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

Introduction

Cognitve Radio

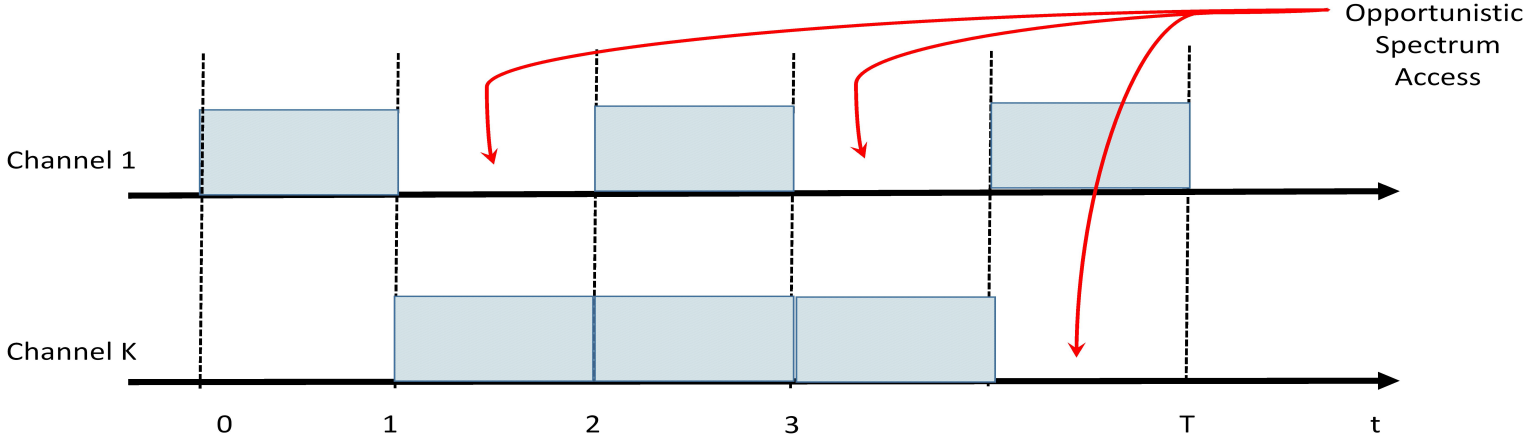
- Cognitve Radio is suggested as one of the solution to mitigate spectrum scarcity problem.
- Opportunistic spectrum access is the dynamic spectrum access mechanism where secondary users opportunistically access the underutilized spectrum.
- The goal of secondary user is to find and subsequently transmit 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].

[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

Cognitve Radio

Sense 1 of K Gilbert-Elliot channels



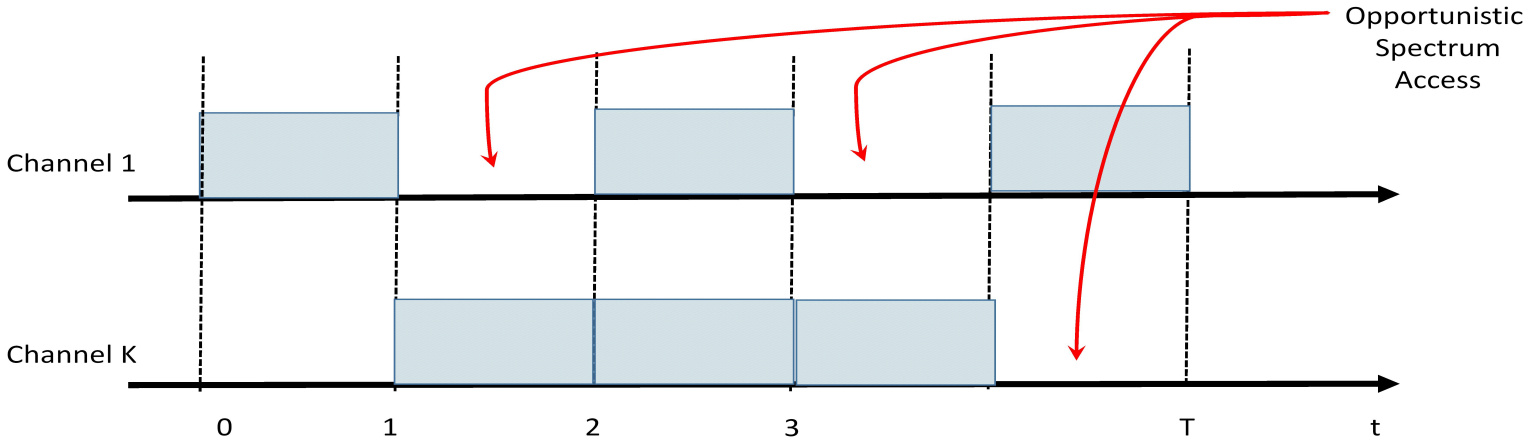
State of the art for spectrum allocation mainly considers:

