Experimental Performance Comparison and Analysis for Various MAB Problems under Cognitive Radio Framework

Navikkumar Modi
Supervised by: Prof. Christophe Moy, Prof. Philippe Mary

navikkumar.modi@supelec.fr

November 5, 2014
Overview

1. Introduction
2. Multi-Armed Bandit Problem
3. Classic Multi-Armed Bandit Problem
   - UCB1 Policy
4. Markovian Multi-Armed Bandit Problem
5. Numerical Analysis
6. Experimental Setup for OSA
7. Take home message
Cognitive Radio is suggested as one of the solution to mitigate spectrum scarcity problem.

Opportunistic spectrum access is the dynamic spectrum access mechanism where secondary users opportunistically access the underutilized spectrum.

The goal of secondary user is to find and subsequently transmit in vacant spectrum with minimal interference to Primary User.

Reinforcement Learning can be used to predict next transmission opportunities.

We have shown that OSA scenario can be modeled as a multi-armed bandit problem

\(^1\)Wassim Jouini et al. “Upper confidence bound based decision making strategies and dynamic spectrum access”. In: International Conference on Communications, ICC’10. May 2010.
Introduction
Cognitive Radio

Sense 1 of $K$ Gilbert-Elliot channels

State of the art for spectrum allocation mainly considers:
- Probabilistic Resource Allocation algorithms
- Genetic Algorithms
Introduction

Cognitive Radio

Sense 1 of $K$ Gilbert-Elliot channels

- Opportunistic Spectrum Access: adapt to time-varying channel state.
- Channel State: free (1) or occupied (0).
- Limited Sensing: can sense and access $M$ channels (1 channel in our work) out of $K$ channels in each slot.

Which channel to sense and subsequently transmit in each slot?
Multi-Armed Bandit Problem

Introduction

- $K$ possible actions (one per machine = arm)
- Reward distribution in general differs from one arm to the another.
- The player must use all his past actions and observations to essentially learn the quality of these arms (in terms of their expected reward).
- You play for period of time to maximize reward in the long run (expected utility)
Multi-Armed Bandit Problem

Introduction

Which is the best action/arm/machine?
What sequence of actions to take to find out optimal machine and to maximize the expected reward?
Multi-Armed Bandit Problem

Exploration Vs Exploitation Dilemma

- Exploration: striving for information
- Exploitation: striving for reward

Suppose, at time $t$ you have arrived at reasonable estimates $\bar{r}(t)$ of the true values $r(t)$

Dilemma:

- You can’t exploit all the time; you can’t explore all the time
- You can never stop exploring; but you could reduce exploring
Multi-Armed Bandit Problem

Application as a Cognitive Radio$^2, 3$

- Choose the best channel to transmit at the next time step based on history.
- User or player = Secondary user
- Slot machines (arms) = Frequency bands
- Reward = channel’s state (e.g., free or occupied)
- Action = Senses a channel

---


i.i.d. Reward Model

Performance Measure: Regret

- \( r^i(t) \): reward achieved by policy \( A \) at time \( t \) from arm \( i \)
- \( r^i(t) \) is assumed to be Bernoulli distributed \( r^i(t) \in \{0, 1\} \)
- \( \mu^i \): expected reward of machine \( i \)
- \( \mu^* \): expected reward of optimal machine

Regret is the expected reward loss after \( n \) sensing due to the fact that the policy does not always sense the optimal channel.

\[
R^A(n) = n\mu^* - \sum_{t=1}^{n} \mathbb{E}[r^i(t)]
\]

Finding a policy which has minimum growth rate of regret \( R^A(t) \)
i.i.d. Reward Model

The UCB1 Algorithm

UCB1 policy is presented in\(^4\).

- Each arm is a frequency band

\[
B_{n,T_i(n)}^i = \frac{1}{n} \sum_{s=1}^{T_i(n)} r^i(T_i(n)) + \sqrt{\frac{\alpha \ln(n)}{T_i(n)}
\]

Where, \(T_i(n)\) is number of times an arm \(i\) has been sensed up to time \(n\). Select an arm with highest \(B_{n,T_i(n)}^i\).

- Sum of an exploration and exploitation term.

- Intuition: Select an arm that has a high probability of being the best, given what has been observed so far

- The \(B_{n,T_i(n)}^i\) index is upper confidence bound on \(\mu^i\)

Markov MAB problem is more suitable for modeling OSA formulation. The state ( Occupied or Free) of a channel is assumed to be evolved as a Markov chain. Reward is a function of the observed state of a channel or Markov chain. Possible to assume observed reward as a channel condition due to non-binary Markovian reward assumption. $M$ channels (1 in our work) out of $K$ channels are sensed.
After a channel $i$ is sensed in state $i \in \{0, 1\}$, the probability that the channel is in state 1 after $t$ slots is given by the $t$-step transition probability $p_{01}^i(t)$ of the Markov chain.

Reward $r_{1}^i(t)$ is observed reward in state 1 of channel $i$ at time $t$.

Independent channels (arms) with fully observable states $S^i(t)$.

