# Enhanced Online Q-Learning Scheme for Energy Efficient Power Allocation in Cognitive Radio Networks

Ismail AlQerm* and Basem Shihada †
* Dept. of Mathematics and Computer Science, University of Missouri St. Louis, USA, Email: alqermi@umsl.edu
† CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia, Email: basem.shihada@kaust.edu.sa

***Abstract*—The considerable growth in demands for wireless services have led to spectrum scarcity challenge. Cognitive radio came into practice to deal with the scarcity problem by granting cognitive users access to the licensed spectrum. However, this solution requires efficient power allocation strategies to guarantee QoS for cognitive system, reduce power consumption, and protect primary users from the cognitive users' interference impact. In this paper, we investigate the energy efficient power allocation problem for cognitive radio networks in underlay mode. We propose a novel approximated online Q-learning scheme for power allocation in which cognitive users learn with conjecture feature to select the most appropriate power level. The power allocation problem is formulated as an optimization problem with the goal to maximize energy efficiency under QoS and interference constraints. The scheme is evaluated using software defined radio testbed and simulations. The evaluation results demonstrate the scheme capability to guarantee SINR for both primary and cognitive systems and mitigate interference with minimum power consumption in comparison with other schemes.**

***Index Terms*—Energy Efficiency, Cognitive Radio, Power Allocation, Online Q-Learning**

## I. Introduction

With the explosive growth of wireless networks, the radio resources become scarce, which restricts the development of wireless services and applications. On the other hand, studies reveal that the licensed spectrum is underutilized since 70% of it is not used [1], while the unlicensed spectrum is overwhelmed with communication devices [2]. Cognitive radio (CR) is envisioned as the solution that improves spectrum efficiency through appropriate spectrum resources allocation. The spectrum resources allocation must not hamper the communications of the primary network. Since CR network is a highly dynamic, interference caused by the CR network transmissions not only impact the primary users but also compromise energy consumption of the cognitive networks. In addition, the transmission of cognitive users over the allocated spectrum should satisfy application QoS requirements. Therefore, energy efficiency is a fundamental problem to consider in the context of cognitive networks. Efficient power allocation techniques are required to mitigate the interference impact of the cognitive users on the primary network, maximize energy efficiency, and maintain signal to interference and noise ratio (SINR) for all users.

There were several proposals in the literature to study the power allocation problem with system capacity maximization goal. For example, the authors in [3] proposed a distributed power allocation to maximize the system capacity, where they consider cognitive Gaussian multiple access channels on which the maximization is formulated as a standard non-convex quadratically constrained quadratic problem. In [4], the cognitive users' throughput is maximized with assurance of primary users data rate. A power control problem in cognitive networks, which maximizes data rate under primary users' interference power constraint is investigated in [5]. In [6], power allocation scheme is proposed in underlay cognitive networks with arbitrary input distributions. Resource allocation with energy consideration has been studies in [7], focusing on the maximization of the system throughput for unit-energy consumption of the transmission of second users while meeting the interference constraints requirement. The authors in [8] proposed multiple sub-algorithms to manage the cognitive interference and improve both the spectrum efficiency and the energy efficiency of cognitive users while maximizing the sum effective capacity of the cognitive network. The power control problem has also been studied in [9] using game theory based on cost function, in which each user tries to minimize its own cost to achieve the target SINR. The cost function in [9] has been defined as a weighted sum of power and square of SINR error. Most of the researchers focus on the development of power control algorithms in the cognitive network and neglect the interaction of primary network, which makes power control techniques similar to ones applied to non-cognitive networks. In addition, all the proposed resource allocation schemes consider specific model of the network and ignore the fact that the cognitive networks are dynamic and unpredictable. Machine learning has been utilized for power allocation in cognitive networks as in [10] and [11]. However, these schemes targeted maximization of the data rate of the cognitive users with relaxed constraints in which primary users interference is not considered. In addition, they use typical Q-learning without approximation which compromises the convergence speed as the computation required increases.

To the best of our knowledge, our scheme is the first for

power allocation using approximated and conjecture based online Q-learning with energy efficiency consideration. Approximated online Q-learning with conjecture feature allows cognitive users to surmise the power allocation strategies of each other without explicit cooperation.

In this paper, we propose a novel distributed power allocation scheme that aims to minimize power consumption of the cognitive network with SINR efficiency tracking. It exploits computation efficient approximated online Q-learning, where each cognitive user conjectures other cognitive users' strategies for power allocation and the Q-value is approximated. This enhances the scheme capability to determine the most appropriate transmission power. The contributions of the paper can be summarized as:

- Energy efficient power allocation scheme in CR with consideration of interference to primary users, the maximum allowed power for the cognitive users and QoS for all systems.
- An approximated online Q-learning algorithm is developed to allocate transmission power in a distributive manner. The approximation of Q-value in the learning reduces the considered state space which minimizes the required computation to reach the ultimate power allocation policy.
- A conjecture feature is developed in the proposed online Q-learning which allows each cognitive user to conjecture the power allocation policy of other cognitive users in the network without explicit cooperation. This will enhance the efficiency of the selected power allocation policy in addition to elimination of the overhead of cooperation among the cognitive users in such dynamic environment.
- The power allocation policy selection probability is derived as a graded function of Q-value, where all the allocation policies are ranked according to their Q-value. Thus, the power allocation policy with the highest Q-value will be selected.

