Cascading Effect in Restless Markov Bandits

Aaryaman Aggarwal, Kavin Chowdary G, Dr. S Swamy Peruru
Department of Electrical Engineering
aarymanag23@iitk.ac.in kavincg23@iitk.ac.in

Introduction

Our research problem is a version of the classic multi-armed bandits. In application to cognitive radios, it is more likely for the current state of the arm to be dependent on the previous state of the arm. In the classic problem, the state in each time-step is assumed to be independent and identically distributed. In our situation, we use the more realistic approach of the state of the arm following a Markov Chain. This simulates the real-life situation more accurately along with the cascading effect of being able to sense multiple arms at once until a success is found.

Objective

- Construct an algorithm which efficiently combines existing algorithms with the Cascading effect.
- Experiment with the learning functions to balance exploration and exploitation.
- Minimize regret, which is the gap between our algorithm and the best possible algorithm in which all the parameters that need to be learnt are known, The ideal policy. obtain a form of regret which grows sub-linearly with respect to T, the time horizon of each game

Whittle Index

The main idea behind the Whittle index policy is that we define what we call as a subsidy(m), which we use to tell us when playing an arm as passive gives us a better reward than playing it as an active arm. This takes into consideration future rewards as well, modelling the events to be dependent on each other rather than iid's.

To solve the Restless Multi-Armed Bandit problem, we use the Whittle Index as our Oracle, which assigns an index to each arm (channel) based on a belief state ω (probability of being in a "good" state).

Whittle Index Definition:

$W(\omega)=\inf\{m: V_{\beta,m}(u=0) \geq V_{\beta,m}(u=1)\}$

Value Functions:

- Passive action (don't select the arm): $V_{\beta,m}(u=0)=m+\beta V_{\beta,m}(T(\omega))$
- Active action (select the arm): $V_{\beta,m}(u=1)=\omega+\beta[\omega V_{\beta,m}(p11)+(1-\omega)V_{\beta,m}(p01)]$

Key Terms:

- ω: Belief state (Pr[channel is good])
- m: Subsidy for passivity
- β: Discount factor (future reward weight)
- T(ω): Updated belief if arm is not observed
- p11,p01: Markov transition probabilities:
  - p11p_{11}p11: good → good
  - p01p_{01}p01: bad → good

Algorithm

Using the whittle index oracle, we implement a UCB algorithm using what we've learnt in our online exploration phase to update the expected rewards and probability transitions of each arm in our offline phase as follows:

Empirical Estimates:

$P_i(j,k)$ and r(i,k) These are empirical estimates of:

- $P^i(j,k)$: Estimate transition probability of the ith arm from state j to k.
- r′(i,k): Estimated reward of arm i at state k.

Confidence Radius Function:

Rad(T): Function bounding error in our empirical estimate.

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

Constructed Reward:

r′(i,k)=r^(i,k)+rad(T)

Constructed Transition Probabilities:

These define the instance R′ with more optimistic (for one) or pessimistic (for the other) bounds.

