Cascading Effect on Restless Markov Chains

Aaryaman Aggarwal , Kavin Chowdary G
IIT Kanpur

(IIT Kanpur SURGE Project)

Project Mentor: Prof. Subrahmanya Swamy Peruru
Electrical Engineering Dept. IIT Kanpur

(Dated: July 10, 2025)

We have studied the online restless bandit problem and have tried to implement the cascading effect to it. In our problem statement, the state of each arm of the Multi-Armed Bandit problem evolves according to a Markov Chain. The reward is based on the arm we pull and the current state of the corresponding Markov Chain. Additionally, unlike the traditional case, we are not restricted to pulling a single arm in each time-step. It is based on the Cascading Model, where up to $K \leq N$ arms can be pulled, where N is the total number of arms present in our case. The application of this concept is in various domains, for example cognitive radios. The importance of our problem statement is that it is more realistic than the case where we assumed the reward from an arm is identically and independently distributed because there is a chance of correlation between the past state and current state. For example: In cognitive radios, if a channel is sensed at some timestep t, and it is occupied, there is a higher probability of it being occupied in the timestep t + 1. In our research, we are attempting to propose an algorithm that efficiently combines the Restless-UCB Algorithm with modifications to adapt to the Cascading nature of our problem, using Whittle's index as the Oracle to define the policy we use.

I. INTRODUCTION

The Multi-armed bandits are defined as a problem with multiple actions, the *Arms* that the agent can choose to interact with the environment.The environment gives feedback via the rewards provided. The rewards from each arm are not affected by each other, i.e the arms are independent of each other. This is a case of reinforcement learning which exemplifies the exploration-exploitation dilemma. The arms have multiple states in which they can be, and the reward for each state could be different. The arms can have a different set of states to be in as well.

At the start of the game, the agent *explores* the environment to determine the specific details of the arm, like the expected reward from the particular arm. In the *exploitation* phase, the agent has information about which arm is expected to provide the best reward according to which the agent can make a better decision and choose to exploit the better set of arms.

In the traditional setting of the game, the rewards for each arm are I.I.Ds; therefore, the reward in each time step is independent of the reward in the previous time steps. This is an ideal case, but in the real life application we have, specifically in cognitive radios, it is much more likely for the current state of each arm, to be dependent on the previous state of the arm. For example: If at the current time step, any arm is observed, and it is busy, it is more likely than not, that the arm will continue to be busy 20 seconds later.

In our setting, there are N arms(or actions) and the state of each arm i according to a Markov Chain M_i. The restless factor is that the state of each arm evolves in each time-step, regardless of it being observed in the particular time-step. At each time slot t the player chooses one arm to pull. Say they pull arm i, and the current state of arm i, is $s_i(t)$ of M_i for which it receives a random reward $x_i(t)$ which is dependent on i and $s_i(t)$. The expected reward over the time horizone T is $\mathbb{E}[\sum_{t=1}^{T} x_{a(t)}(t)]$. This is what we aim to maximize. In each time-step the agent can choose to play up to $K \leq N$ until a success is observed. This is the cascading effect which is the novel attribute we have attempted to integrate in our problem statement.

Most existing works in restless bandits focus on the offline setting where the parameters of the game are known to the player. Our setting is the online restless bandit setting, where the parameters of the game have to be learned (estimated). Obviously the policies in the offline settings do not perform close to the online setting as a part of the time is spent learning the parameters of the game. Second, for the class of Thompson Sampling based algorithms, theoretical guarantees are often established in the Bayesian setting, where the update methods can be computationally expensive when the likelihood functions are complex, especially for prior distributions with continuous support. Third, the existing policy with theoretical guarantee of a sublinear regret upper bound, i.e., colored-UCRL2, suffers from an exponential computation complexity and a regret bound that is exponential in the number of arms.