The rest of the paper is organized as follows, the system model and problem formulation are described in Section II. Section III presents the online Q-learning mechanism for power allocation. The evaluation results are discussed in Section IV and the paper concludes in Section V.

## II. System Description

In this section, the system model of the considered cognitive network and the energy efficiency optimization problem are described.

### A. System Model

We consider the infrastructure topology with orthogonal frequency division multiple access (OFDMA to access the shared spectrum, where there are multiple cognitive users (CUs) associated with a cognitive base station (BS). This cognitive BS operates under the coverage of a primary BS with primary users (PUs) associated with it as in Fig. 1. The focus of this work is on the uplink connection. There are $N$ CUs uniformly distributed and the channel state information is assumed to be captured by cognitive receiver. The transmission

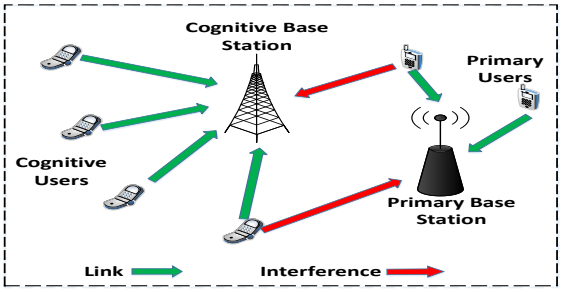



Fig. 1. System Model

power and the channel gain (between CU and its cognitive BS) for each CU $i$ are denoted as $P_i$ and $H_i$ respectively. The SINR at the cognitive BS communicating with $i$th CU is expressed as follows,

$$\gamma_i(P_i) = \frac{P_i H_i}{\sum_{j=1, j\neq i}^{N} P_j H_{ji} + \sum_{k=1}^{K} P_k H_{ki} + \sigma^2} \quad (1)$$

where $P_j$ is the transmission power of other CUs, $H_{ji}$ is the gain between CU $j$ and CU $i$, $P_k$ is the transmission power of the primary user $k$, $H_{ki}$ is the gain between CU $i$ and PU $k$, and $\sigma^2$ is the noise power. The SINR of the primary system is given by,

$$\gamma_k(P_k) = \frac{P_k H_k}{\sum_{i=1}^{N} P_i H_{ik} + \sigma^2} \quad (2)$$

where $H_{ik}$ is the gain between CU $i$ and PU $k$

In this model, we aim to satisfy three objectives: (i) maximize the energy efficiency of the cognitive network. (ii) Limit the interference caused by the cognitive transmissions to the primary network (denominator of (2)) to the interference threshold $I_{Th}$. (iii) Maintain SINR requirements for CUs and PUs.

In CR networks, each CU transmits its information over the air, which is a common medium for all transmitters and this creates the interference between the CUs and other users. The interference plus fading, multi-path and background noise cause signal distortion as it is traveling from the source to the destination. In addition, transmission power is an essential commodity for the CUs as they are battery powered. Therefore, user satisfaction in such network is determined according to transmission power and SINR.

### B. Problem Formulation

In order to achieve objectives stated in the system model, the proposed objective function of the power allocation must account for the following:

- The energy efficiency which is a function of CU's transmission power $P_i$, circuit power, and CU's SINR. The SINR is a function of CU $i$ transmission power $P_i$ and the transmission power of other users $P(-i)$.
- When CU $i$ increases its transmission power, this increases its SINR. However, this will decrease the SINR for other CUs.

Information are sent in the cognitive networks in the form of frames. The achieved throughput $T$ with assumption that

all errors in the received signal can be detected by the system and the incorrect data can be re-transmitted is evaluated as follows,

$$T = R * \phi(\gamma) \tag{3}$$

where $R = B(1 + \log_2(\gamma_i(P_i)))$ is the data rate and $\phi(\gamma)$ is the efficiency function of transmission, which depends on the SINR and $\phi(\gamma) \in [0, 1]$. The energy efficiency of CU $i$ with transmission power $P_i$ is defined as the number of information bits received successfully per joule of the energy consumed as follows,

$$EE_i(P_i, P_{-i}) = \frac{R\phi(\gamma_i)}{P_i + P_c} \tag{4}$$

where $P_c$ is the circuit power consumption.

The efficiency function $\phi(\gamma)$ is defined as a sigmoid function, which is an exponential ratio of the targeted SINR and the achieved SINR as follows,

$$\phi(\gamma_i) = \exp\big(-(\frac{c\Gamma_i}{\gamma_i})^v\big) \tag{5}$$

where $c$ and $v$ are non-negative weighing factors. The efficiency function defined in (5) is sigmoidal function with $\phi(\infty) = 1$ and $\phi(0) = 0$ to maintain $EE_i = 0$ when $P_i = 0$. The energy efficiency function in (4) now is given as follows,

$$EE_i(P_i, P_{-i}) = \frac{R}{P_i + P_c} \exp\big(-(\frac{a\Gamma_i}{\gamma_i})^b\big) \tag{6}$$

The function in (6) shows a tradeoff between the throughput and the power consumed. When the target SINR is fixed, the energy efficiency function still can be tuned using the weighing factor $c$. Consequently, CU $i$ transmission power is adjusted according to the maxima of the energy efficiency function. The energy efficiency increases by decreasing $c$ and reducing the transmission power. However, this will reduce the target SINR of the system. The primary system will broadcast the best value of $c$ to the cognitive network to tune the target SINR according to the experienced interference (i.e. the primary system will broadcast low value of $c$ if the experienced interference reaches the interference threshold.

