Reinforcement-learning algorithm for cognitive users operating as independent agents in uncertain environments

Angeliki V. Kordali, Panayotis G. Cottis

School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), Athens, Greece
kordali@mail.ntua.gr, pcottis@central.ntua.gr

Primary User's Traffic

- Primary Users' (PUs) traffic follows either deterministic patterns, as in TV transmission, or stochastic patterns, as in packet-switched or circuit-switched networks, where the packet arrival time follows the Poisson process [1].
- Many approaches are based on traffic pattern learning to predict the future traffic in PUs channels such as in [2].
- In [3], the authors classify the channels based on the history of collected data and apply constant monitoring whereas in [4] the channels are characterized by the probability of being idle based on statistics collected in a learning phase.
- Both approaches address specific traffic patterns with static characteristics.
- However, the traffic stochastic patterns cannot, in general, reflect the dynamic changes in the communication channels, especially when these channels are accessed by not registered users as the Secondary Users (Sus). The statistics of channel occupancy vary with time due to changes in traffic load. The SUs have to function in a completely unknown environment with no information about either the traffic pattern followed by the PUs or its specific characteristics (utilization level, frequency of state transitions, etc.)

Reinforcement Learning

- The main principle of Reinforcement Learning (RL) is learning by selection and not by instruction [5].
- The environment is represented by a discrete set of states S where decision makers/agents operate. In the general case, an agent receives an input from the environment, chooses an action $\boldsymbol{a}$ from a set of actions $\boldsymbol{A}$ and receives a reinforcement signal/reward $\boldsymbol{r}$, which depends on the action taken and the current state s of the environment.
- When, in addition to determining the immediate reward, the actions of a user/agent have influence on the subsequent environment states and future rewards, the problem is modeled as a Markov decision process (MDP). Similarly to the formulation of the simple RL problem, an agent at state s, $s \in \boldsymbol{S}$, can select an action $\boldsymbol{a}$ from a discrete set of actions $\boldsymbol{A}$. This selection has two consequences: first, it offers a reward according to a reward function $\boldsymbol{R}: \boldsymbol{S} \times \boldsymbol{A} \rightarrow \boldsymbol{R}$, and, second, it leads to a new environment state $\boldsymbol{s'}$, $\boldsymbol{s'} \in \boldsymbol{S}$, following the state transition function $\boldsymbol{T}: \boldsymbol{S} \times \boldsymbol{A} \rightarrow \boldsymbol{\Pi}(\boldsymbol{S})$, where $\boldsymbol{\Pi}(\boldsymbol{S})$ is a probability distribution over the set $\boldsymbol{S}$.

Q-Learning Algorithm

- It constitutes an online learning algorithm [6][7].
- It is based on the following recursive equation:

$$\boldsymbol{Q_{t+1}(s_t, a_t) = \big(1 - l_t(s_t, a_t)\big) Q_t(s_t, a_t) + l_t(s_t, a_t)\Big[r_t + \gamma \max_{a_{t+1}} Q_t(s_{t+1}, a_{t+1})\Big]}$$

where $\boldsymbol{s'}$ and $\boldsymbol{a'}$ denote the state and action taken at the following time instance, $\boldsymbol{l_t(s_t, a_t)}$ is the learning parameter ($\boldsymbol{0 \le l_t(s_t, a_t) \le 1}$), $\boldsymbol{r_t}$ is the reward and $\boldsymbol{\gamma}$ is a discount factor to account for the contribution of future reinforcements ($\boldsymbol{0 \le \gamma \le 1}$).

- RL algorithms are characterized by two main components [8]; the **update rule** which designates how an agent imports the accumulated experience into the update of the Q-values of the actions and the **learning policy** which specifies the selection of the action at each time instance based on the Q-values.
- In general, $\boldsymbol{Q_{t+1}(s_t, a_t) \rightarrow \mathbb{E}[R_t | a_t = a]}$ with probability 1 if the following conditions hold:

$$\boldsymbol{\sum_{t=1}^{\infty} l_t(s_t, a_t) I\{a_t = a\} = \infty} \quad \text{and} \quad \boldsymbol{\sum_{t=1}^{\infty} (l_t(s_t, a_t))^2 = \infty}$$

where $I\{a_t = a\}$ is an indicator function taking value 1, if $a_t = a$ and 0 otherwise.