**Two Formulation:** Rested or Restless
Markovian Reward

Rested Markov Multi-Armed Bandit

- Only sensed channel changes state and offers reward.
- Passive arms remain frozen.
- State in which we next observe an arm is independent of the time elapsed between consecutive actions of that arm.
- UCB1 policy was extended for rested Markov Multi-Armed Bandit Problem\(^5\).

---

Markovian Reward

Restless Markov Multi-Armed Bandit

- Passive arms may change state and offer reward\(^6\).
- State in which we next observe an arm is dependent on the time elapsed between two consecutive actions.
- Optimal Policy is no longer staying with one arm.
- Require to learn optimal way to switch among channels based on past observations (infinite possibilities).
- Optimal policy structure is unknown.
- PSPACE-hard.

---

OSA modeled as multi-armed bandit process.
Consider two different scenarios
- Assume iid reward process with Bernoulli distributed reward.
- Assume Markovian reward process

**Goal:** Select an arm more often which has the highest expected mean reward.
Goal: evaluate $K$ possible channels for transmission.

Which one is most effective?

- $K$ Resource to allocate
- In the later stage of allocation, greater fraction of time should be assigned to a channels, which have found to be vacant more during the earlier stage.
- Bernoulli distributed bounded reward $r_i(t) = \{0, 1\}$.
- Reward $r_i(t) = 0$ if the channel found to be occupied.
- Reward $r_i(t) = 1$ if the channel found to be free.
- Expected mean reward $\mu_i$ of each channel is shown in below table.

<table>
<thead>
<tr>
<th>channel</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>0.12</td>
<td>0.14</td>
<td>0.18</td>
<td>0.22</td>
<td>0.26</td>
<td>0.40</td>
<td>0.55</td>
<td>0.60</td>
<td>0.70</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Result: i.i.d. Rewards

Regret Analysis (Reward Loss)

Regret is the expected reward loss after $n$ sensing due to the fact that the policy does not always sense the optimal channel.

$$R^A(n) = n\mu^* - \mu^i \sum_{i=1}^{K} \mathbb{E}[r^i(t)]$$
Result: i.i.d. Rewards

Result: Successful transmission and Optimal channel Percentage

- **Optimal channel selection percentage:** Number of times given policy played an optimal channel from total number of time steps.

- **Successful transmission percentage (STP):** Number of times vacant slot is detected from total time steps.
System Model: Markovian Rewards

- Reward $r^i_0(t)$ if the channel found to be occupied state $P_0$.
- Reward $r^i_1(t)$ if the channel found to be free state $P_1$.
- State transition probabilities $P^i$ and respective mean reward $\mu^i$ is given below:

<table>
<thead>
<tr>
<th>channel</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{01}$</td>
<td>0.20</td>
<td>0.30</td>
<td>0.40</td>
<td>0.50</td>
<td>0.55</td>
<td>0.60</td>
<td>0.65</td>
<td>0.70</td>
<td>0.75</td>
<td>0.80</td>
</tr>
<tr>
<td>$P_{10}$</td>
<td>0.70</td>
<td>0.65</td>
<td>0.55</td>
<td>0.50</td>
<td>0.45</td>
<td>0.40</td>
<td>0.37</td>
<td>0.35</td>
<td>0.30</td>
<td>0.25</td>
</tr>
<tr>
<td>$r^i_0$</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>$r^i_1$</td>
<td>0.20</td>
<td>0.25</td>
<td>0.30</td>
<td>0.35</td>
<td>0.40</td>
<td>0.45</td>
<td>0.55</td>
<td>0.60</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.12</td>
<td>0.14</td>
<td>0.18</td>
<td>0.22</td>
<td>0.26</td>
<td>0.31</td>
<td>0.38</td>
<td>0.43</td>
<td>0.52</td>
<td>0.63</td>
</tr>
</tbody>
</table>
Result: Markovian Rewards

Result: Regret and selection percentage

Regret Analysis
Result: Markovian Rewards
Result: Regret and selection percentage

Regret Analysis

Best Arm selection
Experimental Setup for OSA

- Left: Primary network transmission.
- Right: One secondary user learning (UCB1 algorithm).
- Energy detector as a sensor at receiver side.

Experimental Results for OSA

- **Left:** Empirical average of vacancy of 8 channels.
- **Right:** UCB1 indexes for each channel.
- **Middle:** UCB1 results.
**Take home message**

- **Bandit problem**: starting point for many application and context-specific tasks.
- Simple and efficient *upper confidence bounds* based policies for the bandit problem as an application on cognitive radio with known bounded support with uniform logarithmic regret.
- Compared to iid assumption Markovian assumption facilitates to consider channel condition.
- Lots of open areas for research
  - Extend single user to the Multiple user with better coordination.
  - What if the reward distribution is non-stationary for Markov multi-armed bandit?
  - Consider a channel quality and other criteria for the channel selection with the goal of energy efficiency.
Thank You!

For further information please refer

- SCEE research team web site:
  - http://www.rennes.supelec.fr/ren/rd/scee/

Acknowledgement

- This work has received a French state support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference Nb. ANR-10-LABX-07-01.
- The authors would also like to thank the Region Bretagne, France, for its support of this work.