II. MODEL SETTING

Consider an online restless bandit problem $\mathcal{R}$ which has one player (decision make) and N arms $\{1, 2, ..., N\}$. Each arm $i \ \epsilon \ \{0, 1, 2, ..., N\}$ is associated with a Markov Chain M_i. Markov chains $\{M_i, i = 1, 2, ..., N\}$ have the same state space $S = \{1, 2, ..., M\}$ but may have different transition matrices $\{\mathbb{P}\beth, i = 1, 2, ..., N\}$,that are unknown to the player, and state dependent rewards $\{r(i, s), \forall i, s\}$ which are known to the player. In our setting, the state space is simply $S = \{0, 1\}$. The rewards for these states are as follows:

$$\mathbf{r(i, s_i = 0) = 0}$$

$$\mathbf{r(i, s_i = 1) = 1}$$

This is essentially a modeling of a simple cognitive radio instance. If the channel we sense is *not* free $(s_i = 0)$ for which the reward is 0. If the channel is $free$ then $(s_i = 0)$ for which the reward is 1. The states of only the observed arms are known, the state of the other arms are unknown as they evolve. This makes it a **Partially Observable Markov Decision Process**.

We use *Regret* to evaluate the efficiency of our learning policy, which is the expected gap between the algorithm and the offline optimal: i.e the Optimal Policy, ***The Whittle's Index***, with all the parameters, ***the probability transition matrices***, are known. Let $\mu(\pi, \mathcal{R})$ denotes the expected average reward under policy π for problem instance $\mathcal{R}$. Essentially $\mu(\pi, \mathcal{R})$ $= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[x^{\pi}(t)]$. Where $x^{\pi}(t)$ is the random reward at time t when applying policy π during the game. We define the optimal average reward as $\mu^*(\mathcal{R}) = sup_{\pi} \mu(\pi, \mathcal{R})$. The regret of policy π is defined as $Reg(T) = T\mu^*(\mathcal{R}) - \sum_{t=1}^{T} \mathbb{E}[x^{\pi}(T)]$ which is what we aim to minimize.

III. WHITTLE'S INDEX POLICY

To introduce indexability and Whittle's index, it suffices to consider a single arm due to the strong decomposability property of Whittle's index. Consider a single-armed bandit process (a single channel) with transition probabilities $\{p_{j,k}\}_{j,k\in\{0,1\}}$ and bandwidth B (we drop the channel index for notational simplicity). In each slot, the user chooses one of two actions, $u \in \{0 \text{ (passive)}, \ 1 \text{ (active)}\}$, to make the arm passive or active. An expected reward of ωB is obtained when the arm is activated at belief state ω. The objective is to decide whether to activate the arm in each slot to maximize the total discounted or average reward. The optimal policy is given by an optimal partition of the state space $[0, 1]$ into a passive set $\{\omega : u^*(\omega) = 0\}$ and an active set $\{\omega : u^*(\omega) = 1\}$, where $u^*(\omega)$ is the optimal action at belief ω.

Whittle's index quantifies how attractive it is to activate an arm based on the concept of a *subsidy for passivity*. Specifically, we construct a single-armed bandit process identical to the one described earlier, except that a constant subsidy m is granted whenever the arm is made passive. This subsidy naturally alters the optimal partition of the passive and active sets. States that remain in the active set under higher subsidy values are more attractive. The minimum subsidy m required to shift a state from active to passive thus reflects how desirable it is to activate that state.

We now present the formal definition of indexability and Whittle's index under the discounted reward criterion. Definitions under the average reward criterion can be obtained in a similar manner.

Let $V_{\beta,m}(\omega)$ denote the value function, representing the maximum expected total discounted reward obtained from a single-armed bandit process with subsidy m, starting from belief state ω. Considering the two possible actions in the first slot, we have:

$$V_{\beta,m}(\omega) = \max\left\{V_{\beta,m}(\omega; u = 0), \ V_{\beta,m}(\omega; u = 1)\right\},$$

where $V_{\beta,m}(\omega; u)$ is the expected total discounted reward when action $u \in \{0, 1\}$ is taken in the first slot, followed by an optimal policy thereafter.

