2015 IEEE 35th International Conference on Distributed Computing Systems

Weighted Restless Bandit and Its Applications

Peng-Jun Wan
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
Email: wan@cs.iit.edu

Xiaohua Xu
Department of Computer Science
Michigan Technological University
Houghton, MI 49931
Email: xiaohuax@mtu.edu

***Abstract*—Motivated by many applications such as cognitive radio spectrum scheduling, downlink fading channel scheduling, and unmanned aerial vehicle dynamic routing, we study two restless bandit problems. Given a bandit consisting of multiple restless arms, the state of each arm evolves as a Markov chain. Assume each arm is associated with a positive weight. At each step, we select a subset of arms to play such that the weighted sum of the selected arms cannot exceed a limit. The reward of playing each arm varies according to the arm's state. The exact state of each arm is only revealed when the arm is played. The problem weighted restless bandit aims to maximize the expected average reward over the infinite horizon. We also study an extended problem called multiply-constrained restless bandit where each time there are two simultaneous constraints on the selected arms. First, the weighted sum of the selected arms cannot exceed a limit; Second, the number of the selected arms is at most a constant K. The objective of multiply-constrained restless bandit is to maximize the long term average reward. Both problems are partially observable Markov decision processes and have been proved to be PSPACE-hard even in their special cases. We propose constant approximation algorithms for both problems. Our method involves solving a semi-infinite program, converting back to a low-complexity policy, and accounting for the average reward via a Lyapunov function analysis.**

I. Introduction

Multi-armed bandit (MAB) models the sequential resource allocation in many scientific disciplines such as sensor management, queuing and communication networks, clinical trials, control theory, and search theory. MAB is a paradigm of capturing the fundamental trade-off between exploration and exploitation. In the basic version of MAB, given a collection of arms, a player decides which arms to play, how many times to play each arm, and in which order to play them. The reward of playing each arm depends on the state of the arm. The objective is to maximize the reward earned through a sequence of plays.

In 1930s, Thompson *et al.* [23] initiated the first study on the MAB problem. Since then, different classes of MAB have been extensively studied such as [3], [4], [9], [25]. The first class is stochastic bandit [3]. In this class, the reward of each arm is a real distribution associated with a mean value, only a limited time is allocated to learn about this bandit to determine the arm(s) to play. Each arm of a stochastic bandit is a non-Bayesian and unknown process with fixed playing reward. Variants of stochastic bandit have been studied in [5], [6]. To maximize the reward, we need to strike a balance between probing an arm and exploiting an arm with known reward.

The second class of MAB is non-stochastic or adversarial bandit [4]. In this class, the reward of each arm is non-stochastic and each arm's state may quickly change after the arm is probed. For both stochastic bandit class and adversarial bandit class, the difficulty is informational and the performance metric is termed as *regret*. The regret can capture the difference between the optimal expected reward by playing consistently the best arm(s) and the player's actual reward.

The third class of MAB is Markov bandit where each arm is considered as a finite-state discrete time Markov chain. In a Markov bandit, the exploration and exploitation occur simultaneously by playing an arm. It admits an optimum solution without learning. This means that we can possibly compute the optimum reward without a genie. Thus, the difficulty for a Markov bandit problem is computational. In early versions of Markov bandit, passive arms (arms which are not played) are assumed to be rested. The rested case of MAB problem had been open for 40 years until Gittins [9] showed that the optimal policy has an index structure and a priority index can be assigned to each state of each arm and the optimal action at each step is to play an arm of the largest index. As a generalization of the rested case, the restless bandit assumes that passive arms also change states and multiple arms can be played simultaneously. Restless bandit is well-motivated in cognitive radio scheduling [14], downlink fading channel scheduling [19], unmanned aerial vehicle (UAV) dynamic routing [18], and real time multicast scheduling [21].

We will consider the weighted restless bandit problem as follows. Given a bandit of N independent restless arms, the state of each arm evolves according to a Markov rule. Each arm's state is only observed when the arm is played. Assume each arm is associated with a positive weight. At each step, the arms selected to play need to satisfy some resource constraint, *i.e.*, an upper bound on the weighted sum of selected arms. The reward of playing each arm varies according to the underlying state. The problem of weighted restless bandit aims to maximize the expected average reward over the infinite horizon.

We will also study an extended version called multiply-constrained restless bandit problem. Given a bandit of N independent restless arms, each time the selected arms to play need to satisfy two simultaneous constraints. First, the number of arms played each time cannot exceed a constant K; second, each time the weighted sum of selected arms cannot exceed some constant. To the best of our knowledge, all previous works on restless bandit only consider single constraint. In this work, we will generalize the study on restless bandit with

1063-6927/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDCS.2015.58

IEEE computer society

TABLE I. Comparison of Approximation Bounds

	Variants	Previous Methods	Our Policy
Cognitive Radio Opportunistic Spectrum Access	original	2	
	weighted	none	5
	multi-constraints	none	6
UAV Dynamic Routing	original	good empirical performance without approximation	
	weighted	none	5
	multi-constraints	none	6
Downlink Scheduling Fading Channel Scheduling	original	good empirical performance without approximation	
	weighted	none	5
	multi-constraints	none	6

two simultaneous constraints. Both the weighted restless bandit and multiply-constrained restless bandit have been proved to be PSPACE-hard even in their special cases [20]. The notoriously hard problems belongs to partially observable Markov decision processes [22] and constrained Markov decision processes [2] which are intractable generally.

For restless bandit, Whittle [25] defines a heuristic index that generalizes the Gittins index [9]. Whittle index has been extensively studied such as [14], [18] and has demonstrated excellent empirical performance. However, there are no theoretical guarantees in general. Recently, Guha *et al.* [13] developed a constant approximation policy for the restless bandit in its unweighted version. In this paper, we study the weighted restless bandit where we can play multiple arms simultaneously and each arm is associated with a weight.

Main Contributions: To the best of our knowledge, the weighted restless bandit has not been addressed, we are the first to address the weighted restless bandit problem and propose 5-approximation method for the problem. Our method includes solving the Lagrangian relaxation of the Linear Program (LP) which is a semi-infinite program and designing a feasible policy based on the LP solution. In our policy, based on the updated partial information of each arm, encapsulated by the last observed state and the number of steps since last play, we decide which arms to play next. We use a Lyapunov (potential) function analysis to account for the average reward.

Note that our method obviates the need for a heavy computation of Whittle index. Thus, the time complexity of our method is highly reduced. In addition, our method removes the necessity of establishing the indexability [25] of the bandit. Whittle index is feasible only in the premise of *indexability* [25] while the indexability property is notoriously hard to establish [14], [18]. In the general context of restless bandit, there are extensive works on indexability. See [10] for specific examples of the indexable restless bandit and [16], [17] for a numerical approach of testing indexability. For the weighted restless bandit considered here, our method prevents the proof of indexability.