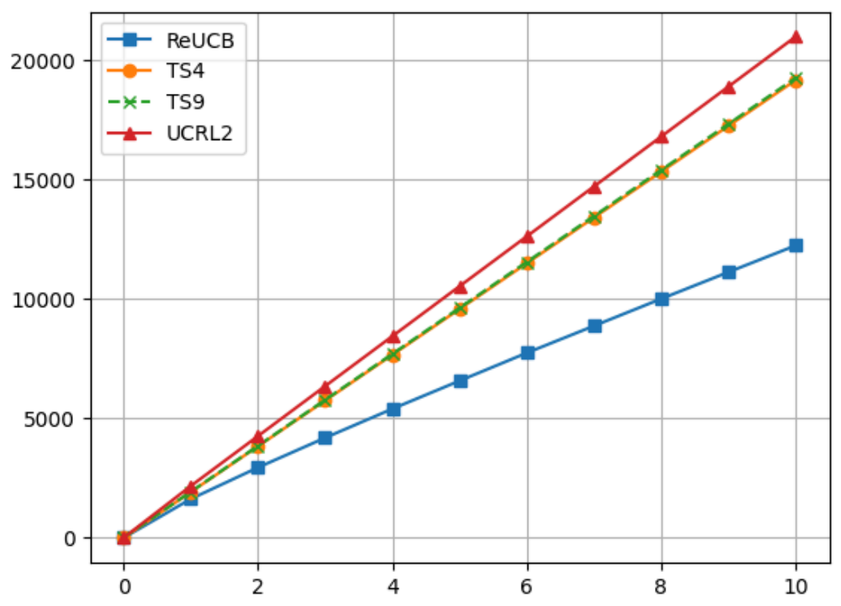
- Transition from state 0 to 1:
  Pi′(0,1)=P^i(0,1)+rad(T)
- Transition from state :
  Pi′(1,0)=P^i(1,0)-rad(T)
- Boundary examples:
  Pi′(1,1)=P^i(1,1)+rad(T)
  Pi′(0,0)=P^i(0,0)−rad(T)

Algorithm:

1. There is an exploration phase where each state of every arm is observed $m(T)$ times.
2. These observations are used to create the empirical estimates which are then used to create the constructed transition probabilities and rewards for offline instance *R'*.
3. Whittle's index policy is used to determine the optimal policy for the offline instance *R'*
4. The policy is then followed for the rest of the game which returns K arms with the highest expected reward based on the Whittle Index.

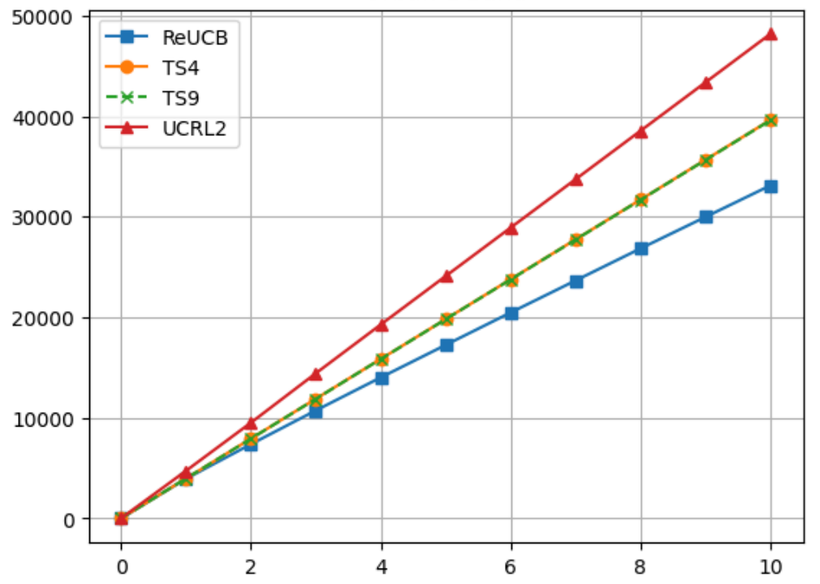
Observations

As we can see, the UCB algorithm with cascading, utilizing Whittle's index, gives a significantly lower regret in comparison to the other algorithms such as the Thomson Sampling algorithm and the UCRL algorithm.



Suburban L-Band, Number of arms = 3

Urban-L Band



Urban L-Band, Number of arms = 4

Results

The results below for cumulative regret among different algorithms shows that our proposed algorithm, ReUCB, outperforms the others

References

- Restless-UCB, an Efficient and Low-complexity Algorithm for Online Restless Bandits, Siwei Wang, Longbo Huang, John C.S Lui
- Indexability of Restless Bandit Problems and Optimality of Whittle's Index for Dynamic Multichannel Access, Keqin Liu, Qing Zhao
- Regret Bounds for Restless Markov Bandits, Ronald Ortner, Daniil Ryabko, Peter Auer, Remi Munos