The energy efficient power allocation problem is defined as follows,

$$\max_{P_i \in \mathbf{P}} EE_i(P_i, P_{-i}) = \frac{R}{P_i + P_c} \exp\big(-(\frac{a\Gamma_i}{\gamma_i})^b\big) \;\; s.t \tag{7}$$

$$C1 : \sum_{i=1}^{N} P_i \leq P_{max}$$

$$C2 : \sum_{i=1}^{N} P_i H_{ik} + \sigma^2 \leq I_{Th}$$

$$C3 : \gamma_i(P_i) \geq \Gamma_i, \;\; \gamma_k(P_k) \geq \Gamma_k$$

The constraint C1 ensures the total transmit power of each cognitive user is lower than its corresponding power budget. Constraint C2 guarantees that the permissible interference power and noise level generated by cognitive transmission do not exceed the tolerance interference $I_{Th}$ for the primary system. The QoS requirements of cognitive system and primary system are represented by their achieved SINRs that are maintained through C3, where $\Gamma$ is the SINR threshold.

## III. Online Q-Learning Energy Efficient Power Allocation Mechanism

In this section, we assume that all CUs learn in team game with a common goal of finding a power allocation policy to maximize the objective function given in (7) using approximated online Q-learning [12]. Online Q-learning is a model-free reinforcement learning technique. Specifically, it can be used to find an optimal action-selection policy. It works by learning an action-value function that ultimately gives the expected utility of taking a given action in a given state and following the optimal policy thereafter. The considered network environment includes a discrete set of states $\mathbf{X}$ and a discrete set of actions $A$. At each time step $t$, the learning agent acquires network state information $\mathbf{X}$ and selects certain action to perform. Consequently, the environment makes a transition to state $X'$ at time step $t + 1$ with probability $T_{XX'}(a)$ and receive certain reward $RW = RW(X, a)$. This process is iterative and repeated infinitely to converge to an optimal decision-making policy $\pi$ that maximizes the total received reward. This policy is a mapping from environment states to probability distributions over actions.

Each CU has the role of a learning agent, which aims to reach optimal power allocation strategy for different network states. The online learning parameters are defined as follows:

- **State** since there is no cooperation among the competing CUs, they only rely on the local observation to define their environment state at certain time instant $t$. The state observed by CU $i$ is defined as,

$$X_i^t = (i, P_i, \gamma_i, \Gamma_i, \Gamma_k) \tag{8}$$

- **Action:** the action is defined as the CU $i$ transmission power $(P_i)$.
- **Reward:** the reward function is defined for the state/action pair as the energy efficiency objective function $RW_i(X_i, P_i, P_{-i}) = EE_i(P_i, P_{-i})$.

We evaluate the optimal Q-value of CU $i$ as the current expected reward plus a future discounted reward when all other CUs follow the optimal strategy as follows,

$$Q_i^*(X_i, P_i) = E[RW_i(X_i, P_i, P_{-i}(X_i))] + \beta \sum_{X_i' \in \mathbf{X}_i} T_{X_i, X_i'}(P_i, P'_{-i}(X_i)) \max_{P_i' \in \mathbf{P}_i} Q_i^*(X_i', P_i') \tag{9}$$

where $T_{X_i, X_i'}(.)$ is the state transition probability, $P_i'$ is the action associated with state $X_i'$, and $\beta$ is the discount factor. The employed online Q-learning scheme aims to reach the optimal Q-value defined in (9) in a recursive way using the information $(P_i, X_i, X_i')$ with the two states $X_i = X_i^t$ and $X_i' = X_i^{t+1}$ observed at the time instant $t$ and $t + 1$ respectively. The update rule for the online Q-learning employed to reach the optimal Q-value is given by,

$$Q_i^{t+1}(X_i, P_i) = (1 - \zeta^t) Q_i^t(X_i, P_i) + \zeta^t$$

$$\Big\{ \sum_{P_{-i}\in \mathbf{P}_{-i}} [RW_i(X_i, P_i, P_{-i}) + \beta \max_{P_i^* \in \mathbf{P}_i} Q_i^t(X_i', P_i')\Big\} \quad (10)$$