- Probabilistic Resource Allocation algorithms
- Genetic Algorithms

Introduction

Cognitve 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

[2] Wassim Jouini et al. "Upper confidence bound based decision making strategies and dynamic spectrum access". In: *International Conference on Communications, ICC'10*. May 2010.

[3] Wassim Jouini, Christophe Moy, and Jacques Palicot. "Decision making for cognitive radio equipment: analysis of the first 10 years of exploration". In: *EURASIP Journal on Wireless Communications and Networking* 2012.26 (Jan. 2012).

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 $r^i(t) \in \{0, 1\}$
- μ^i: expected reward of machine i
- μ^*: 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

$$B^i_{n,T^i(n)} = \frac{1}{n}\sum_{s=1}^{T^i(n)} r^i(T^i(n)) + \sqrt{\frac{\alpha \ln(n)}{T^i(n)}}$$

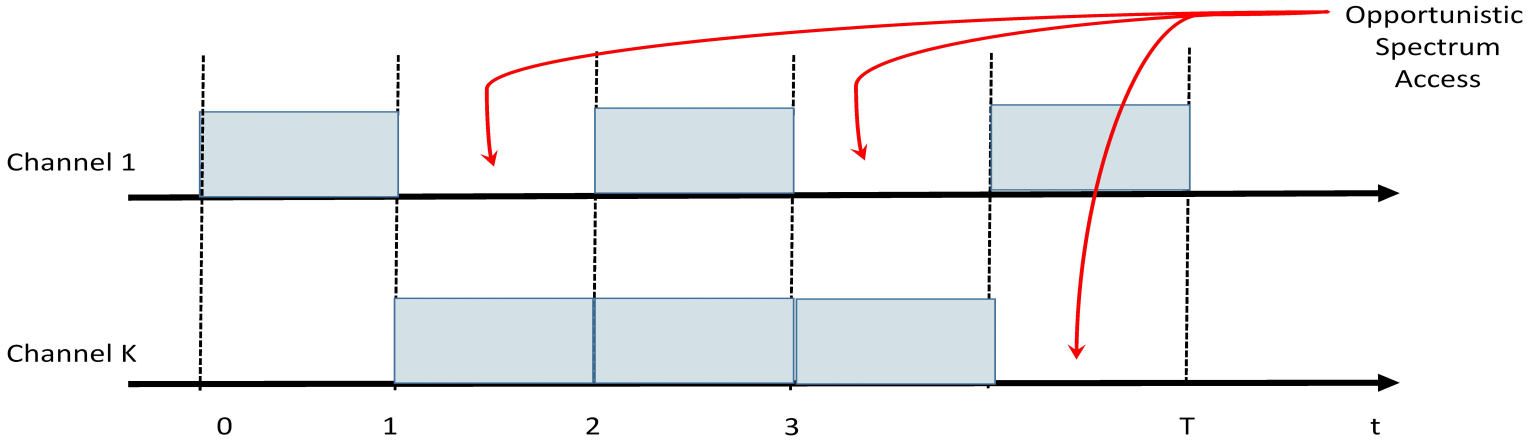
Where, $T^i(n)$ is number of times an arm i has been sensed up to time n.
Select an arm with highest $B^i_{n,T^i(n)}$.

- 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 $B^i_{n,T^i(n)}$ index is upper confidence bound on μ^i

[4] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. "Finite-time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* 47.2-3 (May 2002), pp. 235–256.