For the problem multiply-constrained restless bandit where the number of arms played each time cannot exceed K. At the same time, the arms selected to play need to satisfy the weighted sum constraint, we propose a similar LP-based method that can achieve a 6-approximation guarantee.

Furthermore, our method sheds light on its observed superior performance in specific contexts. We show its applications in cognitive radio scheduling [14], downlink fading channel scheduling [19], and Unmanned Aerial Vehicle (UAV) discrete dynamic routing [18].

For cognitive radio opportunistic spectrum access [14], [26], there is a secondary user searching for spectrum temporarily unused by primary users in N channels. The state (idle/busy) of each channel models the occupancy of this channel by primary users and determines the gain of accessing this channel. Sensing a channel yields reward if the channel is idle. The objective is to select a subset of channels to sense and transmit on every time step, that maximizes the long-term channel utilization. We show how to map this problem to weighted restless bandit and multiply-constrained restless bandit and get constant approximation policies.

In the UAV dynamic routing scenario [18], there are multiple targets and vehicles. Each target has an information state which is not observed if the target is not visited by some vehicle. The rewards are obtained depending on the states observed, and the information state is updated. Once the rewards have been collected, the states of the targets evolve. The goal for the vehicles is to collect maximum rewards obtained when they visit the targets in a particular state. We can reformulate UAV dynamic routing as equivalent restless bandit problems and get constant approximation policies.

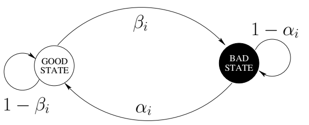
For downlink fading channel scheduling [19], in the presence of energy budget constraint, the objective is to schedule transmissions both to exploit those channels with up-to-date channel-state-information (CSI) and to explore the current state of those with outdated CSI to maximize the throughput. We convert downlink fading channel scheduling to weighted restless bandit and multiply-constrained restless bandit and achieve constant-approximations to maximize long term transmission rate.

To summarize, we present a side-by-side comparison for the above applications in Table I.

The rest of the paper is organized as follows. Section II presents the model of the problems. Section III and Section IV are devoted to the policy design for weighted restless bandit and the multiply-constrained restless bandit respectively. Section V demonstrates the applications of our results. Section VI presents the related work. Section VII concludes the paper.

II. Model and Preliminaries

There is a bandit consisting of N independent restless arms $\{1, 2, \cdots N\}$. Each arm i has two internal states, *i.e.*, good state and bad state. The good state of arm i yields reward r_i while the bad state yields no reward. We divide time into time slots. Let $s_i(T)$ be the arm i's internal state at time step T. The state transition of each arm i is illustrated in Fig. II, for instance, α_i represent the probability that the arm i transits to the good state, given the condition that its previous state is bad; then,



End Begin	Good	Bad
Good	$1-\beta_i$	β_i
Bad	α_i	$1-\alpha_i$

Fig. 1. The state transition of arm i can be represented by a 2×2 probability transition matrix.

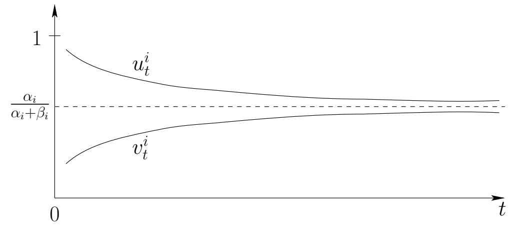


Fig. 2. The monotone property of u_t^i and v_t^i.

$1-\alpha_i$ is the probability that the arm i remains in bad state. We assume that for each arm i, $\alpha_i+\beta_i<1$ is satisfied, *i.e.*, all arms are *positive correlated* [14]. The reward of playing each arm depends on the internal state of the arm. The true internal state of each arm is only revealed when the arm is played.

We define a policy as selecting a subset of arms to play each time over the infinite horizon. For a policy, let $a_i(T)$ denote that whether arm i is played ($a_i(T)=1$) or not ($a_i(T)=0$) at step T. Given a constant W, assume each arm i is associated with a positive weight w_i, the problem **weighted restless bandit** aims to design a policy that maximizes the expected infinite horizon average reward, *i.e.*,

$$\lim_{\mathcal{T}\to+\infty} E\Big\{\frac{\sum_{T=0}^{\mathcal{T}}\sum_{i=0}^{N} a_i(T)\cdot s_i(T)\cdot r_i}{\mathcal{T}}\Big\}$$

subject to the constraint that $\sum_{i=1}^{N} w_i \cdot a_i(T) \le W : \forall T$.

In the problem **multiply-constrained restless bandit**, each time the total weight of selected arms to play cannot exceed W. In addition, the number of arms played each time cannot exceed K where K is a constant. The objective is to design a policy that maximizes the expected infinite horizon average reward, *i.e.*,

$$\lim_{\mathcal{T}\to+\infty} E\Big\{\frac{\sum_{T=0}^{\mathcal{T}}\sum_{i=0}^{N} a_i(T)\cdot s_i(T)\cdot r_i}{\mathcal{T}}\Big\}$$

subject to two simultaneous constraint that (1) $\sum_{i=1}^{N} w_i \cdot a_i(T) \le W$ and (2) $\sum_{i=1}^{N} a_i(T) \le K : \forall T$.

For any arm i, let u_t^i be the probability that the initial state is good and the state remains good after t steps; let v_t^i be the probability that the initial state is bad and the state is good after t steps. The values of u_t^i and v_t^i are given as follows and their monotone properties are illustrated in Fig. II.

$$u_t^i = \frac{\alpha_i}{\alpha_i+\beta_i} + \frac{\beta_i}{\alpha_i+\beta_i}\cdot(1-\alpha_i-\beta_i)^t$$
$$v_t^i = \frac{\alpha_i}{\alpha_i+\beta_i}\cdot\big(1-(1-\alpha_i-\beta_i)^t\big).$$

For each arm i whose last observed state is bad, given a constant t_i, if it is observed $t \ge t_i$ steps ago, the arm is called as t_i-*ready* or simply *ready*. If an arm i is not played when it is observed in bad state some $t \ge t_i$ steps ago, the arm is called as t_i-*blocked* or simply *blocked*.

