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 from a real-world maternal and childcare dataset.

Introduction

Restless multi-armed bandits (RMABs) (Whittle 1988) generalize multi-armed bandits by introducing states for each arm. RMABs are commonly used to model sequential scheduling problems with limited resources such as in 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 bandits (Chen, Wang, and Yuan 2013), the RMAB learner must repeatedly pull K out of N arms at each timestep. Unlike stochastic bandits, the reward distribution of each arm in an RMAB depends on that arm's state, which transitions based on a Markov decision process (MDP) depending 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 rewards 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" for 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 strategically query arms to learn the underlying transition probabilities while simultaneously achieving high reward. Accordingly, in this paper we focus on the challenge of online learning in RMABs with unknown transitions. We focus on the Whittle index policy due to its scalability and consider a fixed-length episodic RMAB setting.

Main contributions We present *UCWhittle*, an upper confidence bound (UCB) algorithm that uses the Whittle index policy to achieve the first sublinear frequentist regret guarantee for RMABs. Our algorithm maintains confidence bounds for every 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, greatly reducing the computational 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 the worst-case 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 per-episode time horizon. Our result extends the analysis of Bayesian regret in RMABs (Jung and Tewari 2019) to frequentist regret by removing the need to assume a prior distribution. Finally, we evaluate UCWhittle against other online RMAB approaches on real maternal and child healthcare data (Mate et al. 2022b) and two synthetic settings, showing that UCWhittle achieves lower frequentist regret empirically as well.

Background

Offline planning for RMABs When the transition dynamics are given, an RMAB is an optimization problem in a sequential setting. 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 asymptotic optimality (Weber and Weiss 1990) of the Whittle index policy with 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 and transitions known.

Online learning for RMABs When the transition dynamics are unknown, an RMAB becomes an online learning problem in which the learner must simultaneously learn the transition probabilities (exploration) and execute high-reward actions (exploitation), with the objective of minimizing regret with respect to a chosen benchmark. Dai et al. (2011) achieve a regret bound of $O(\log T)$ benchmarked against an optimal policy from a finite number of potential policies. Xiong, Li, and Singh (2022) use a Lagrangian relaxation and index-based algorithm, but require access to an 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 a UCB-based algorithm to achieve $O(\log T)$ frequentist regret.

Recent works introduce oracle-based policies for the non-combinatorial setting in which the learner pulls a single arm in each round, receiving bandit feedback and observing only the state of the pulled arm. Jung and Tewari (2019) use a Thompson sampling–based algorithm which achieves a Bayesian regret bound $O(\sqrt{T \log T})$ under a given prior distribution. Wang, Huang, and Lui (2020) use separate exploration and exploitation phases to achieve frequentist regret $O(T^{2/3})$. These works assume some policy oracle is given, thus benchmark regret with the policy given by the oracle with knowledge of the true transitions. 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 regret guarantees. 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, evaluating 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.

Frequentist versus Bayesian regret The regret definition that we consider is *frequentist* regret, measuring worst-case regret under unknown transition probabilities. The other regret notion is Bayesian regret: the expected regret over a prior distribution over possible transition functions. Bayesian regret, such as from Thompson sampling–based methods, relies on a prior and does not provide worst-case guarantees (Jung and Tewari 2019; Jung, Abeille, and Tewari 2019).

Restless Bandits and Whittle Index Policy

An instance of a restless multi-armed bandit 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 (a set of "arm pulls") 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 $\sum_{i \in [N]} a_{h,i} \leq K$.

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.

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 $\gamma \in (0, 1)$:

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

where $\boldsymbol{s} \sim \boldsymbol{P}$ indicates $s_{h,i} \sim P_i(\cdot \mid s_{h-1,i}, \pi_i(\boldsymbol{s}_{h-1}))$ and $\boldsymbol{a} \sim \pi$ indicates $a_i \sim \pi_i(\boldsymbol{s})$.

Lagrangian Relaxation

Equation 1 is intractable to 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 also considers 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 candidate penalty values λ and find the infimum $\lambda^\star = \arg\min_\lambda U_{\star}^{\boldsymbol{P},\lambda}$ afterward.