For the passive action $u = 0$, the value function is:

$$V_{\beta,m}(\omega; u = 0) = m + \beta V_{\beta,m}(T(\omega)),$$

where m is the subsidy received and $T(\omega)$ is the updated belief state as defined earlier. For the active action $u = 1$, the value function is:

$$V_{\beta,m}(\omega; u = 1) = \omega + \beta\left[\omega V_{\beta,m}(p_{11}) + (1 - \omega) V_{\beta,m}(p_{01})\right].$$

The optimal action under subsidy m for belief state ω is defined as:

$$u^*_m(\omega) = \begin{cases} 1, & \text{if } V_{\beta,m}(\omega; u = 1) > V_{\beta,m}(\omega; u = 0), \\ 0, & \text{otherwise.} \end{cases}$$

The passive set under subsidy m is then given by:

$$\mathcal{P}(m) = \{\omega : u^*_m(\omega) = 0\} = \{\omega : V_{\beta,m}(\omega; u = 0) \geq V_{\beta,m}(\omega; u = 1)\}.$$

If an arm is indexable, its Whittle's index $W(\omega)$ at belief state ω is defined as the infimum subsidy m for which it becomes optimal to make the arm passive in state ω. Equivalently, it is the smallest subsidy value that makes the passive and active actions equally rewarding. Formally, Whittle's index is given by:

$$\begin{aligned} W(\omega) &= \inf_m \{m : u^*_m(\omega) = 0\} \\ &= \inf_m \{m : V_{\beta,m}(\omega; u = 0) = V_{\beta,m}(\omega; u = 1)\}. \end{aligned}$$

IV. ALGORITHM

Following is the algorithm we have proposed to tackle our problem setting.

Algorithm 1 Restless-UCB Policy

```
Input:   Time horizon $T$, learning function $m(T)$.
Until $v_i(s = k) = m(t)\ \forall i = 1, 2, ...N$ and $k = 0, 1$ where $i$ is the arm and $s$ is the state of the arm do:
    Choose up to $K$ arms until there a success found i.e $s_i = 1$ where $i \leq K$.
Let $\hat{P}_i(j, k)$'s be the empirical values of $P_i(j, k)$'s.
Construct instance $\mathcal{R}'$ with $P_i'(k, k+1) = \hat{P}_i(k, k+1) + rad(T)$, $P_i'(k, k) = \hat{P}_i(k, k)$, $P_i'(k, k-1) = \hat{P}_i(k, k-1) - rad(T)$. Specifically, $P_i'(1, 1) = \hat{P}_i(1, 1) + rad(T)$ and $P_i'(M, M) = \hat{P}_i(M, M) - rad(T)$.
Find the optimal policy $\pi^{*\prime}$ for problem $\mathcal{R}'$, i.e., $\pi^{*\prime} = \texttt{Oracle}(\mathcal{R}')$. Which is determined using $Whittle's\ Index$ Policy with the modification of choosing up to $k$ arms or $i$ less than equal to $k$ if a success (i.e $s_i = 1$) is found on the $i$th arm
while true do
    Follow $\pi^{*\prime}$ for the rest of the game.
end while
```

Here, v is a vector which records the number of times the state $(s = k)$ has been observed for arm i. K is the maximum number of arms we can sense in a single timestep according to our cascading model.

Exploration Phase :
The objective here is to accurately estimate the transition probabilities $P_i(j \mid k)$ and rewards $r(i, k)$ for each arm $i \in \{1, 2, \ldots, N\}$ and state k. Each arm is pulled until, for every state-action pair, we have at least $m(T)$ observations of transitions and corresponding rewards. With high probability, these empirical values are within a confidence radius

$$\mathrm{rad}(T) = \sqrt{\frac{\log T}{2m(T)}}$$

from the true parameters. The function $m(T)$ must be carefully selected to trade off between estimation accuracy and sample complexity.

Exploitation Phase :
After exploration, the algorithm constructs an *estimated offline instance* and invokes an offline oracle to compute the optimal policy. Instead of relying directly on empirical means, we build a confidence-adjusted model as follows:

- Set $\hat{r}(i, k) = r(i, k) + \mathrm{rad}(T)$
- Set $\hat{P}_i(k \mid k+1) = P_i(k \mid k+1) - \mathrm{rad}(T)$
- Set $\hat{P}_i(k \mid k) = P_i(k \mid k)$
- Set $\hat{P}_i(k \mid k-1) = P_i(k \mid k-1) + \mathrm{rad}(T)$

This optimistic construction enables the oracle to compute a policy that is robust yet performs well in practice. The total computational complexity remains $O(N)$, making the algorithm efficient for large-scale settings.