where $\zeta \in [0, 1)$ is the learning rate. The considered online Q-learning model here is a stochastic approximation method that solves the Bellman's optimality equation associated with the discrete time environment. Online Q-learning does not require explicit state transition probability model and it converges with probability one to an optimal solution if $\sum_{t=1}^{\infty} \zeta^t$ is infinite, $\sum_{t=1}^{\infty} (\zeta^t)^2$ is finite, and all state action pairs are visited infinitely often [13]. Balancing exploration and exploitation is an essential issue in the stochastic learning process. Exploration aims to try new allocation strategies so it does not only apply the strategies it already knows to be good but also explore new ones. Exploitation is the process of using well-established strategies. The most common technique to achieve exploration vs exploitation balance is to use the $\epsilon$-greedy selection [14], where $\epsilon$ is the percent of the time that an agent takes a randomly selected action rather than taking the action that is most likely to maximize its reward given what it knows so far. It usually starts with a lot of exploration (i.e. a higher value for $\epsilon$). Over time, as the agent learns more about the environment and which actions yield the most long-term reward, it steadily reduce $\epsilon$ as it settles into exploiting what it knows. However, $\epsilon$-greedy selects equally among the available actions i.e. ( the worst action is likely to be chosen as the best one). In order to overcome this drawback, the action selection probabilities are varied as a graded function of the Q-value. The best power level is given the highest selection probability, while all other levels are ranked according to their Q-values. The learning algorithm exploits Boltzmann probability distribution [15] to determine the probability of the resource allocation action that fulfills the energy efficiency maximization constraints in C1 to C3. Thus, the action $P_i$ in state $X_i$ is selected at $t$ with the following probability,

$$\pi_i^t(X_i, P_i) = \frac{e^{Q^t(X_i,P_i)/\tau}}{\sum_{P_i'\in \mathbf{P}_i} e^{Q^t(X_i,P_i')/\tau}} \quad (11)$$

where $\tau$ is a positive integer that controls the selection probability. With high value of $\tau$, the action probabilities become nearly equal. However, low value of $\tau$ causes big difference in selection probabilities for actions with different Q-values. One issue to report is that the CR environment has a large space. Therefore, the curse of dimensionality increases the required computations and makes it unfeasible to use the typical online Q-learning methodology to maintain the Q-value for each state/action pair, which slows the system convergence. Therefore, we propose a brief representation for the Q-values in which they are approximated as a function of much smaller set of variables to account for the curse of dimensionality. The brief representation of Q-value focuses on a countable state space $\mathbf{X}^*$ using the function $Q' : \mathbf{X}^* \times Y$, which is referred as a function approximator. The parameter vector $\xi = \{\xi_z\}_{z=1}^{Z}$ is adopted to approximate the Q-value by minimizing the metric of difference between $Q^*(X_i, P_i)$ and $Q'(X_i, P_i, \xi)$ for all $(X_i, P_i) \in \mathbf{X}^* \times \mathbf{P}_i$. Thus, the approximated $Q'$ value is formalized as follows,

$$Q'(X_i, P_i, \xi) = \sum_{z=1}^{Z} \xi_z \psi_z(X_i, P_i) = \xi \psi^T(X_i, P_i) \quad (12)$$

where $T$ denotes the transpose operator and the vector $\psi(X_i, P_i) = [\psi_z(X_i, P_i)_{z=1}^{Z}]$ with a scalar function $\psi_z(X_i, P_i)$ defined as the basis function (BF) over $\mathbf{X}^* \times \mathbf{P}_i$, and $\xi_z(z = 1, ...., Z)$ are the associated weights. A gradient function $\psi(X_i, P_i)$, which is a vector of partial derivative with respect to the elements of $\xi^t$, is used to combine the typical online Q-learning model defined in (10) with the linearly parametrized approximated online learning proposed.

The Q-value update rule in (10) is reconstructed to include the parameter vector updates as follows,

$$\xi^{t+1}\psi^T(X_i, P_i) = \big\{(1-\alpha^t)\xi^t\psi^T(X_i, P_i) + \alpha^t\big[RW_i(X_i, P_i, P_{-i}) + \beta \max_{P_i'\in \mathbf{P}_i} \xi^t\psi^T(X_i', P_i')\big]\big\}\psi(X_i, P_i) \quad (13)$$

The probability of selecting certain action presented in (11) is updated with the Q-value approximation as follows,

$$\pi^t(X_i, P_i) = \frac{e^{\xi^t\psi^T(X_i,P_i)/\tau}}{\sum_{P_i'\in \mathbf{P}_i} e^{\xi^t\psi^T(X_i,P_i)/\tau}} \quad (14)$$

The optimal Q-function, $Q_i^*(X_i, P_i)$ for all $(X_i, P_i) \in \mathbf{X} \times \mathbf{P}_i$, defines the optimal joint power allocation strategy and captures the team game structure. For each network state $X_i \in \mathbf{X}$, the CUs play a team stage game $\Psi_x = [N, \{\mathbf{P}_i\}, Q^*(\mathbf{X}, .)]$ and consider $Q^*(\mathbf{X}, .)$ to be independent. Note that the action in the game is jointly generated by the $N$ independent CUs in a distributed manner. A joint power allocation action $P_i$ is optimal if $Q_i^*(X_i, P_i) \geq Q_i^*(X_i, P_i')$ for all $P_i' \in \mathbf{P}_i$. We assume that the power allocation strategies of different CUs do not change significantly in the same network states. The initial network state process $\{X_i(t)\}$ evolves following irreducible and Harris recurrent Markov chain [16].