For any policy, we define a performance measure of arm i as follows:

$$\begin{aligned} x_{gt}^i = & \Pr\{\text{arm } i \text{ is last played } t \text{ steps ago}\}\\ & \times \Pr\{\text{arm } i \text{ is last observed to be in good state}\}\\ & \times \Pr\{\text{arm } i \text{ is played currently}\}.\end{aligned}$$

$$\begin{aligned} x_{bt}^i = & \Pr\{\text{arm } i \text{ is last played } t \text{ steps ago}\}\\ & \times \Pr\{\text{arm } i \text{ is last observed to be in bad state}\}\\ & \times \Pr\{\text{arm } i \text{ is played currently}\}.\end{aligned}$$

Here, x_{gt}^i, x_{bt}^i are expected values which capture the fraction of time distribution of states when executing a policy. x_{gt}^i, x_{bt}^i will be used for LP formulations in Section III and IV.

III. Algorithm Design for weighted restless bandit

In this section, we design an algorithm for the weighted restless bandit problem. We partition arms into two subset, one subset consists of arms whose weights are at most $W/2$ and the other one consists of arms whose weights are greater than $W/2$. We design a policy for each subset and calculate the average reward output by each subset respectively. We select the subset with the larger reward.

A. Arms with weight $w_i \le W/2$

For the weighted restless bandit problem with each arm of weight $w_i \le W/2$, we formulate it as a relaxed LP.

$$\begin{aligned} \max \quad & \sum_{i=1}^{N}\sum_{t\ge1} r_i x_{gt}^i \\ s.t. \quad & \sum_{i=1}^{N}\sum_{t\ge1} w_i\cdot\big(x_{gt}^i + x_{bt}^i\big) \le W \qquad (1)\\ & \sum_{t\ge1} t\big(x_{gt}^i + x_{bt}^i\big) \le 1, \forall i \\ & \sum_{t\ge1} x_{bt}^i v_t^i = \sum_{t\ge1} x_{gt}^i(1-u_t^i), \forall i \\ & x_{gt}^i, x_{bt}^i \ge 0, \forall i, t\ge 1 \end{aligned}$$

Lemma 1: The optimal objective to the above LP, OPT, is at least the value of the optimal policy for weighted restless bandit.

Proof: In the execution of the optimal policy for the weighted restless bandit, the constraint in Equation (1) is satisfied since at each step, we play a subset of arms whose total weight is at most W.

Fix an arm i, any play with interval t maps to $t-1$ no-plays with interval t. The play rate with play interval t is $x_{gt}^i + x_{bt}^i$, the probability of no-play with interval t is $(t-1)\,\big(x_{gt}^i + x_{bt}^i\big)$. The total probability of play (play rate) is $\sum_{t\ge1}\big(x_{gt}^i + x_{bt}^i\big)$; the total probability of no-play is $\sum_{t\ge1}(t-1)\cdot\big(x_{gt}^i + x_{bt}^i\big)$.

The summation of two terms should be 1 if the arm has at least one play; 0 when no-play. So the second constraint is satisfied in the execution of the optimal policy.

We call a play on good state when last play is on bad state as an ascending play; we call a play on bad state when last play is on good state as a descending play. The probability of ascending play is equal to the probability of descending play. So the third constraint is satisfied in the execution of the optimal policy.

The value of the optimal policy is $\sum_{i=1}^{N}\sum_{t\geq 1} r_i x^i_{gt}$ which is at most OPT, the maximum possible objective value satisfying the LP constraints. ■

For an arbitrary $\lambda \geq 0$, the Lagrangian relaxation of the above LP is as follows.

$$\begin{aligned} &\max \sum_{i=1}^{N}\sum_{t\geq 1} r_i x^i_{gt} \\ &\quad + \lambda\Big(W - \sum_{i=1}^{N}\sum_{t\geq 1} w_i \cdot \big(x^i_{gt} + x^i_{bt}\big)\Big) \\ &= \max\ W\lambda + \sum_{i=1}^{N}\sum_{t\geq 1}\Big((r_i - w_i\lambda)x^i_{gt} - w_i\lambda x^i_{bt}\Big) \end{aligned}$$

In the Lagrangian relaxation of the LP, the arms are decoupled. Thus, we reduce an N-dimension problem to N independent 1-dimension problems and reduce the complexity of finding the optimal policy from exponential with N to linear with N. For each arm i, let $\lambda_i = w_i \cdot \lambda$, the corresponding LP is as follows.

$$\begin{aligned} \mathtt{LP}^i_{\lambda_i}: \quad &\max && \sum_{t\geq 1}\Big(r_i x^i_{gt} - \lambda_i(x^i_{gt} + x^i_{bt})\Big) \qquad (2) \\ &s.t. && \sum_{t\geq 1} t\big(x^i_{gt} + x^i_{bt}\big) \leq 1 \\ &&& \sum_{t\geq 1} x^i_{bt} v^i_t = \sum_{t\geq 1} x^i_{gt}(1 - u^i_t) \\ &&& x^i_{gt}, x^i_{bt} \geq 0, t \geq 1. \end{aligned}$$

For $\mathtt{LP}^i_{\lambda_i}$, Guha *et al.* [13] provides a method to solve it.

Lemma 2: [13] Given a Lagrangian multiplier λ_i, the optimal value of $\mathtt{LP}^i_{\lambda_i}$ is

$$H^i_{\lambda_i} = \frac{(r_i - \lambda_i)v^i_{t_i} - \lambda_i\beta_i}{v^i_{t_i} + t_i\beta_i}$$

where

$$t_i = \arg\max_t \frac{(r_i - \lambda_i)v^i_t - \lambda_i\beta_i}{v^i_t + t\beta_i}.$$

Since we have $\forall\lambda \geq 0, W\lambda + \sum_i H^i_{\lambda_i} \geq OPT$ and $H^i_{\lambda_i}$ decreases with λ, we can find a λ such that $\frac{W\lambda}{2} = \sum_i H^i_{\lambda_i} \geq OPT/3$.

The policy for weighted restless bandit is described in Table II. Note that at the initial step, we assume that all arms' last observed states are bad. This assumption has no impact on the average reward over the long term.

TABLE II. POLICY FOR WEIGHTED RESTLESS BANDIT

```
Remove any arm $i$ with $H^i_{\lambda_i} \leq 0$;
Each step, play a subset of arms so that
  the total weight is at most a limit $W$
  according to the following priority:
⋆ Play each arm which is last observed in good state;
    assume the total weight of played arms is $W'$.
⋆ Select ready arms of total weight up to $W - W'$
    in an arbitrary order.
```

Lemma 3: The policy in Table II outputs a 3-approximation for the weighted restless bandit problem with each arm of weight $w_i \leq W/2$.

Proof: At step T, if arm i move to bad state y steps ago, the potential $\phi^i_T = H^i_{\lambda_i}\big(\min(y, t_i) - 1\big)$. In good state, the potential $\phi^i_T = \frac{w_i\lambda + t_i H^i_{\lambda_i}}{v^i_t}$.