System Description

- M available channels occupied by M Primary Users.
- $\boldsymbol{S = (y_{c_1}, y_{c_2}, .., y_{c_M})}$ is the set of the 2^M possible states of the M available channels.
- $\boldsymbol{A = (a_{c_1}, a_{c_2}, .., a_{c_M})}$ is the set of the possible actions an SU can take. An action represents the channel chosen by an SU at a specific time instance, i.e. a_{c_k}=1 denotes that channel c_k is chosen for transmission.
- The SUs are equipped with only one transceiver; hence, parallel transmissions are not feasible and only one channel can be used at a time.
- Time is divided into periods for sensing and transmission.
- The Q-values, kept by the SUs, characterize solely their actions, i.e. their channel choices, and are independent of the current state, i.e. $\boldsymbol{Q_t(s_t, a_{c_k}) = Q_t(a_{c_k})}$
- In case of a successful transmission, the received reward $\boldsymbol{r_t(s_t, a_{c_k})}$ is the throughput related to the specific transmission as quantified by the number of successfully transmitted packets divided by the transmission duration.
- The goal of the SU is to set the channels in a preference order based on the probability of being vacant and the estimated duration of the vacant period.

Proposed Algorithm

Step 1: Initialize Q-values $Q_0(s_0, a_{c_k}) = 1$
Step 2: Evaluate $P(a_{c_i}) = \frac{\exp\{Q(s,a_i)/Temp\}}{\sum_{j=1}^{m} \exp\{Q(s,a_j)/Temp\}}$, $i = 1..m$
Step 3: Set sensing order $O: \{a_{c_1}, a_{c_2}, .., a_{c_M}\} \rightarrow \Re$ based on the probability function P, $j = 1$
Step 4: Execute action a_{c_k} which corresponds to order o_j
If channel c_k is vacant:
Transmit
Receive reward $r_t(s_t, a_{c_k})$
else
go to Step 4 with $j = j + 1$
Step 5: Update Q-values according to the update rule and go to Step 2.

Update Rules for learning procedure

Two update rules are considered for the learning procedure of Step 5:

- ***Learning with constant learning parameter (L-learning)***

When a successful transmission is completed, the Q-value of the probed channel is updated following the Q-learning model, i.e.:

$$\boldsymbol{Q_{t+1}(s, a_{c_k}) = (1 - L) \cdot Q_t(s, a_{c_k}) + L \cdot r_t(s, a_{c_k})}$$

where $\boldsymbol{r_t(s, a_{c_k})}$ is the reward and $\boldsymbol{L}$ is the learning parameter quantifying the weight assigned to the latest information whereas $\boldsymbol{1 - L}$ is the weight assigned to the already accumulated experience. No future rewards are taken into account.

- ***Learning with discounted learning parameter (Time-Learning)***

It constitutes a modified RL-based update rule, where the SU counts the attempts made to access a primary channel $\boldsymbol{c_k}$; this count is denoted $\boldsymbol{\pi_{c_k}}$. The learning rate $\boldsymbol{L}$ is not an a priori defined parameter. Instead, it is related to $\boldsymbol{\pi_{c_k}}$ via

$$\boldsymbol{L = 1/\pi_{c_k}}$$

thus $\boldsymbol{Q_{t+1}(s, a_{c_k}) = ((\pi_{c_k} - 1)/\pi_{c_k}) \cdot Q_t(s, a_{c_k}) + (1/\pi_{c_k}) \cdot r_t(s, a_{c_k})}$

And $\boldsymbol{\pi_{c_k}}$ is the number of times the SU has attempted to access channel $\boldsymbol{c_k}$.

Learning policy

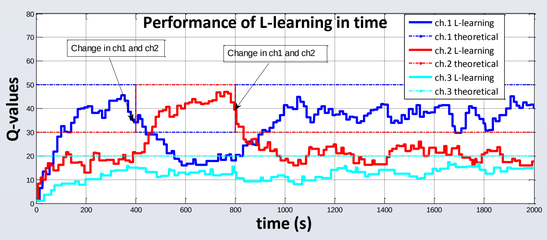
- In the proposed scheme the Boltzmann strategy is employed for the selection of a future action, i.e. which channel to access.