The similarity between two network states $X_i$ and $X_i'$ can be determined in term of Hamming distance [17], which is denoted by $D(X, X')$. Thus, each CU can conjecture the power allocation strategies employed by other CUs for the current network state through making use of historical knowledge. This knowledge up to time instant $t$ is given by,

$$F(t) = (\{X_i(b), P_i(b)\}_{b=1}^{t}, \{RW_i(X_i(bz), P_i(b))\}_{b=1}^{t-1}) \quad (15)$$

In each time instant $t$, every CU checks the Hamming distance between the current state $X_i(t)$ and the state $X_i(b)$ in $F(t)$ and obtains a set $\mathbf{X}_F^*(X_i(t), F(t))$, which includes the $F$ most recent observations from $F(t)$ that minimizes $\sum_{f=1}^{F} D(X_i(t), X_i(b))$. Let us define $E(X_i(t), .)$ as the common reward that all CUs receive after they perform the joint power allocation action $P_i \in \mathbf{P}_i$ and is set to be 1 if $P_i = \max_{P_i'\in \mathbf{P}_i} Q_i(X_i(t), P_i')$ and 0 otherwise. Since the CUs learn in a distributed manner, we choose $\mathbf{P}_i^{"}(X_i(t))$ for each CU $i$ to denote the set of joint actions that output the payoff 1 in state $X_i(t)$. The mechanism of power allocation

for CU $i$ that maximizes its energy efficiency is illustrated in Algorithm 1. We assume that each CU $i$ updates its transmission power at time instances $T_i = \{t, t+1, ....\}|$ where $t < t+1$. Suppose two integers $l$ and $q$ that satisfy $1 \leq l \leq F \leq q$. The algorithm starts by checking the learning

**Algorithm 1** Energy Efficient Power Allocation Algorithm

**Require:** CU $i$, $t = 1$, power vector $\mathbf{P} = [P_1, .....P_N]$, Network state $X_i(t)$
**Ensure:** proper $P_i(t)$ that maximizes the function in (7)
**BEGIN**
initialization of Learning
**for** each($X_i$, $P_i \in \mathbf{P}_i$) **do**
  initialize power allocation strategy $\pi^t(X_i, P_i)$;
  initialize approximated Q-value $\xi^t \psi^T(X_i, P_i)$;
**end for**
**if** $(t < q+1)$ **then**
  Select action $P_i$ according to $\pi^t(X_i, P_i)$ in (14);
  **if** (C1 to C3 are satisfied ) **then**
    $RW_i(X_i, P_i)$ is achieved
  **else**
    $RW_i(X_i, P_i) = 0$
  **end if**
**else**
  Update $\mathbf{P}_i^{"}(X_i(t)) = \{P_i | E(X_i(t), P_i) = 1\}$ for $X_i(t)$
  Randomly select $\mathbf{P}_F(\mathbf{X}^*(X_i^t, F(t)))$ out of $F$ joint actions associated with $\mathbf{X}^*(X_i^t, F(t))$
  Calculate $E'(X_i(t), P_i')\}$ according to (16) and populate $\mathbf{P}_i^o(X(t))$
  **if** (i.1) and (i.2) are satisfied **then**
    select action from $P_i \in \mathbf{P}_F(\mathbf{X}_F^*(X_i(t), F(t))) \cap \mathbf{P}_i^{"}(X_i(t))$
  **else**
    select action from $\mathbf{P}_i^o(X_i(t))$
  **end if**
**end if**
Update $\xi_i^{t+1}\psi^T(X_i, P_i)$ according to (13)
Update $\pi_i^{t+1}(X_i, P_i)$ according to (14)
$X_i = X_i^{t+1}$
$t = t+1$
**END**

condition, if $t < q+1$, all CUs select their transmission power according to the probability in (14). From $t = q+1$, each CU $i$ selects $l$ records $\mathbf{P}_F(\mathbf{X}_F^*(X_i(t), F(t)))$ from the $F$ joint actions with respect to $\mathbf{X}_F^*(X_i(t), F(t))$. If (i.1) there is a joint power allocation action $P = (P_i, P_{-i}) \in \mathbf{P}_i^{"}(X(t))$ such that $P'_{-i} = P_{-i}$ for all $P' = (P_i', P'_{-i}) \in \mathbf{P}_F(\mathbf{X}_F^*(X_i(t), F(t)))$. (i.2) there exists at least one joint action $P_i$ such that $P_i \in \mathbf{P}_F(\mathbf{X}_F^*(X_i(t), F(t))) \cap \mathbf{P}_i^{"}(X(t))$, then, CU $i$ selects the power allocation action $P_i(b^*)$ where $b^* = \max_b\{b|P_i(b) \in \mathbf{P}_F(\mathbf{X}_F^*(X_i(t), F(t))) \cap \mathbf{P}_i^{"}(X(t))\}$. If the conditions (i) and (ii) are not met, CU $i$ selects the action from $\mathbf{P}_i^o(X(t)) = \{P_i | P_i = \max_{P_i'} E'(X_i(t), P_i')\}$, where,

$$E'(X_i(t), P_i') = \sum_{P_{-i}} E(X_i(t), P_i) \frac{A_i^t(X_i(t), P_{-i})}{q} \quad (16)$$

The expected value in (16) is calculated using the $q$ records selected from $F$ most recently performed actions. $A_i^t(X_i(t), P_{-i})$ denotes the number of times the other CUs perform the joint action $P_{-i}$ in state $X_i(t)$.