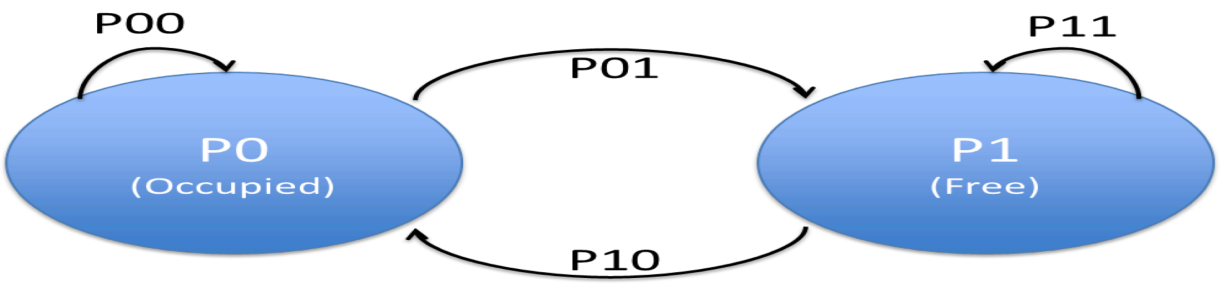
Markovian Multi-Armed Bandit Problem

Introduction



- 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.

Markovian Reward



State transition probability

- 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].

[5] Cem Tekin and Mingyan Liu. "Online algorithms for the multi-armed bandit problem with markovian rewards". In: *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE. 2010, pp. 1675–1682.

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.

[6] Haoyang Liu, Keqin Liu, and Qing Zhao. "Learning in a Changing World: Restless Multiarmed Bandit With Unknown Dynamics". In: *IEEE Transactions on Information Theory* 59.3 (2013), pp. 1902–1916.

Numerical Analysis

- 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.

Numerical Analysis

System Model: i.i.d. Rewards

- 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 μ^i of each channel is shown in below table.

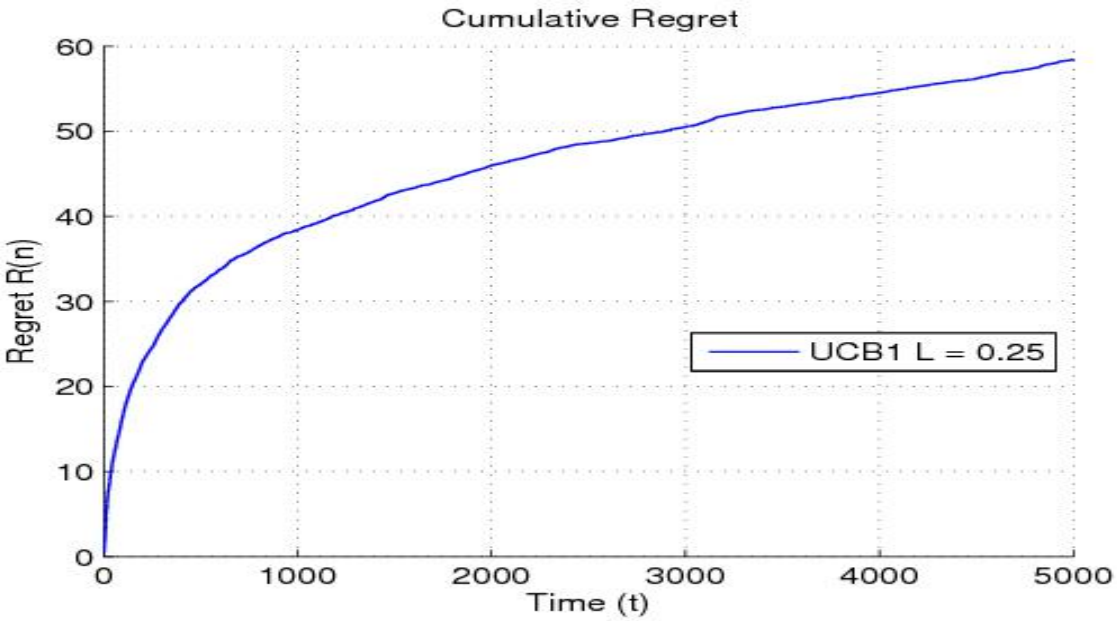
channel	1	2	3	4	5	6	7	8	9	10
μ	0.12	0.14	0.18	0.22	0.26	0.40	0.55	0.60	0.70	0.85