In this *Optimistic Problem Instance* we use Whittle's Index where the agent knows the probability estimates, and continues to play this according to that policy for the rest of the game.

V. RESULTS

As we can see, the UCB algorithm with cascading, utilizing **Whittle's index**, exhibits a significantly lower cumulative regret when compared to other algorithms such as *Thompson Sampling (TS)* and *UCRL2*. The cascading effect, in this context, refers to the structured decision-making process where choices made earlier in the sequence influence the outcomes and available decisions in subsequent rounds.

This is especially relevant in environments with limited resources or partially observable systems. Whittle's index efficiently prioritizes which arms to activate at each round by quantifying their attractiveness, while the cascading structure helps eliminate suboptimal decisions earlier. This synergy reduces exploration overhead and leads to faster convergence to optimal policies.

The following plots present the cumulative regret over time for different settings:

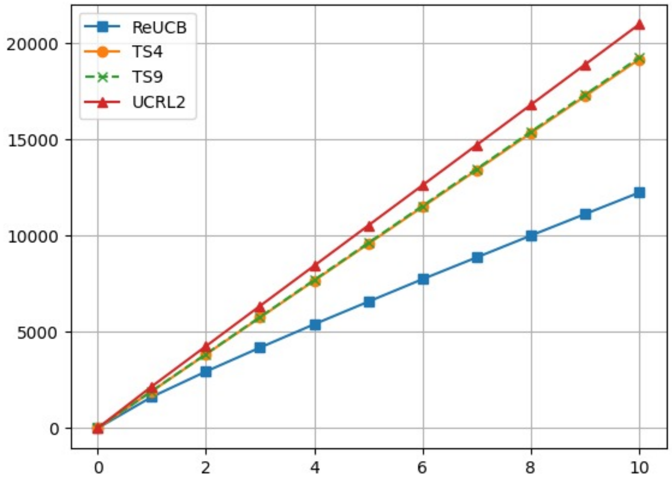

FIG. 1: Cumulative Regret – Suburban L-Band, Number of Arms = 3

From the Suburban L-Band setting, it is clear that the cascading-UCB method results in a much flatter regret curve, indicating more efficient arm selection and faster adaptation to the environment. In the Urban L-Band scenario, despite the increased number of arms and more complex dynamics, the cascading strategy continues to maintain a noticeable advantage.

These results underscore the strength of combining Whittle's index with cascading policies in bandit settings, offering practical improvements in both structured and unstructured environments.

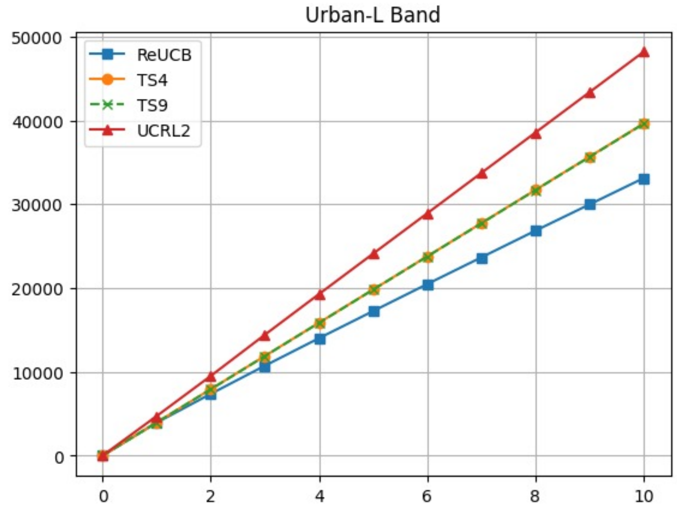

FIG. 2: Cumulative Regret – Urban L-Band, Number of Arms = 4