## IV. Performance Evaluation

In this section, we demonstrate the performance of our proposed power allocation scheme. The performance is evaluated using simulation and testbed implementation using Software Defined Radio (SDR).

### A. Simulation Results

We consider a system that comprises a single-cell cognitive radio and a single primary BS with 15 cognitive users associated with the cognitive BS and uniformly distributed. The system assumes fixed frame size and no coding for forward error correction. The propagation model considered in the simulation includes path gains with deterministic function and path loss component $\kappa$ with the distance between the CU $i$ and the cognitive BS as follows,

$$H_i = \frac{K}{d_i^\kappa} \quad (17)$$

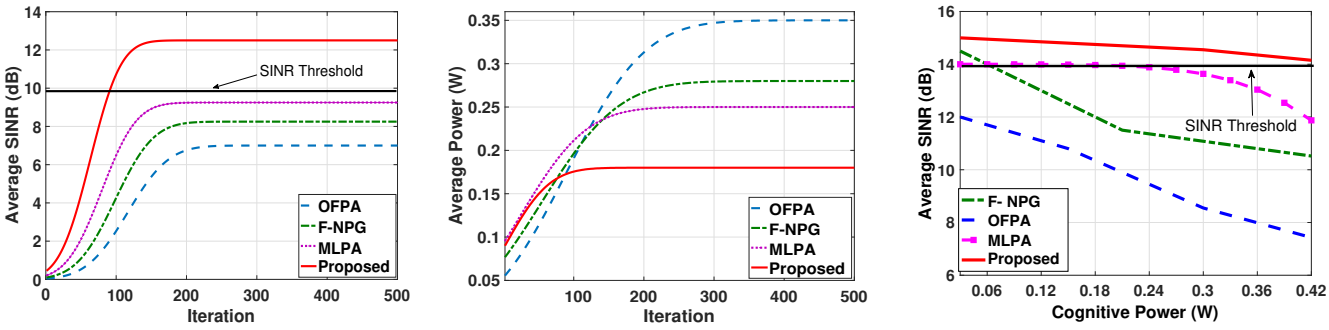
where $d_i$ is the distance between the CU $i$ and the cognitive BS, $\kappa = 4$, and $K = 0.097$ is a constant. This value of $K = 0.097$ is selected to establish a transmit power of 0.5 W for a CR terminal operating at 1140 meters from the cognitive BS. All CUs start with initial transmission power $P_i = 2.2 \times 10^{-16}$ W and $\tau = 10^{-5}$. The weighing factors are tuned according to the primary system feedback to achieve the target SINR. The rest of simulation parameters are stated in Table I. The results

| Parameter | Value |
|---|---|
| $P_i^{max}$ | 0.5 W |
| Number of bits per frame $L$ | 80 |
| Cognitive system SINR threshold $\Gamma_i$ | 10 dB |
| number of primary users | 3 |
| Noise power ($N_0$) | -86 dBm |
| Data rate | 10 Kbps |
| Discount factor $\beta$ | 0.7 |
| Learning rate $\zeta$ | 0.3 |

TABLE I
Simulation parameters

obtained are compared with the performance of the OFDMA power allocation scheme (OFPA) proposed in [8], the machine learning power allocation scheme (MLPA) proposed in [10], and the fair power control algorithm using game theory (F-NPG) proposed in [18]. In this simulation, we evaluate the performance of the power allocation scheme in terms of the average SINR achieved for the cognitive system, its average power consumption, the average SINR of the primary system, and the average energy efficiency of the cognitive system.

The average cognitive users SINR is plotted as a function of the number of iterations in Fig. 2 (a). The figure shows that the proposed algorithm maintains the average SINR of the cognitive system at 12 dB, which is above the threshold. In addition, it shows that our proposed scheme is the fastest in convergence compared to other competing algorithms including the legacy Q-learning. Fig. 2 (b) presents the power consumed for transmission for the proposed scheme compared to other algorithms. We can observe that it has a significant saving in the power consumed. The result obtained from Fig. 2 (b) indicates that the amount of interference measured at




(a) Average SINR of the cognitive system

(b) Average cognitive user power level for all algorithms

(c) Average SINR for the primary system

Fig. 2. Simulation evaluation for cognitive SINR, cognitive power consumption, and SINR of primary system