Result: i.i.d. Rewards

Regret Analysis (Reward Loss)

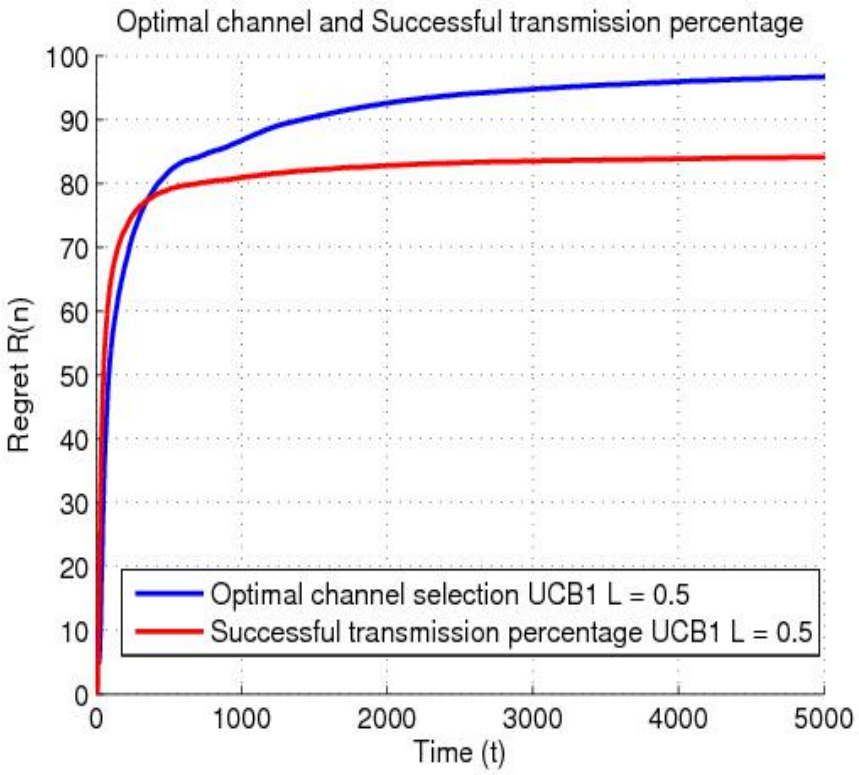
- Regret is the expected reward loss after n sensing due to the fact that the policy does not always sense the optimal channel.

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



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.