Let r^i_T denote the reward accrued from arm i until time T.

$$\Delta\phi^i_T = \phi^i_{T+1} - \phi^i_T$$
$$\Delta r^i_T = r^i_T - r^i_{T-1}$$

Consider an arm i at any step $T+1$, if the arm i is blocked, then there is no reward from this arm, and the potential of arm remains the same, *i.e.*, $\phi^i_T = \phi^i_{T+1} = H^i_{\lambda_i}(t_i - 1)$, we have

$$E[\Delta r^i_T + \Delta\phi^i_T] = 0.$$

If the arm i is observed in bad state some $t < t_i$ steps The potential increases by $H^i_{\lambda_i}$ this step, the reward collected by this arm is zero, we have

$$E[\Delta r^i_T + \Delta\phi^i_T] = H^i_{\lambda_i}.$$

If the arm i is played and the arm is last observed in bad state, assume the arm is last observed y steps ago, then $y \geq t_i$. Since v^i_t is monotonically increasing with t, with probability $q \geq v^i_{t_i}$, the observed state is good and the reward $\Delta r^i_T = r_i$ and the change in potential is $\frac{w_i\lambda + t_i H^i_{\lambda_i}}{v^i_t} - H^i_{\lambda_i}(t_i - 1)$. With probability $1 - q$ the observed state is bad and the change in potential is $H^i_{\lambda_i}(t_i - 1)$ and there is no change in reward. Thus, in this case, since $q \geq v^i_{t_i}$ and $\frac{w_i\lambda + t_i H^i_{\lambda_i}}{v^i_t} \geq 0$, we have:

$$\begin{aligned} E[\Delta r^i_T + \Delta\phi^i_T] &= q\Big(r_i + \frac{w_i\lambda + t_i H^i_{\lambda_i}}{v^i_t}\Big) - H^i_{\lambda_i}(t_i - 1) \\ &\geq w_i\lambda + H^i_{\lambda_i} \end{aligned}$$

If an arm i which was last observed in good state and played in the last step, with probability $1 - \beta_i$ the increase in reward is r_i and the potential is unchanged. With probability β_i, the potential will decrease by $\frac{w_i\lambda + t_i H^i_{\lambda_i}}{v^i_t}$. We have

$$E[\Delta r^i_T + \Delta\phi^i_T] = (1 - \beta_i)r_i - \beta_i\frac{w_i\lambda + t_i H^i_{\lambda_i}}{v^i_t} \geq w_i\lambda + H^i_{\lambda_i}$$

Now, consider the N arms together at time T. Let

$$\Phi_T = \sum_{i=1}^{N} \phi^i_T$$

Let $R_T = \sum_{i=1}^{N} r_T^i$ denote the total reward accrued until T.

$$\Delta\Phi_T = \Phi_{T+1} - \Phi_T = \sum_{i=1}^{N} \Delta\phi_T^i$$

$$\Delta R_T = R_T - R_{T-1} = \sum_{i=1}^{N} \Delta\phi_T^i$$

We will verify the following fact under different scenarios.

$$E[\Delta R_T + \Delta\Phi_T | \Phi_T] \geq OPT/3$$

If there exists blocked arm(s), since we select a subset of ready arms of total weight up to $W - W'$ in a greedy manner s.t. $t \geq t_i$, this means that if we add any blocked arm (assume its weight is w) to the existing subset of ready arms, the total weight will exceed $W - W'$, thus the weighted sum of selected ready arms at this step is at least $W - W' - w$. Then, the weighted sum of played arms at this step is at least $W' + (W - W' - w) = W - w \geq \frac{W}{2}$, suppose the played arms are $\{j_1, j_2, \cdots, j_L\}$. Note that the values of $\Delta r_T^i + \Delta\phi_T^i$ of other plays arms are non-negative.

$$\begin{aligned} E[\Delta r_T^i + \Delta\phi_T^i] &= \sum_{i=1}^{N} E[\Delta r_T^i + \Delta\phi_T^i] \\ &\geq \sum_{i=j_1}^{j_L} E[\Delta r_T^i + \Delta\phi_T^i] \geq \sum_{i=j_1}^{j_L} w_i\lambda + \sum_{i=j_1}^{j_L} H_{\lambda_i}^i \\ &\geq W/2 \cdot \lambda \geq OPT/3 \end{aligned}$$

If no arm is blocked, then each arm i is either played or last observed $t < t_i$ steps ago, in both cases, we have

$$E[\Delta r_T^i + \Delta\phi_T^i] \geq H_\lambda^i.$$

Thus, for all N arms, we have

$$E[\Delta R_T + \Delta\Phi_T | \Phi_T] \geq \sum_{i=1}^{N} H_\lambda^i \geq OPT/3$$

Thus, no matter whether any arm is blocked or not, at each step we have

$$E[\Delta R_T + \Delta\Phi_T | \Phi_T] \geq OPT/3.$$

Since we have

$$\Phi_T = \sum_{i=1}^{N} H_{\lambda_i}^i (\min(y, t_i) - 1 \leq \sum_{i=1}^{N} H_{\lambda_i}^i \left(t_i - 1\right),$$

the value of Φ_T is bounded, we have

$$\lim_{t\to\infty} \frac{R_T}{T} \geq OPT/3.$$

This finishes the proof. ■

TABLE III. POLICY FOR THE RB PROBLEM [13]

Each step play one arm according to the following priority: ⋆ Play an arm which is last observed in good state; ⋆ Otherwise select a ready arm to play.

B. Arms with weight $w_i > W/2$

In this section, we focus on the arms whose weights are greater than $W/2$. As the total weight cannot exceed W, the total number of arms played each time is exactly one. This is the restless bandit problem studied in [13] (noted as RB problem). For the RB problem, Guha *et al.* [13] proposed a policy with a 2-approximation (Lemma 4).

Lemma 4: [13] The policy in Table III outputs a 2-approximation for RB problem.

By Lemma 4, the policy in Table III can achieve 2-approximation for the weighted restless bandit problem with each arm of weight $w_i > W/2$.

C. Main Theorem

Recall that in our algorithm design, we divide all arms into two groups, one with small weight and the other with large weight. For each group, there is a constant-approximation factor policy. We calculate the average reward by using the corresponding policy for each group and focus only on the group with the larger reward. Let S be the larger one of those two rewards.

Theorem 5: S is a 5-approximation for the weighted restless bandit problem.