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 $\pi(\boldsymbol{s}) = [\pi_i(\boldsymbol{s}_i)]_{i\in[N]}$, where each arm policy $\pi_i : \mathcal{S} \to \mathcal{A}$ specifies the action for arm i at state s_i. The value function is then:

$$V_{\pi_i}^{P_i,\lambda}(s_{1,i}) \coloneqq \mathop{\mathbb{E}}_{(s_{1,i},a_{1,i},s_{2,i},a_{2,i},\ldots)\sim(P_i,\pi_i)} \sum\nolimits_{h\in\mathbb{N}} \gamma^{h-1} \left(R(s_{h,i}, a_{h,i}) - \lambda \left(\pi_i(s_{h,i}) - K \right) \right) . \tag{3}$$

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 as good as not pulling it:

Definition 1. *Given transition probabilities P_i and state s_i, the* Whittle index *W_i 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 the Q-function $Q^{m_i}(s_i, a_i)$ and value-function $V^{m_i}(s_i)$ are the solutions to the Bellman equation with penalty m_i for pulling action $a_i = 1$:

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

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 for Equation 3 is to pull that 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 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:*

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

which pulls all arms with Whittle indices larger than λ.

The Whittle index threshold policy maximizes the relaxed Lagrangian in Equation 2 under penalty λ, but may violate the budget constraints in Equation 1. In practice, we pull only the arms with the top K Whittle indices to respect the strict budget constraint.

Problem Statement: Online Learning in RMABs

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.

At the beginning of each episode $t \in [T]$, the learner starts the RMAB instance (timestep $h = 1$) from $\boldsymbol{s}_1 = \boldsymbol{s}_{\text{init}}$ and selects a new policy $\pi^{(t)}$. We consider the following setting:

- Each episode has an 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 $\boldsymbol{s} \in \mathcal{S}$.

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

Definition 3 (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 $\pi^\star = \pi_{W(\boldsymbol{P}^\star),\lambda}$:*

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

However, the relaxed Lagrangian in Equation 2 with a randomly chosen penalty λ may not be a good proxy to the

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

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 (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:*

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

The expected regret is approximated using the regret from the relaxed Lagrangian in Equation 2 as defined in Definition 3 and Definition 4.

UCWhittle: Optimistic Whittle Index Threshold Policy

A key challenge to UCB-based online learning in RMABs is that the estimated transitions impact estimates of future reward, so optimistic estimates of transition probabilities do not correspond to optimistic estimates of reward. We introduce a method, UCWhittle, to compute optimistic Whittle indices that account for highest future value.

Confidence Bounds of Transition Probabilities

To compute confidence bounds for every unknown transition probability in the RMAB instance, we maintain counts $N_i^{(t)}(s,a,s')$ for every state, action, and next state transition observed by episode t.

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

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

where $N_i^{(t)}(s,a) \coloneqq \sum_{s'\in\mathcal{S}} N_i^{(t)}(s,a,s')$. With these confidence bounds, the ball $\boldsymbol{B}$ of possible values for transition probabilities $\boldsymbol{P}$ is

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

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:

$$\max_{V,Q,P_i\in\boldsymbol{B}_i^{(t)}} V(s_i) \qquad \text{s.t. } V(s) = \max_{a\in\mathcal{A}} Q(s,a) \qquad (\mathcal{P}_V)$$
$$Q(s,a) = -\lambda a + R(s,a) + \gamma \sum\nolimits_{s'\in\mathcal{S}} P_i(s,a,s')V(s')$$

Algorithm 1: UCWhittle

Input: N arms, budget K, episode horizon H
Initialize counts $N_i^{(t)}(s,a,s') = 0$ 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. 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*

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

We use the optimistic transition $P_i^\dagger$ to compute the corresponding *optimistic Whittle index* $W_i^\dagger = W(P_i^\dagger, s_i)$. 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 $\boldsymbol{s}$.

UCWhittle Algorithm

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

Alternative Formulation for Whittle Index Upper Bound

Equation $(\mathcal{P}_V)$ provides optimistic transition probabilities but requires separately solving for optimistic Whittle indices afterwards. Computing a Whittle index involves binary search, solving value iteration at every step, so is quite computationally expensive. We thus formulate a heuristic which solves for the highest *Whittle index* directly (instead of highest *future value*) at the current state $s_{h,i}$:

$$\max_{m_i,V,Q,P_i,\in\boldsymbol{B}_i^{(t)}} m_i \qquad (\mathcal{P}_m)$$
$$\text{s.t. } V(s) = \max_{a\in\mathcal{A}} Q(s,a),\ \ Q(s,a{=}0) = Q(s,a{=}1)$$
$$Q(s,a) = -m_i a + R(s,a) + \gamma \sum\nolimits_{s'\in\mathcal{S}} P_i(s,a,s')V(s')$$

Solving Equation $(\mathcal{P}_m)$ directly yields 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 $(\mathcal{P}_V)$. The theoretical analysis does not hold for $(\mathcal{P}_m)$, but empirically, we

show that this heuristic achieves comparable performance with significantly lower computation.

