analysis of Markov chain convergence from Chapter 10 in (Spielman 2007) by defining:

$$\omega_2 = \max_{\pi \in \Pi} \sigma_2(P_\pi)$$

where $\sigma_2(P)$ is the magnitude of the second largest eigenvalue of the random walk matrix P_π induced by the policy π. In practice, ω_2 can be upper bounded by 1 if the MDP satisfies some properties, e.g., laziness of the Markov chain induced from the MDP (Chapter 10.2 in (Spielman 2007)).

Let v be the corresponding stationary distribution of the random walk matrix P_π with the policy π that maximizes the second largest eigenvalue. We know that v is strictly positive by ergodicity. When $\sigma_2 < 1$, we can write $r = \min_i v_i > 0$ and choose $\varepsilon = \frac{1}{2} r > 0$.

Let w be an arbitrary initial distribution. By applying Theorem 10.4.1 from (Spielman 2007) (the directed graph version), for every $t > H = \log_{\omega_2}(\frac{1}{2} r^{3/2}) = \log_{\omega_2}(\sqrt{2}\varepsilon^{3/2})$, we have:

$$|v - P_\pi^t w|_1 \leq \sqrt{\frac{1}{\min_i v_i}} \omega_2^t \leq \frac{r}{2}$$

which implies that the minimum value of $P_\pi^t w$ and the minimum value of v, i.e., r, differ by at most $\frac{r}{2}$. This implies that the minimum value of $P_\pi^t w$ is at least $\frac{r}{2} = \varepsilon$ for any initial distribution w. This choice of ε and H satisfies our requirement mentioned in Appendix D.4.

E ARMMAN: Maternal and Child Health Data

In the maternal mobile health program operated by ARMMAN, each instance is composed of a set of beneficiaries who participated in the program for 10 or more weeks. The dataset contains the states of each beneficiary, actions taken to schedule a service call to the beneficiary or not, and the beneficiary's next states after receiving the calling actions. This dataset is used to construct a set of empirical estimates of the transition probabilities and build an interactive RMAB environment to interact with. Our online learning algorithm then interacts with the environment to learn the transition and optimize total engagement. The experiments were all done in simulation.

Specifically, this problem is modelled as a 2-state (Engaging and Non-Engaging) RMAB problem where we do not know each beneficiary's transition behavior — transition between Engaging and Non-Engaging state, determined by whether the beneficiary listens to an automated voice message (average length 1 minute) for more than 30 seconds. The goal of the online learning challenge is to simultaneously learn the missing transition and optimize the overall engagement of all beneficiaries under budget constraints. The ARMMAN data is also abstracted out and contains no personally identifiable information and demographic feature related to the beneficiaries.

In the following sections, we provide more detailed information about consent related to data collection, analyzing data, data usage and sharing.

E.1 Secondary Analysis and Data Usage

This study falls into the category of secondary analysis of the aforementioned dataset shared by ARMMAN. We randomly sampled from the previously collected engagement probabilities of different beneficiaries participating in the service call program to simulate online learning environment. This paper does not involve deployment of the proposed algorithm or any other baselines to the service call program. As noted earlier, the experiments are secondary analysis with approval from the ARMMAN ethics board.

E.2 Consent for Data Collection and Sharing

The consent for collecting data is obtained from each of the participants of the service call program. The data collection process is carefully explained to the participants to seek their consent before collecting the data. The data is anonymized before sharing with us to ensure anonymity. Data exchange and use was regulated through clearly defined exchange protocols including anonymization, read-access only to researchers, restricted use of the data for research purposes only, and approval by ARMMAN's ethics review committee.

E.3 Universal Accessibility of Health Information

To allay further concerns: this simulation study focuses on improving quality of service calls. Even in the intended future application, all participants will receive the same weekly health information by automated message regardless of whether they are scheduled to receive service calls or not. The service call program does not withhold any information from the participants nor conduct any experimentation on the health information. The health information is always available to all participants, and

participants can always request service calls via a free missed call service. In the intended future application our algorithm may only help schedule *additional* service calls to help beneficiaries who are likely to drop out of the program.

F Experiment Details

F.1 Whittle Index Implementation Speedups

We introduce a number of implementation-level improvements to speed up the computation of Whittle indices. To our knowledge these approaches are novel for Whittle index computation.