Proof: Assume $\mathcal{P}^*$ is the optimal solution for all arms. Let S^* be the average reward obtained by executing $\mathcal{P}^*$ on all arms. Let S_1 be the average reward obtained by executing $\mathcal{P}^*$ on arms whose weights are at most $W/2$. Let S_2 be the average reward obtained by executing $\mathcal{P}^*$ on arms whose weights are greater than $W/2$. Clearly, $S^* = S_1 + S_2$. Let S_1^* be the optimal average reward obtained for arms whose weights are at most $W/2$. Let S_2^* be the optimal average reward obtained for arms whose weights are greater than $W/2$. Then we have

$$S_1 \leq S_1^*, S_2 \leq S_2^*.$$

By Lemma 3 and 4, the output reward S of the proposed policy satisfies that

$$3S \geq S_1^*, 2S \geq S_2^*.$$

Thus, we have

$$S \geq \frac{S^*}{5}.$$

This finishes the proof. ■

IV. ALGORITHM DESIGN FOR MULTIPLY-CONSTRAINED RESTLESS BANDIT

A. Arms of weight $w_i \leq W/2$

In this section, we focus on the problem multiply-constrained restless bandit with each arm of weight $w_i \leq W/2$. The problem can be formulated as a LP relaxation as follows.

$$\begin{aligned}
\max \quad & \sum_{i=1}^{N}\sum_{t\geq 1} r_i x_{gt}^i \\
s.t. \quad & \sum_{i=1}^{N}\sum_{t\geq 1}\left(x_{gt}^i + x_{bt}^i\right) \leq K \qquad\qquad (3)\\
& \sum_{i=1}^{N}\sum_{t\geq 1} w_i \cdot \left(x_{gt}^i + x_{bt}^i\right) \leq W \\
& \sum_{t\geq 1} t\left(x_{gt}^i + x_{bt}^i\right) \leq 1, \forall i \\
& \sum_{t\geq 1} x_{bt}^i v_t^i = \sum_{t\geq 1} x_{gt}^i (1-u_t^i), \forall i \\
& x_{gt}^i, x_{bt}^i \geq 0, \forall i, t \geq 1 \\
& w_i \leq W/2, \forall i
\end{aligned}$$

Lemma 6: The optimal objective to the above LP, OPT, is at least the value of the optimal policy for multiply-constrained restless bandit.

Proof: The proof is similar to that of Lemma 1. We need to verify that all the constraints in the LP are satisfied during the execution of the optimal policy for the multiply-constrained restless bandit problem. The only difference is Equation (3). This inequality holds for the optimal policy since we can play at most K arms each step on average in addition to the weight constraint. So the inequality in Equation (3) is satisfied. This means that OPT, the maximum possible objective value satisfying the LP constraints, is at least the optimal value of the problem. ■

For an arbitrary $\lambda \geq 0$, let us consider the Lagrangian relaxation of the LP by moving the first two constraints into the objective as follows.

$$\begin{aligned}
&\max \sum_{i=1}^{N}\sum_{t\geq 1} r_i x_{gt}^i + \lambda\left(W - \sum_{i=1}^{N}\sum_{t\geq 1} w_i \cdot \left(x_{gt}^i + x_{bt}^i\right)\right) + \\
&\frac{W\lambda}{2K}\left(K - \sum_{i=1}^{N}\sum_{t\geq 1}\left(x_{gt}^i + x_{bt}^i\right)\right) \\
= \;& \frac{3W\lambda}{2} + \max \sum_{i=1}^{N}\sum_{t\geq 1}\left(r_i x_{gt}^i - \left(w_i\lambda + \frac{W\lambda}{2K}\right)\left(x_{gt}^i + x_{bt}^i\right)\right)
\end{aligned}$$

Then the arms can be decoupled. We consider the LP restricted to each arm. For each arm i, let $\lambda_i = w_i\lambda + \frac{W\lambda}{2K}$, we consider $\texttt{LP}_{\lambda_i}^i$ shown in Equation (2). The LP $\texttt{LP}_{\lambda_i}^i$ is restricted to arm i with a Lagrangian multiplier λ_i.

Given any $\lambda \geq 0$, we can obtain λ_i for each arm i and corresponding the values of $H_{\lambda_i}^i$ and t_i by Lemma 2. We have $\frac{3W\lambda}{2} + \sum_i H_\lambda^i \geq OPT$. At the same time, for each arm i, $H_{\lambda_i}^i$ decreases with λ_i, thus $H_{\lambda_i}^i$ decreases with λ. Considering all arms together, there exists a λ such that $\frac{W\lambda}{2} = \sum_i H_\lambda^i \geq OPT/4$. We find such a λ via a binary search.

We describe our policy in Table IV. Note that at the initial step, we also assume that all arms' last observed states are bad.

TABLE IV. POLICY FOR MULTIPLY-CONSTRAINED RESTLESS BANDIT

```
Remove any arm $i$ with $H_{\lambda_i}^i \leq 0$;
Each step play up to $K$ arms
  according to the following priority:
  ⋆ Play each arm which is last observed in good state;
     assume the number of played arms is $K'$
     and assume the total weight of played arms is $W'$.
  ⋆ Select up to $K - K'$ ready arms and prune these selected arms
     in an arbitrary order until the total weight of remaining selected
     arms is at most $W - W'$. Play all the remaining selected arms.
```

We will analyze the performance of the policy in Table IV.

Lemma 7: The policy in Table IV outputs a 4-approximation for the multiply-constrained restless bandit problem with each arm of weight $w_i \leq W/2$.

Proof: At step T, if arm i move to bad state y steps ago, we define the potential $\phi_T^i = H_{\lambda_i}^i\big(\min(y, t_i) - 1\big)$. In good state, the potential $\phi_T^i = \frac{\lambda_i + t_i H_{\lambda_i}^i}{v_t^i}$.

Let r_T^i denote the reward accrued from arm i until time T.

$$\begin{aligned}
\Delta\phi_T^i &= \phi_{T+1}^i - \phi_T^i \\
\Delta r_T^i &= r_T^i - r_{T-1}^i
\end{aligned}$$