Regret Analysis

We analyze the regret of our UCWhittle algorithm to provide the first frequentist regret analysis for RMABs. In this section, we use the Lagrangian objective as a proxy to the reward received from the proposed policy. Section first assumes an arbitrary penalty λ is given to define the regret (Definition 3). Section generalizes by defining the regret of the optimal Lagrangian objective based on the unknown optimal penalty $\lambda^\star$ (Definition 4). Section provides an update rule for updating the penalty $\lambda^{(t)}$ after each episode. Full proofs are given in Appendix .

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 1. *Given $\delta > 0$ and $t \geq 1$, 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, even 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 1 (Regret decomposition)**.** *Given the penalty λ and $\boldsymbol{P}^\star \in \boldsymbol{B}^{(t)}$ for all t, we have:*

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

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

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

Theorem 1 enables us to bound our regret by the difference between two future values under the same policy $\pi^{(t)}$.

Definition 5 (Bellman operator)**.** *Define the* Bellman operator *as:*

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

Using Theorem 1 and the Bellman operator, we can further decompose the regret as:

Theorem 2 (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:*

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

Theorem 2 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 3. *Assume the penalty term $\lambda^{(t)} = \lambda$ is given and the RMAB instance is ε-ergodicity after H timesteps. Then with probability $1-\delta$, the cumulative regret in T episodes is:*

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

Proof sketch. We focus on bounding the regret when the confidence bounds hold. By Theorem 1 and Theorem 2, we estimate the right-hand side of Equation 11 to bound the total regret by the L^1-difference in the transition probability:

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

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 1 to conclude the proof. □

When the penalty term λ is given, Theorem 3 bounds the frequentist regret with a constant term depending on the ergodicity ε of the underlying true MDPs.

Regret Bound with Unknown Optimal Penalty

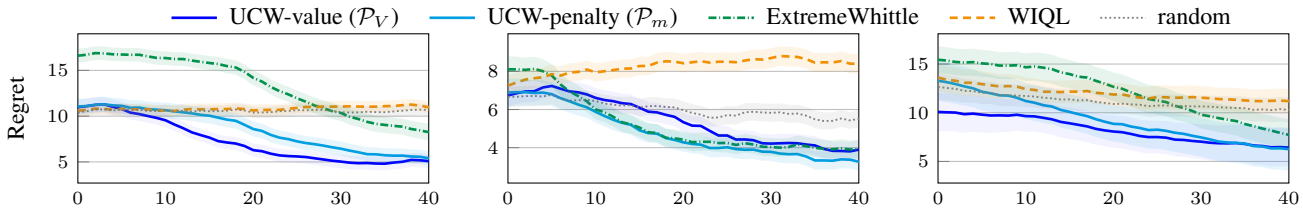
The analysis in Theorem 1 assumes a fixed and given penalty λ. Now, we generalize to regret defined in terms of the optimal but unknown penalty $\lambda^\star$ (Definition 4). We show that updating penalty $\lambda^{(t)}$ in Algorithm 1 achieves the same regret bound without requiring knowledge of the true transitions $\boldsymbol{P}^\star$ or optimal penalty $\lambda^\star$:

Theorem 4 (Regret bound with optimal penalty)**.** *Assume the penalty $\lambda^{(t)}$ in Algorithm 1 is updated by a saddle point $(\lambda^{(t)}, \boldsymbol{P}^{(t)}, \pi^{(t)}) = \arg\min_\lambda \max_{\boldsymbol{P},\pi} U^{\boldsymbol{P},\lambda}_\pi(\boldsymbol{s}_1)$ 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 1 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 $(\lambda^{(t)}, \boldsymbol{P}^{(t)}, \pi^{(t)}) = \arg\min_\lambda \max_{\boldsymbol{P},\pi} U^{\boldsymbol{P},\lambda}_\pi(\boldsymbol{s}_1)$ and $\lambda^\star = \inf_\lambda U^{\boldsymbol{P}^\star,\lambda}_{\pi^\star}(\boldsymbol{s}_1)$ 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 1 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 Theorem 1 without requiring knowledge of the optimal penalty $\lambda^\star$. The rest of the proof follows the same argument in Theorem 2 and Theorem 3 with the same regret bound. □

Penalty Update Rule

Theorem 4 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, the bilinear objective of $\mathcal{P}_V$ — where the transition probability and value function variables are being multiplied together — is difficult to solve in a minimax problem. 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 4.