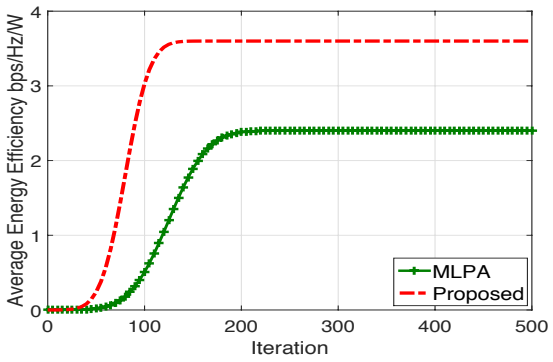



Fig. 3. Average energy efficiency of the cognitive system

the primary system for the proposed algorithm is the lowest. This feature makes our scheme the best for maximizing the spectrum sharing and QoS guarantees in both primary and cognitive systems. The QoS of the primary system is evaluated as the average SINR against the transmission power of the cognitive users. The primary system SINR threshold is chosen to be $14dB$. Fig. 2 (c) shows the average primary SINR achieved by our scheme compared to other algorithms. We notice that the SINR is kept above the threshold regardless of the increase of the average cognitive user transmission power, which is not the case for other algorithms. Fig. 3 presents the achieved energy efficiency of the cognitive system comparing our proposed scheme with the machine learning scheme (MLPA). The figure shows that our proposed scheme outperforms MLPA not only in the achieved energy efficiency but also in the convergence speed.

As a result, the proposed online Q-learning power allocation scheme shows tremendous performance improvement in terms of SINR, power consumption, and interference mitigation. The use of the conjecture feature in online Q-learning allows the system to be aware of other cognitive users power allocation strategies and supports the appropriate selection of the transmission power.

### *B. Testbed Implementation*

The testbed is implemented using GNU Radio and USRP-N210 SDR platform [19]. The USRP-N210 is employed to obtain spectrum occupancy information represented by the power level of the primary users. This information is sent from the SDR platform to the GNU blocks for processing via Ethernet interface. Data transmission and information acquisition is processed separately. Therefore, GNU time frame is divided into two periods: one for data processing and one for data transmission. The network model comprises primary and cognitive users transmit in the same band, which is 2.4 GHz ISM band. The number of available channels is assumed to be equal to the number of primary users. The frequency starts at 2.404GHz, which is channel 1 and ends at 2.444 GHz, which is channel 11. The optimization part from game theory functions to adapt transmission power in order to allocate channels with minimum interference. The bitrate of the primary system is 500kb/s with 1 MHz bandwidth. The size of the transmitted frames is 1500 bytes. The traffic of primary users is based on ON-OFF mechanism as in [20] where frames are generated every 30ms within the ON time.

The implementation setup incorporates four USRP-N210, they are labeled as A, B, C and D. USRP A acts as a cognitive transmitter, B as a spectrum monitor, C as a cognitive receiver and D as a primary transmitter. Daughter board used in this implementation is RFX2400, which covers frequencies from 2.3GHz to 2.9GHz. As we follow the underlay spectrum access paradigm, we have only one channel for testing and it was channel 2. The testing was conducted in isolated area where there is no external interference. The spectrum information is collected via USRP B, which represents the spectral occupancy of the primary system and the corresponding power level. Cognitive users with their resource allocation scheme are capable to adapt their radio transmission parameters according to the information obtained from USRP B. We demonstrate the capability of our scheme through the obtained results in this implementation. The evaluation aims to demonstrate the scheme capability to maintain the average SINR for both cognitive and primary systems with variable transmission power of both. As we can see in Fig. 4, The SINR achieved at the cognitive receiver remains above the threshold regardless of the primary user transmission power even at low transmission power level where primary detection is difficult. Moreover, when primary user power is high, the cognitive user increases its power as the primary user can tolerate more interference.

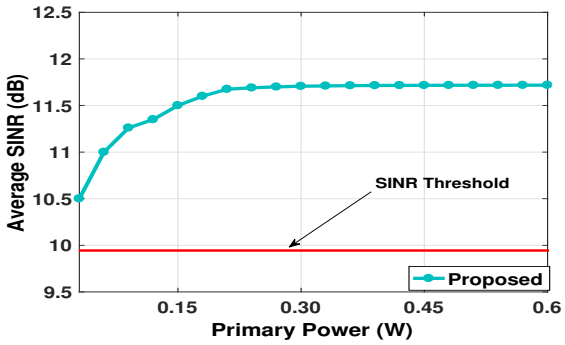



Fig. 4. Average cognitive system SINR

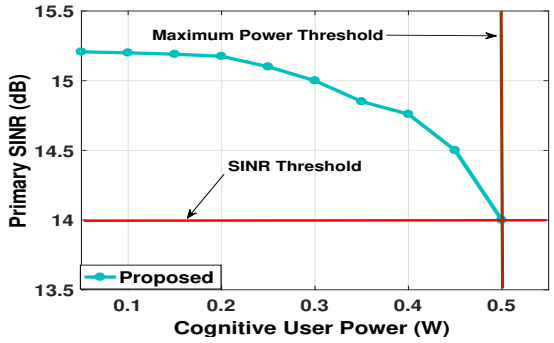



Fig. 5. Average primary system SINR

Fig. 5 presents the average SINR of the primary system as a function of the cognitive user transmission power. The primary SINR is maintained above the threshold. Due to the hardware resources limitation, the figure focus is on the area below the cognitive user power threshold. The reason is that cognitive user transmission power cannot exceed the threshold which is 0.5 W. Otherwise, it will interfere with the primary user transmission.