Consider an arm i at any step $T+1$, if the arm i is blocked, then there is no reward from this arm, and the potential of arm remains the same, *i.e.*, $\phi_T^i = \phi_{T+1}^i = H_{\lambda_i}^i(t_i - 1)$, we have

$$E[\Delta r_T^i + \Delta\phi_T^i] = 0.$$

If the arm i is observed in bad state some $t < t_i$ steps ago, The potential increases by $H_{\lambda_i}^i$ this step, the reward collected at this step by this arm is 0, we have

$$E[\Delta r_T^i + \Delta\phi_T^i] = H_{\lambda_i}^i.$$

If the arm i is played and the arm is last observed in bad state, assume the arm i is last observed y steps ago, then we have $y \geq t_i$. With probability $q = v_y^i$, the observed state is good and the reward $\Delta r_T^i = r_i$ and the change in potential is $\frac{\lambda_i + t_i H_{\lambda_i}^i}{v_t^i} - H_{\lambda_i}^i(t_i - 1)$. Since v_t^i is monotonically increasing with t, we have $q \geq v_{t_i}^i$. With probability $1 - q$ the observed state is bad and the change in potential is $H_{\lambda_i}^i(t_i - 1)$ and there is no change in reward. Thus, in this case, since $q \geq v_{t_i}^i$ and $\frac{\lambda_i + t_i H_{\lambda_i}^i}{v_t^i} \geq 0$, we have

$$\begin{aligned}
E[\Delta r_T^i + \Delta\phi_T^i] &= q\left(r_i + \frac{\lambda_i + t_i H_{\lambda_i}^i}{v_t^i}\right) - H_{\lambda_i}^i(t_i - 1) \\
&\geq \lambda_i + H_{\lambda_i}^i.
\end{aligned}$$

If an arm i is last observed in good state, then the arm is played this step when executing the policy. With probability $1 - \beta_i$ the increase in reward is r_i and the potential is unchanged. With probability β_i, the potential will decrease by $\frac{\lambda_i + t_i H_{\lambda_i}^i}{v_t^i}$. We have

$$E[\Delta r_T^i + \Delta\phi_T^i] = (1 - \beta_i) r_i - \beta_i \frac{\lambda_i + t_i H_{\lambda_i}^i}{v_t^i} \geq \lambda_i + H_{\lambda_i}^i$$

Now, consider the N arms together at time T. Let $\Phi_T = \sum_{i=1}^N \phi_T^i$ denote the total potential. Let $R_T = \sum_{i=1}^N r_T^i$ denote the total reward accrued until T. Let $\Delta\Phi_T = \Phi_{T+1} - \Phi_T$ and $\Delta R_T = R_T - R_{T-1}$. Then, we have

$$\Delta\Phi_T = \sum_{i=1}^N \Delta\phi_T^i, \quad \Delta R_T = \sum_{i=1}^N \Delta\phi_T^i$$

We will verify the following fact under different scenarios.

$$E[\Delta R_T + \Delta\Phi_T | \Phi_T] \geq OPT/4$$

If no arm is blocked, then each arm i is either played or last observed $t < t_i$ steps ago. In both cases, we have

$$E[\Delta r_T^i + \Delta\phi_T^i] \geq H_{\lambda_i}^i.$$

Thus, for all N arms, we have

$$E[\Delta R_T + \Delta\Phi_T | \Phi_T] \geq \sum_{i=1}^N H_{\lambda_i}^i \geq OPT/4$$

If there exists blocked arm(s), either (1) there are K played arms at this step or (2) if we add any blocked arm to the existing subset of selected arms, the total weight will exceed W. In Case (1), suppose the K arms are $\{j_1, j_2, \cdots, j_K\}$. Note that the values of $\Delta r_T^i + \Delta\phi_T^i$ of non-blocking arms is non-negative.

$$\begin{aligned} & E[\Delta r_T^i + \Delta\phi_T^i] = \sum_{i=1}^N E[\Delta r_T^i + \Delta\phi_T^i] \\ & \geq \sum_{i=j_1}^{j_K} E[\Delta r_T^i + \Delta\phi_T^i] \geq \sum_{i=j_1}^{j_K} \lambda_i \\ & \geq \sum_{i=j_1}^{j_K} \frac{W\lambda}{2K} \geq \frac{W\lambda}{2} \geq OPT/4 \end{aligned}$$

In Case (2), if we add any blocked arm (assume its weight is $w \leq \frac{W}{2}$) to the existing subset of ready arms, the total weight will exceed W, thus the weighted sum of selected ready arms at this step is at least $W - w \geq \frac{W}{2}$. Then we have

$$\begin{aligned} & E[\Delta r_T^i + \Delta\phi_T^i] = \sum_{i=1}^N E[\Delta r_T^i + \Delta\phi_T^i] \\ & \geq W/2 \cdot \lambda \geq OPT/4 \end{aligned}$$

Thus, no matter whether any arm is blocked or not, at each step we have

$$E[\Delta R_T + \Delta\Phi_T | \Phi_T] \geq OPT/4.$$

Observe that

$$\Phi_T = \sum_{i=1}^N H_{\lambda_i}^i (\min(y, t_i) - 1 \leq \sum_{i=1}^N H_{\lambda_i}^i \, (t_i - 1) \, ,$$

the value of Φ_T is bounded, we have

$$\lim_{t\to\infty} \frac{R_T}{T} \geq OPT/4.$$

This finishes the proof. ∎

B. Arms with weight $w_i > W/2$

In this section, we focus on the multiply-constrained restless bandit problem with each arm of weight $w_i > W/2$. As the total weight cannot exceed W, the total number of arms played each time is exactly one. This is exactly the RB problem. By Lemma 4, the policy in Table III achieves a 2-approximation for the multiply-constrained restless bandit problem with each arm of weight $w_i > W/2$.

C. Main Theorem

Recall that we divide all arms into two groups in the beginning. One group consists of arms of weights at most $W/2$ and the other consists of arms of weights great than $W/2$. For both groups, there is a constant-approximation method. We can calculate the average reward for each group respectively and select the larger reward for these two groups. Let S be the large one of those two rewards.

Theorem 8: S is a 6-approximation for the multiply-constrained restless bandit problem.

Proof: The proof is similar to that of Theorem 5. Given an instance of multiply-constrained restless bandit problem, let $\mathcal{P}^*$ be the corresponding optimal policy. Let S^* be the optimal reward that can be obtained. Let S_1 be the optimal reward obtained by executing $\mathcal{P}^*$ on arms of weights at most $W/2$. Let S_2 be the average reward obtained by executing $\mathcal{P}^*$ on arms of weights greater than $W/2$. Clearly, $S^* = S_1 + S_2$. Let S_1^* be the optimal average reward obtained for arms of weights at most $W/2$. Let S_2^* be the optimal average reward obtained for arms of weights greater than $W/2$. Then we have

$$S_1 \leq S_1^*, S_2 \leq S_2^*.$$

By Lemma 7, the output reward S of the proposed policy satisfies $4S \geq S_1^*$. Furthermore, we have $2S \geq S_2^*$ Thus, we have

$$S \geq \frac{S^*}{6}.$$

This finishes the proof. ∎

V. Applications

In this section, we present applications of the weighted restless bandit and multiply-constrained restless bandit in Cognitive Radio (CR) Networks, Unmanned Aerial Vehicle (UAV) routing, and downlink scheduling setting respectively.