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 in Appendix , and details about implementation (including novel techniques to speed up computation) and experiments are in Appendix .[2]

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).

ARMMAN is a non-profit based in India that disseminates health information to pregnant women and mothers to reduce maternal mortality. Twice a week, ARMMAN sends automated voice messages to enrolled mothers relaying critical preventative health information. To improve listenership, the organization provides service calls to a subset of mothers; the challenge is selecting which subset to call to maximize engagement. We use real, anonymized data of 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.

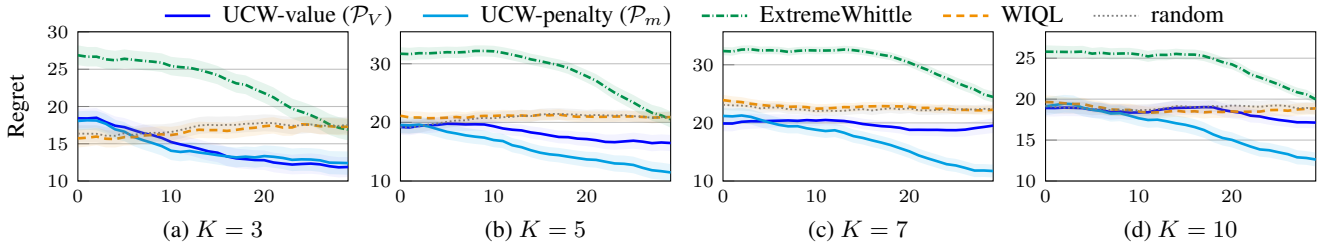
Wide Margin We randomly generate transition probabilities with high variance, while respecting the constraints specified above.

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 ($\mathcal{P}_m$).

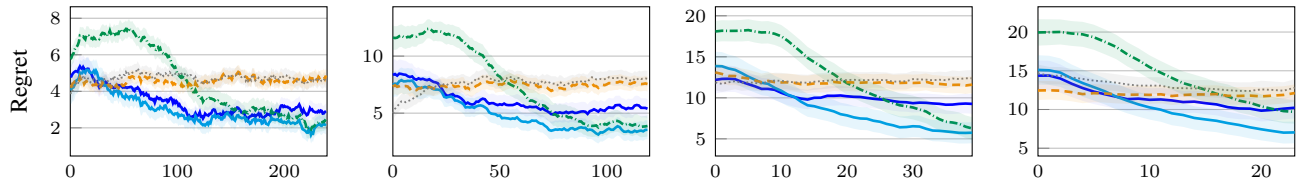
In this paper, we focus on frequentist regret, thus we exclude the Bayesian regret baselines, e.g., Thompson sampling (Jung and Tewari 2019), because their regret bounds are averaged over a prior. We consider the following three regret baselines: *ExtremeWhittle* is similar to the the approach by Wang et al. (2019): estimate Whittle indices from the extreme points of the unknown transition probabilities, 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 at 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.

[2]Code available at https://github.com/lily-x/online-rmab



(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.

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 3) 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 Equation ($\mathcal{P}_m$) — 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.

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 online learning using an optimistic Whittle index threshold policy. Theoretically, our work pushes the boundary of existing frequentist regret bounds in RMABs while enabling scalability using the Whittle index threshold policy to decompose the solution approach.

Acknowledgments

Kai Wang was supported by W911NF-17-1-0370 and ARO Grant Number W911NF-18-1-0208. Lily Xu was supported by ARO Grant Number W911NF-18-1-0208. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARO or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Kai Wang was also supported by Siebel Scholars. Lily Xu was also supported by a Google PhD fellowship. The authors thank Elias Khalil and anonymous reviewers for their thoughtful comments.

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.

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 (ICML)*, 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 *Proceedings of the Thirty-Sixth 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)*.