Early termination The key insight is that the Whittle index threshold policy will pull the arms with the K largest Whittle indices. As we compute Whittle indices for each of the N arms, after we have computed the first K Whittle indices, any future arm selected would have to have Whittle index at least as high as the K-th largest seen so far in order to be pulled. Let us notate the K-th largest value seen so far as `top-k`.

Whittle indices are computed using a binary search procedure (Qian et al. 2016), which at each iteration tracks the upper bound $\overline{\lambda}$ and lower bound $\underline{\lambda}$ of the index. Once the upper bound falls below that of the minimum value of the K largest indices so far $\overline{\lambda} < \texttt{top-k}$, then we can terminate the binary search procedure as we are guaranteed that we would not act on that arm anyways. We implement the tracking of the K largest indices so far with a priority queue.

Similarly, we implement early termination to solve the bilinear programs ($\mathcal{P}_V$) and ($\mathcal{P}_m$) as callbacks in the Gurobi solver, in which we check the value of the current objective bound.

Memoization We memoize every Whittle index result computed throughout execution to track the index resulting from each pair of probabilities P_i and current state s_i as we perform calculations for each arm i. We implement this memoizer as a dictionary where the key is a tuple (P_i, s_i) with P_i recorded to four decimal places.

To implement the bilinear programs ($\mathcal{P}_V$) and ($\mathcal{P}_m$), we similarly memoize using the lower confidence bound (LCB) and upper confidence bound (UCB) that comprise the space $\boldsymbol{B}_i^{(t)}$.

F.2 Synthetic Data

The synthetic datasets are created by generating transition probabilities $P^i_{s,a,s'}$ sampled uniformly at random from the interval $[0, 1]$ for each arm i, starting state s, action a, and next state s'. Specifically we select transition probabilities for the probability of transitioning to a good state $P^i_{s,a,s'=1}$, then set $P^i_{s,a,s'=0} = 1 - P^i_{s,a,s'=1}$.

To ensure the validity constraints that acting is always helpful and starting in the good state is always helpful, we apply the following: for all arms $i \in [N]$:

- *Acting is always helpful:* If this requirement is violated with $P^i_{s,a=1,1} < P^i_{s,a=0,1}$, then $P^i_{s,a=0,1} = P^i_{s,a=1,1} \times \eta$ where η is uniform noise sampled between $[0, 1]$.
- *Starting from good state is always helpful:* If this requirement is violated with $P^i_{s=1,a,1} < P^i_{s=0,a,1}$, then $P^i_{s=0,a,1} = P^i_{s=1,a,1} \times \eta$ where η is uniform noise sampled between $[0, 1]$.

The *thin margin* dataset is created by mirroring the procedure described above but then constraining the probability of transitioning to a good state $P^i_{s,a,s'=1}$ to the interval $[0.2, 0.4]$. Thus the probabilities of transitioning to the bad state $P^i_{s,a,s'=0}$ are all between $[0.6, 0.8]$.

F.3 Acting in Low-Budget Settings

The potential impact of effectively allocating one resource is greater in low-budget settings. As one example, the ARMMAN setting from our experiments helps distribute a small number of healthcare workers across a group of pregnant women for preventative health care. We study real data from ARMMAN to show that the performance gap between approaches is wider in low-budget settings.

Using one actual instance from ARMMAN, we consider distributing healthcare workers across mothers (arms). Using the true transition probabilities, we calculate the (sorted) Whittle indices of an optimal policy as: 0.42, 0.39, 0.28, 0.23, 0.19, 0.11, 0.07, 0.

In the table below, we first show the expected reward of the optimal action and a random action (baseline) as we increase budget in the ARMMAN problem. We then calculate the difference in reward between the optimal action and random action for each budget level, normalized per worker. It is clear that the potential impact over the baseline of effectively allocating one worker is greater in low budget settings.

F.4 Computation Infrastructure

All results are averaged over 30 random seeds. Experiments were executed on a cluster running CentOS with Intel(R) Xeon(R) CPU E5-2683 v4 @ 2.1 GHz with 8GB of RAM using Python 3.9.12. The bilinear program solved using Gurobi optimizer 9.5.1.

	Reward		Reward gap per worker
K	Optimal	Random	$(\text{Opt} - \text{Random})/K$
1	0.42	0.211	0.209
2	0.81	0.423	0.194
3	1.09	0.634	0.152
4	1.32	0.845	0.119
5	1.51	1.056	0.091
6	1.62	1.268	0.059
7	1.69	1.479	0.030
8	1.69	1.690	0.000

Table 4: Reward contribution from each worker