A. Opportunistic Spectrum Access in CR Networks

Opportunistic Spectrum Access (OSA) has received great research interests in cognitive radio networks [14], [26]. In the OSA problem, there is a secondary user searching for spectrum temporarily unused by primary users. Given multiple independent channels $\{1, 2, \cdots, N\}$, for each channel i at time-slot t, its state $S_i(T) \in \{0, 1\}$ models the occupancy probability of this channel by primary users and determines the reward of accessing this channel as well. If $S_i(T) = 1$, this means that there is a spectrum access opportunity for channel i at time T; otherwise, there is no opportunity. The state of each channel i varies according to a Gilbert-Elliot model [8], *i.e.*, there is a 2×2 probability transition matrix

$$\left(\begin{array}{cc} p_{11}^{i} & 1-p_{11}^{i} \\ 1-p_{01}^{i} & p_{01}^{i} \end{array} \right).$$

Here p_{11}^{i} represents the transition probability from idle state to idle state and p_{01}^{i} represents the transition probability from busy state to idle state for the channel i. Note that the Gilbert-Elliot model is commonly used to abstract physical channels with memory.

At each slot, the secondary user selects a subset of channels to sense. If the channel state is idle, the secondary user transmits in this channel and collects some reward. If the channel state is busy, the secondary user collects no reward. We assume there is a power budget associated with the secondary user and the cost to sense each channel varies. The objective is to sequentially select channels subject to the power budget in each slot that maximizes the expected average reward (utilization) in the long term.

By a careful examination of this problem, we show how to map it to the weighted restless bandit problem. Let us associate a secondary user with a bandit of multiple arms. Each channel corresponds to an arm. Playing the arm corresponds to sensing and transmitting on the channel, generating reward if the transmission is successful, and at the same time revealing to the secondary user the current state of the channel. The weight of each arm is the power cost to sense the corresponding channel. The power budget associated with the secondary user is the limit on the total weight of selected arms to play each time. Let $U(T)$ denote the set of arms (correspondingly channels) chosen in slot T. Let $S_i(T)$ be the state of arm i (correspondingly the i-th channel) at time-slot T. The reward obtained in slot T is $\sum_{i \in U(T)} S_i(T)B_i$ where B_i is a parameter belonging to arm i and its value equals to the bandwidth of channel i. The objective is to design a policy to maximize $\sum_{i \in U(T)} S_i(T)B_i$.

After formulating the OSA problem as a weighted restless bandit, we apply the policy in Section III to solve it. Thus, we can achieve a 5-approximation for the OSA problem. At the same time, the complexity of the policy is very low as we do not need to compute the Whittle index in closed form existing in previous methods such as [14]. Thus, we have Theorem 9.

Theorem 9: There is a polynomial time 5-approximation for power-budgeted cognitive radio opportunistic spectrum access problem.

Now let us consider an extension of the OSA problem. Each time the number of channels sensed cannot exceed some constant K in addition to the power budget constraint. We call this new problem as multiply-constrained cognitive radio opportunistic spectrum access problem. Similarly, this problem can be mapped to the multiply-constrained restless bandit, by applying the policy for multiply-constrained restless bandit in Section IV, we get a 6-approximation.

Theorem 10: There is a polynomial time 6-approximation for the multiply-constrained cognitive radio opportunistic spectrum access problem.

B. UAV Dynamic Routing Problem

We present a UAV routing problem and show how it can be viewed as a weighted restless bandit problem. We also present briefly variants of the problem which lead to different variants of the restless bandit problem.

For the weighted UAV routing problem, the energy cost of visiting each target (site) varies. Each target's state is not observed if the target is not visited by some vehicle. The goal is to visit a collection of targets sequentially to collect maximum expected average reward. This weighted UAV routing problem can be mapped to the weighted restless bandit exactly. Consider each target as an arm, the weight of each target is the energy cost to visit and observe the target. The energy budget for the vehicles is the weight limit W associated with the weighted restless bandit problem. By reformulating the problem, we can apply the policy for weighted restless bandit in Section III to obtain a 5-approximation.

Theorem 11: There is a polynomial time 5-approximation for the weighted UAV routing problem.

Next we consider a UAV routing problem called multiply-constrained UAV routing. In the problem, there are K vehicles and N targets ($K \leq N$). Given the current information state, we make the decision as to observe up to K targets where K is a small constant. Additionally, each time the total energy cost of visiting the selected targets cannot exceed a budget. The rewards are obtained depending on the states observed, and the information state is updated. Once the rewards have been collected, the states of the sites evolve. The goal is to determine a subset of targets to observe sequentially to collect expected average rewards obtained in the long run.

The problem described above can be formulated as an equivalent multiply-constrained restless bandit problem. Associate the UAV routing instance with a bandit process. Each target can be considered as an arm. The state of each arm varies according to a two-state Markov chain. The goal is to maximize the long term average reward. By applying the policy for multiply-constrained restless bandit, we get a 6-approximation for the UAV routing problem.

Theorem 12: There is a 6-approximation for the multiply-constrained UAV routing problem.

We call the above described model the basic model of UAV routing. Changes in the number and types of vehicles or the target dynamics result in problems that can be transformed into one of the variants of the restless bandit problem. For instance, when there is a setup cost or delay for switching the vehicle from one site to another and everything else is the same as in the basic model, the resulting problem can be formulated as a restless bandit problem with *switching penalties*; By considering new available vehicles brought to the system, we can obtain an arm-acquiring bandit variant of the restless bandit problem.

C. Downlink Scheduling with Power Budget

For downlink fading channel scheduling [19], we consider the CSI acquisition via a practical Automatic Repeat reQuest (ARQ)-based feedback mechanism. The ON/OFF state of each channel is revealed at the end of only scheduled users' transmissions.

In the downlink scheduling problem, there are multiple channels with temporally-correlated channel evolution. The desired scheduler must optimally balance the exploitation-exploration trade-off, whereby it schedules transmissions both to exploit those channels with up-to-date CSI and to explore the current state of those with outdated CSI. The problem is similar to the opportunistic spectrum access problem described in Section V-A. We assume there is a power cost to transmit over each channel. Each time the total power used for transmissions cannot exceed a limit. The goal is to select a subset of channels to transmit on every time step, that maximizes the long-term transmission rate. Similarly, we can map the downlink fading channel scheduling problem to a Weighted restless bandit problem. Each channel can be considered as an arm. The state of the arm is the state of the channel. Playing the arm corresponds to transmitting on the channel, yielding reward if the channel state is ON, and at the same time revealing to the transmitter the current state of the channel. The power budget is the limit on the weighted sum of selected arms at each time, *i.e.*, W. For the problem, by applying the policy for the weighted restless bandit, we get a 5-approximation for the weighted downlink fading channel scheduling problem.