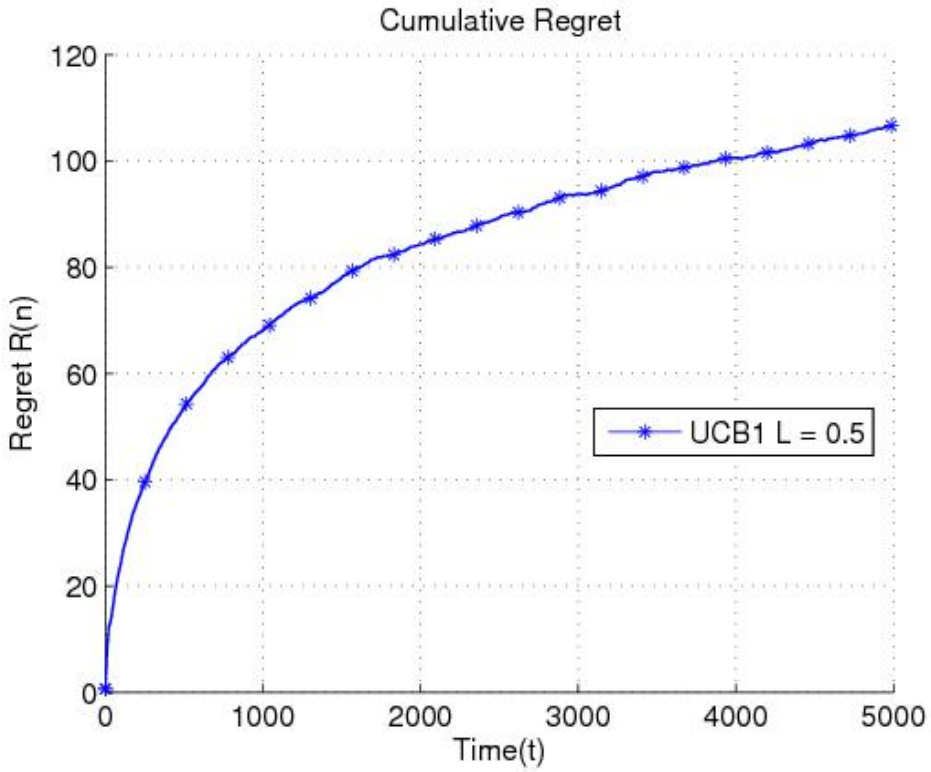
System Model: Markovian Rewards

- Reward $r_0^i(t)$ if the channel found to be occupied state P_0.
- Reward $r_1^i(t)$ if the channel found to be free state P_1.
- State transition probabilities P^i and respective mean reward μ^i is given below:

channel	1	2	3	4	5	6	7	8	9	10
P_{01}	0.20	0.30	0.40	0.50	0.55	0.60	0.65	0.70	0.75	0.80
P_{10}	0.70	0.65	0.55	0.50	0.45	0.40	0.37	0.35	0.30	0.25
r_0^i	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
r_1^i	0.20	0.25	0.30	0.35	0.40	0.45	0.55	0.60	0.70	0.80
μ	0.12	0.14	0.18	0.22	0.26	0.31	0.38	0.43	0.52	0.63