## V. Conclusion

We presented in this paper an online Q-learning based power allocation scheme that is non-cooperative in an underlay spectrum sharing paradigm, where primary user and cognitive user can share the spectrum with the constraint of minimum interference to primary users. The scheme not only considers power allocation that assures protection to primary users activities, but also addresses QoS issues when cognitive users access the spectrum. In addition, the scheme exploits online Q-learning to conjecture the power allocation of other cognitive users in the system, which brought significant improvement in the achieved performance. The performance of the scheme is demonstrated by both simulation and testbed implementation. The results show that our scheme achieved the maximum SINR, and minimum power consumption compared to other schemes.

## References

[1] S. Haykin, "Cognitive radio: brain-empowered wireless communications", *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, Feb 2005.

[2] J. Mitola and Jr. Maguire, G.Q., "Cognitive radio: making software radios more personal", *Personal Communications, IEEE*, vol. 6, no. 4, pp. 13–18, Aug 1999.

[3] S. w. Han, H. Kim, Y. Han, J. M. Cioffi, and V. C. M. Leung, "A distributed power allocation scheme for sum-rate maximization on cognitive gmacs", *IEEE Transactions on Communications*, vol. 61, no. 1, pp. 248–256, January 2013.

[4] F. E. Lapiccirella, X. Liu, and Z. Ding, "Distributed control of multiple cognitive radio overlay for primary queue stability", *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 112–122, January 2013.

[5] J. Wang, J. Chen, Y. Lu, M. Gerla, and D. Cabric, "Robust power control under location and channel uncertainty in cognitive radio networks", *IEEE Wireless Communications Letters*, vol. 4, no. 2, pp. 113–116, April 2015.

[6] G. Ozcan and M. C. Gursoy, "Optimal power control for underlay cognitive radio systems with arbitrary input distributions", *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4219–4233, Aug 2015.

[7] M. Naeem, K. Illanko, A. Karmokar, A. Anpalagan, and M. Jaseemuddin, "Optimal power allocation for green cognitive radio: fractional programming approach", *IET Communications*, vol. 7, no. 12, pp. 1279–1286, Aug 2013.

[8] M. Liu, T. Song, L. Zhang, and J. Hu, "Subchannel and power allocation for ofdma-based mobile cognitive radio networks", in *2016 8th International Conference on Wireless Communications Signal Processing (WCSP)*, Oct 2016, pp. 1–6.

[9] O. L. A. Lopez, S. M. Sanchez, S. B. Mafra, E. M. G. Fernandez, G. Brante, and R. D. Souza, "Power control and relay selection in cognitive radio ad hoc networks using game theory", *IEEE Systems Journal*, vol. PP, no. 99, pp. 1–12, 2017.

[10] O. van den Biggelaar, J. Dricot, P. De Doncker, and F. Horlin, "Power allocation in cognitive radio networks using distributed machine learning", in *2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications - (PIMRC)*, Sept 2012, pp. 826–831.

[11] Jerzy Martyna, "Power allocation in cognitive radio networks by the reinforcement learning scheme with the help of shapley value of games", in *Internet of Things, Smart Spaces, and Next Generation Networking*, Sergey Andreev, Sergey Balandin, and Yevgeni Koucheryavy, Eds., Berlin, Heidelberg, 2012, pp. 316–327, Springer Berlin Heidelberg.

[12] Richard S. Sutton and Andrew G. Barto, *Introduction to Reinforcement Learning*, MIT Press, Cambridge, MA, USA, 1st edition, 1998.

[13] C. J. C. H. Watkins and P. Dayan, "Q-learning", *Mach. Learn*, vol. 8, no. 3, pp. 279–292, 1992.

[14] Eduardo Rodrigues Gomes and Ryszard Kowalczyk, "Dynamic analysis of multiagent q-learning with $\epsilon$-greedy exploration", in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, ICML '09, pp. 369–376, ACM.

[15] A. D. Tijsma, M. M. Drugan, and M. A. Wiering, "Comparing exploration strategies for q-learning in random stochastic mazes", in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec 2016, pp. 1–8.

[16] Gareth O. Roberts and Jeffrey S. Rosenthal, "Harris recurrence of metropolis-within-gibbs and trans-dimensional markov chains", *Ann. Appl. Probab.*, vol. 16, no. 4, pp. 2123–2139, 11 2006.

[17] L. N. de Castro and F. J. Von Zuben, "Learning and optimization using the clonal selection principle", *Trans. Evol. Comp*, vol. 6, no. 3, pp. 239–251, June 2002.

[18] X. Xie, H. Yang, A. V. Vasilakos, and L. He, "Fair power control using game theory with pricing scheme in cognitive radio networks", *Journal of Communications and Networks*, vol. 16, no. 2, pp. 183–192, April 2014.

[19] Universal Software Radio Peripheral (USRP N210), "https://www.ettus.com/product/details/un210-kit".

[20] Yiyang Pei, Anh Tuan Hoang, and Ying-Chang Liang, "Sensing-throughput tradeoff in cognitive radio networks: How frequently should spectrum sensing be carried out?", in *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, Sept 2007, pp. 1–5.