Theorem 13: For the weighted downlink fading channel scheduling, there is a 5-approximation policy for this problem.

Next, we consider an extended version of downlink fading channel scheduling problem called multiply-constrained downlink fading channel scheduling. Given N independent channels, each time the total power used for transmissions cannot exceed a limit, at the same time, the number of transmissions on each step cannot exceed a constant K. The goal is to select a subset of channels to transmit on every time step without violating both constraints, that maximizes the long-term transmission rate. By applying the policy for multiply-constrained restless bandit, we get a 6-approximation for the problem.

Theorem 14: For the multiply-constrained downlink fading channel scheduling problem where there is a power cost to transmit over each channel and each time the power used for transmissions cannot exceed a limit, we can obtain a 6-approximation for this problem.

VI. Related Work

The structure of the optimal policy for a restless bandit had been pursued for decades, while an index policy was shown by Gittins in 1960s to be optimal for the classical bandit problems [9]. In 1988, Whittle proposed a Gittins-like heuristic index policy [25] for restless bandit. In Whittle's pioneering work [25], Whittle defined the index as the minimum subsidy that makes playing arms or not equally attractive. Whittle index policy is asymptotically optimal in terms of the number of arms in certain limiting regime as shown by Weber and Weiss in 1990 [24]. However, little is known about the structure of the optimal policies for a restless bandit. Even the indexability of a restless bandit is often difficult to establish [16]. In [18], Le Ny *et al.* have considered restless bandit motivated by the applications of target tracking. They have established the indexability and obtained the closed-form expressions for Whittle index under the discounted reward criterion. At the same time, Liu and Zhao [14] considered both discounted and average reward criteria for restless bandit, developed heuristic algorithms and computed the upper bound on the optimal performance, and established the semiuniversal structure, the optimality, and the performance of Whittle index policy for stochastically identical arms.

There are various approximation algorithms, see for example [16] [17] for near-optimal heuristics, as well as conditions for certain policies to be optimal for special cases of the restless bandit problem [1], [7], [15]. Constant factor approximation algorithms for restless bandits have been explored in the literature [11]–[13]. Guha *et al.* [13] developed a constant factor approximation for restless bandit where only one arm is allowed to play each time.

VII. Conclusion

We have proposed constant approximations to the weighted restless bandit and multiply-constrained restless bandit. We believe the techniques can be used for a large number of variants of restless bandit problems. A major open problem is whether we can still achieve a constant approximation for the weighted restless bandit problem with distinct demands. Specifically, given a set of arms, the total weight of arms activated is at most W each time. In addition, each arm has a demand associated with it, the objective is to satisfy as many demands as possible over the infinite horizon.

VIII. Acknowledgement

This work was supported in part by the National Science Foundation of USA under grants CNS-1219109, CNS-1454770, and ECCS-1523965.

References

[1] Ahmad, S., Liu, M., Javidi, T., Zhao, Q., and Krishnamachari, B. Optimality of myopic sensing in multichannel opportunistic access. *Information Theory, IEEE Transactions on 55*, 9 (2009), 4040–4050.

[2] Altman, E. *Constrained Markov decision processes*, vol. 7. CRC Press, 1999.

[3] Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning 47*, 2-3 (2002), 235–256.

[4] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing 32*, 1 (2002), 48–77.

[5] Bubeck, S., and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721* (2012).

[6] Cesa-Bianchi, N., and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

[7] Ehsan, N., and Liu, M. Server allocation with delayed state observation: Sufficient conditions for the optimality of an index policy. *Wireless Communications, IEEE Transactions on 8*, 4 (2009), 1693–1705.

[8] Gilbert, E. N., et al. Capacity of a burst-noise channel. *Bell Syst. Tech. J 39*, 9 (1960), 1253–1265.

[9] Gittins, J., Glazebrook, K., and Weber, R. *Multi-armed Bandit Allocation Indices*. Wiley. com, 2011.

[10] Glazebrook, K., Ruiz-Hernandez, D., and Kirkbride, C. Some indexable families of restless bandit problems. *Advances in Applied Probability* (2006), 643–672.

[11] Guha, S., and Munagala, K. Approximation algorithms for budgeted learning problems. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing* (2007), ACM, pp. 104–113.

[12] Guha, S., and Munagala, K. Model-driven optimization using adaptive probes. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (2007), Society for Industrial and Applied Mathematics, pp. 308–317.

[13] Guha, S., Munagala, K., and Shi, P. Approximation algorithms for restless bandit problems. *Journal of the ACM (JACM) 58*, 1 (2010), 3.

[14] Liu, K., and Zhao, Q. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *Information Theory, IEEE Transactions on 56*, 11 (2010), 5547–5567.

[15] Lott, C., and Teneketzis, D. On the optimality of an index rule in multichannel allocation for single-hop mobile networks with multiple service classes. *Probability in the Engineering and Informational Sciences 14*, 03 (2000), 259–297.

[16] Nino-Mora, J. Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability 33*, 1 (2001), 76–98.

[17] Niño-Mora, J. Dynamic priority allocation via restless bandit marginal productivity indices. *Top 15*, 2 (2007), 161–198.

[18] Ny, J. L., Dahleh, M., and Feron, E. Multi-uav dynamic routing with partial observations using restless bandit allocation indices. In *American Control Conference, 2008* (2008), IEEE, pp. 4220–4225.

[19] Ouyang, W., Eryilmaz, A., and Shroff, N. B. Asymptotically optimal downlink scheduling over markovian fading channels. In *INFOCOM, 2012 Proceedings IEEE* (2012), IEEE, pp. 1224–1232.

[20] Papadimitriou, C. H., and Tsitsiklis, J. N. The complexity of optimal queuing network control. *Mathematics of Operations Research 24*, 2 (1999), 293–305.

[21] Raghunathan, V., Borkar, V., Cao, M., and Kumar, P. Index policies for real-time multicast scheduling for wireless broadcast systems. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE* (2008), IEEE, pp. 1570–1578.

[22] Sondik, E. J. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations Research 26*, 2 (1978), 282–304.

[23] Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika 25*, 3/4 (1933), 285–294.

[24] Weber, R. R., and Weiss, G. On an index policy for restless bandits. *Journal of Applied Probability* (1990), 637–648.

[25] Whittle, P. Restless bandits: Activity allocation in a changing world. *Journal of applied probability* (1988), 287–298.

[26] Zhao, Q., and Sadler, B. M. A survey of dynamic spectrum access. *Signal Processing Magazine, IEEE 24*, 3 (2007), 79–89.