Result: Markovian Rewards

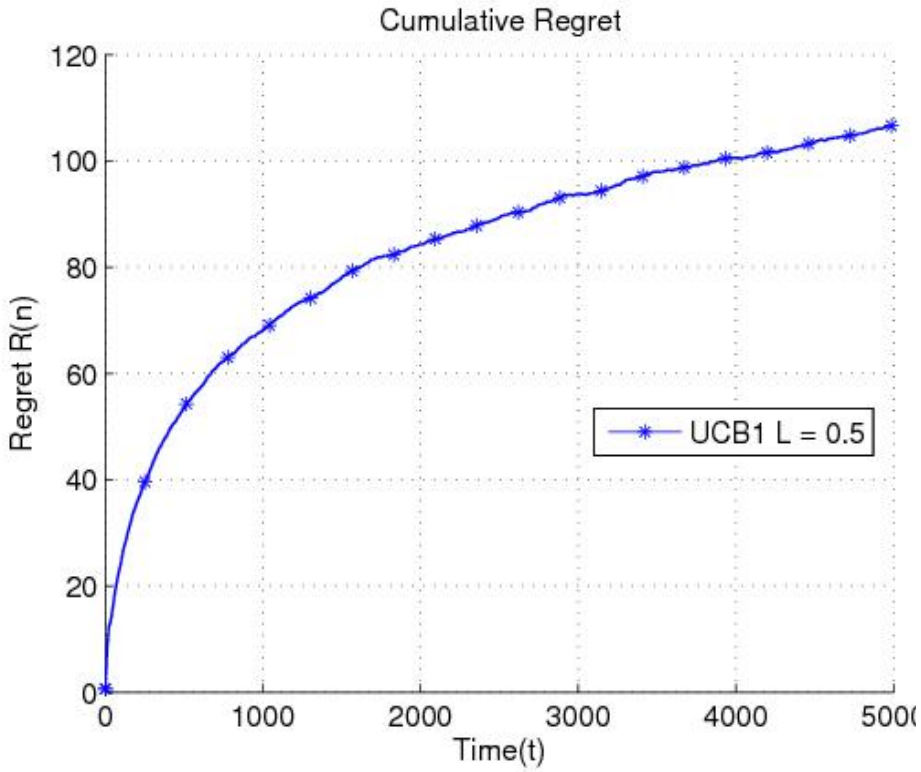
Result: Regret and selection percentage



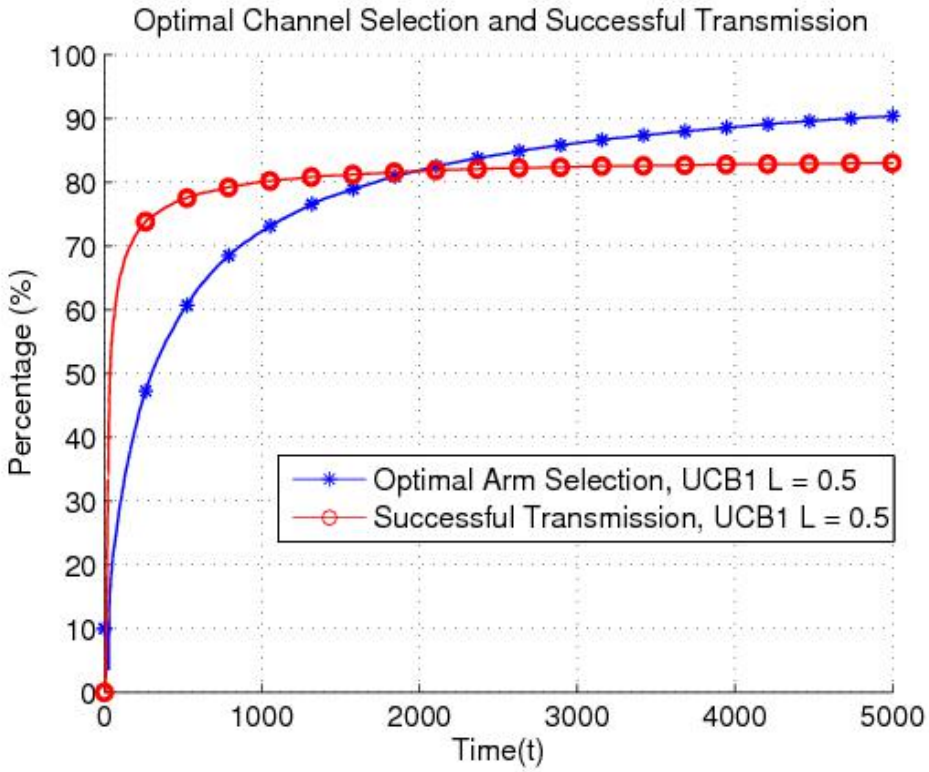
Regret Analysis

Result: Markovian Rewards

Result: Regret and selection percentage

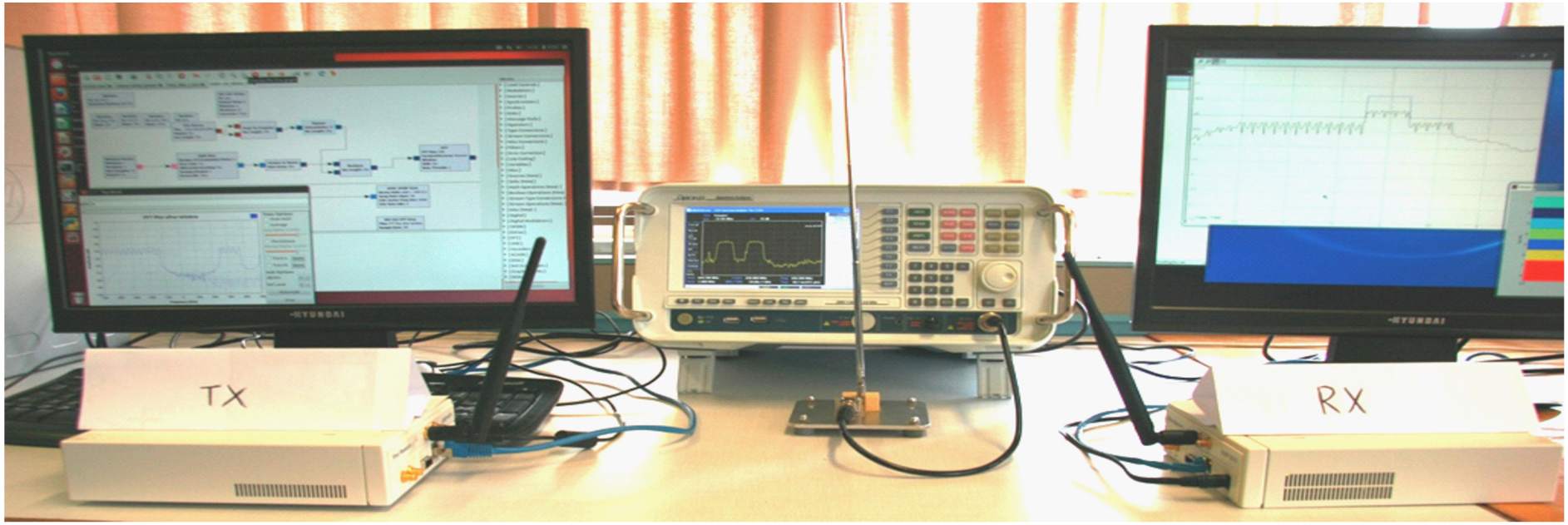


Regret Analysis



Best Arm selection

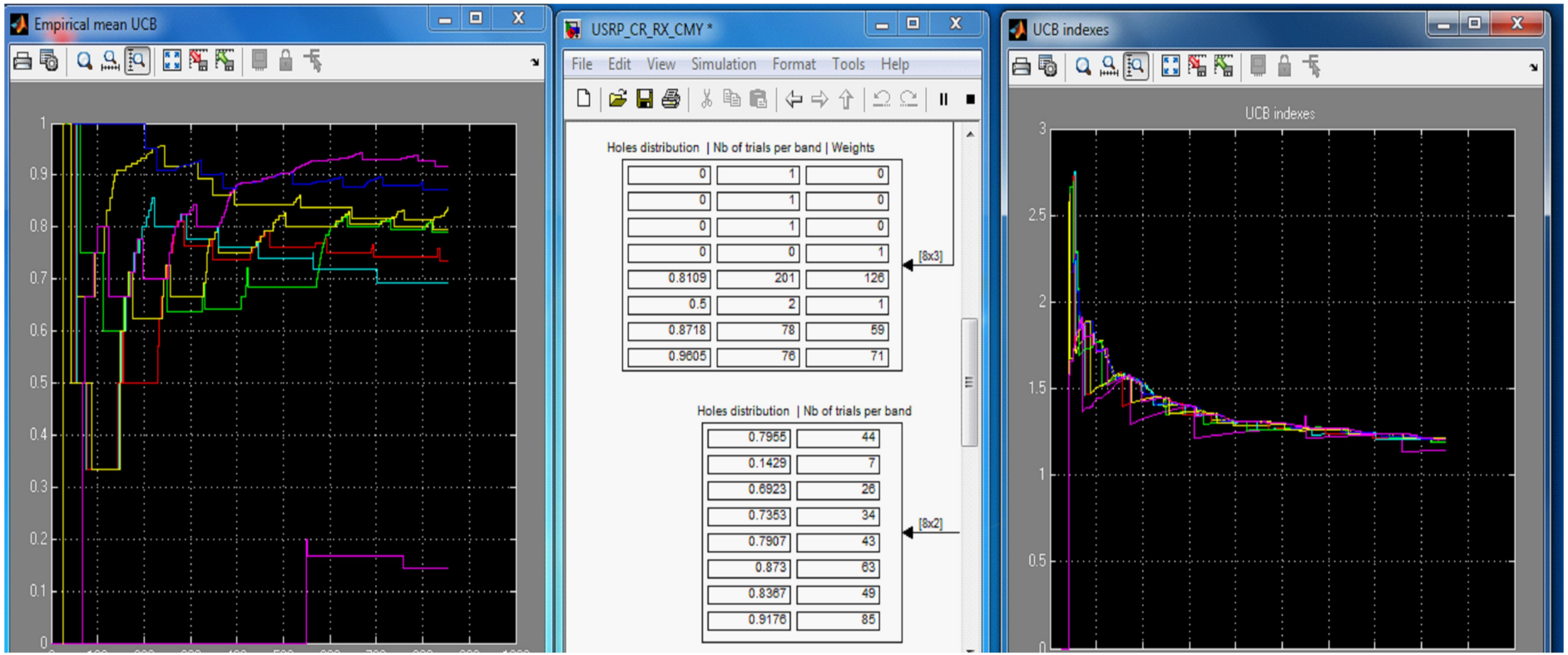
Experimental Setup for OSA[7]



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

[7] Clément Robert, Christophe Moy, and Honggang Zhang. "Opportunistic Spectrum Access Learning Proof of Concept". In: *SDR-WinnComm'14*. Schaumburg, United States, Mar. 2014, 8 pages.

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.