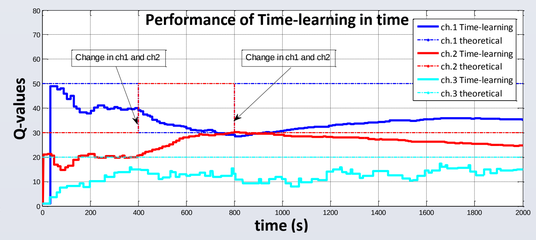
$$\boldsymbol{P(a_{c_k}) = \frac{\exp\{Q(a_{c_k})/Temp\}}{\sum_{j=1}^{m} \exp\{Q(a_{c_j})/Temp\}}}$$

- **Temp** is the temperature parameter which is related to the variance of the Gumbel errors in a Logit discrete choice model. High **Temp** values favor exploration by reducing the importance of the variations of the Q values and low **Temp** values favor exploitation.

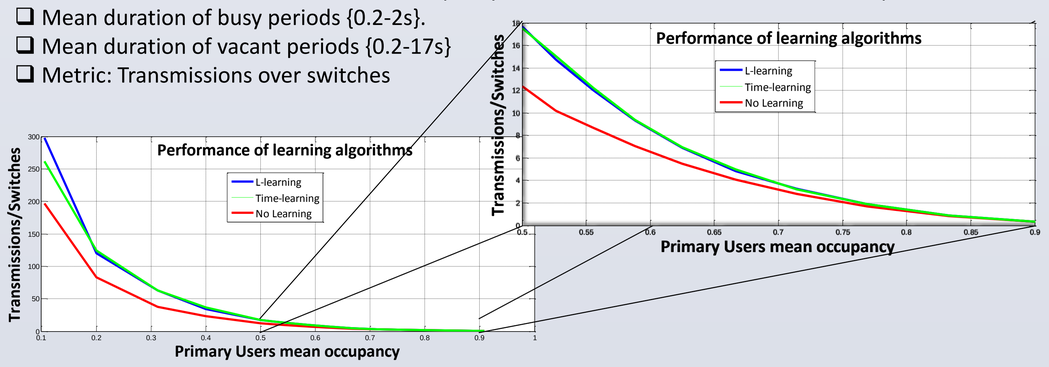
Simulations

- 3 available channels of different mean duration of vacant periods
- Mean duration of vacant periods {1s, 0.6s, 0.4s}, {0.6s, 1s, 0.4s},{1s, 0.6s, 0.4s}.
- Evolution of Q-values in time:





- 10 available channels of same mean occupancy and different mean duration of vacant periods
- Mean duration of busy periods {0.2-2s}.
- Mean duration of vacant periods {0.2-17s}
- Metric: Transmissions over switches



Conclusions

- Both L-learning and Time-Learning offer high exploitation of the available opportunities.
- The ratio of transmits/switches is significantly higher than the case of no learning.
- The suggested algorithm works with any traffic pattern of Primary Users.
- Awareness is achieved based only on information collected by the SU.

Selected References

1) S. Haykin, "Cognitive radio: Brain-empowered wireless communications", *IEEE Journal on Selected Areas in Communications, Vol. 25,* pp. 201–220, 2005.
2) Xiukui Li and Seyed A. Reza Zekavat, "Traffic Pattern Prediction Based Spectrum Sharing for Cognitive Radios", *Cognitive Radio Systems*, Wei Wang (Ed.), ISBN: 978-953-307-021-6, InTech, DOI: 10.5772/7838, 2009.
3) Canberk, B., Akyildiz, I. F., & Oktug, S., "Primary User Activity Modeling Using First-Difference Filter Clustering and Correlation in Cognitive Radio Networks", *IEEE/ACM Transactions on Networking, Vol.19, No. 1*, 2000.
4) Yun, G., Grammenos, R. C., Yang, Y., & Wang, W., "Performance Analysis of Selective Opportunistic Spectrum Access With traffic Prediction", *IEEE Transactions on Vehicular Technology, Vol.59, No.4*, 2000.
5) R.S. Sutton, A.G. Barto, "Reinforcement Learning", *MIT Press, Cambridge*, *MA*, 1998.
6) C. J. C. H. Watkins and P. Dayan, "Technical note: Q-learning", *Machine Learning, 8(3/4)*:279–292,May 1992.
7) L.P. Kaelbling, M.L. Littman, A.W. Moore, "Reinforcement learning: A survey", J. *Artificial Intelligence* Res., 4, pp. 237–285, 1996.
8) S. Singh, T. Jaakkola, M.L. Littman, C. Szepesvári, "Convergence results for single-step on-policy reinforcement-learning algorithms", *Machine Learning*, 38, pp. 287–308, 2000 .

Acknowledgment

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.