Opportunistic Spectrum Access with Discrete Feedback in Unknown and Dynamic Environment：A Multi-agent Learning Approach

Zhan Gao[1,2] , Junhong Chen[2], and Yuhua Xu[2]
[1] The State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System（CEMEE），Luoyang，471003，China
[2] PLA University of Science and Technology, China
[e-mail: gzck111@sina.com, junhongchen0526@126.com, yuhuaenator@gmail.com]
*Corresponding author: Yuhua Xu

Received November 5, 2014; revised January 19, 2015; revised May 8, 2015; accepted August 17, 2015; published October 31, 2015

Abstract

This article investigates the problem of opportunistic spectrum access in dynamic environment, in which the signal-to-noise ratio (SNR) is time-varying. Different from existing work on continuous feedback, we consider more practical scenarios in which the transmitter receives an Acknowledgment (ACK) if the received SNR is larger than the required threshold, and otherwise a Non-Acknowledgment (NACK). That is, the feedback is discrete. Several applications with different threshold values are also considered in this work. The channel selection problem is formulated as a non-cooperative game, and subsequently it is proved to be a potential game, which has at least one pure strategy Nash equilibrium. Following this, a multi-agent Q-learning algorithm is proposed to converge to Nash equilibria of the game. Furthermore, opportunistic spectrum access with multiple discrete feedbacks is also investigated. Finally, the simulation results verify that the proposed multi-agent Q-learning algorithm is applicable to both situations with binary feedback and multiple discrete feedbacks.

Keywords: Opportunistic spectrum access, multi-agent learning, distributed channel selection, potential game, and discrete feedback

This research was supported by The Open Project in the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System (CEMEE) No. CEMEE 2015K0203B, and by the National Science Foundation of China under Grant No.61401508.
This paper was presented in part at the 5th International Conference on Game Theory for Networks (GameNets), Beijing, China, Nov., 2014.

1. Introduction

With the rapidly growing demand for wireless spectrum resources, spectrum shortage is becoming quite a serious problem. Some experts have pointed out that the shortage is in fact due to the inefficient spectrum access method [1], rather than the physical scarcity of the spectrum. Often, a portion of the available spectrum is licensed to some specific users for their exclusive usage. However, these licensed users only occupy the spectrum for a certain duration of time, resulting in a low efficiency of spectrum usage. For example, the average spectral occupancy is only 6.2% in urban Auckland, New Zealand in 2007 [2]. This inefficient usage of the spectrum highlights the need to find a highly efficient spectrum access method.

Opportunistic spectrum access (OSA) has become increasingly popular due to its potential to improve the efficiency of spectrum usage [3]-[4]. However, many authors studying the problem of OSA assumed that the wireless environment is static and does not vary with time. This assumption is not realistic since the wireless environment is affected by many factors, such as fading, and hence the spectrum is always time-varying in cognitive radio networks (CRNs).

Based on previous work, some authors began to study the problem of OSA considering a dynamic environment [5]-[6]. However, all this work ignored the following feature of CRNs: that the feedback is not continuous in nature. Influenced by fading and the dynamic nature of the environment, the signal-to-noise ratio (SNR) at the receiver is time-varying. To demodulate the received information correctly, the instantaneous received SNR at the receiver should be larger than a threshold value. Specifically, if the received SNR is larger than the required threshold, the transmitter receives an Acknowledgment (ACK), indicating the successful transmission. Otherwise, it receives a Non-Acknowledgment (NACK), indicating the failed transmission. That is, the feedback is discrete. Considering this discrete feedback, the approaches in the existing work designed for continuous feedback are not applicable to deal with the problem of OSA. Instead a new alorithm considering the dynamic environment and discrete feedback needs to be investigated.

Firstly, the problem of distributed channel selection is formulated as a non-cooperative game, and then it is proved that this game is a potential game which has at least one pure strategy Nash equilibrium. A multi-agent Q-learning algorithm considering the dynamic environment and discrete feedback is then proposed. Users learn to adjust their channel selection strategies according to their received random and discrete feedbacks, and it is also proved that the algorithm can converge to Nash equilibrium (NE) in the unknown and dynamic environment.

To summarize, the main contributions of this article are:

1) We formulate the problem of opportunistic spectrum access with discrete feedback in the dynamic spectrum environment as a non-cooperative game, where the utility function is defined as the expected feedback of each user. In addition, it is also proved that this non-cooperative game is a potential game which has at least one pure strategy Nash equilibrium.
2) We propose a multi-agent Q-learning algorithm where users learn to adjust the channel selection strategy to achieve the pure NE points of the game. Since users only need the current feedback to adjust the channel selection strategies and do not need information about other players, the proposed multi-agent Q-learning algorithm is fully distributed and autonomous. Furthermore, it is also proved that this proposed multi-agent Q-learning algorithm can converge to a Nash equilibria with discrete feedback.

The rest of the article is organized as follows. In Section 2, the related work is discussed. In Section 3, the system model is presented and the problem is formulated. In Section 4, we present the problem of opportunistic spectrum access as a non-cooperative game and investigate the properties of its NE. In Section 5 we propose a multi-agent learning algorithm for achieving the maximum system utility value and prove that the algorithm can achieve NE. The multiple discrete feedbacks are also investigated in this section. In Section 6, simulation results are presented and finally the conclusion is presented in Section 7.

2. Related Work

There are many solutions for OSA including game-theoretic, Markovian decision process, optimal stopping problem and the multi-armed bandit problem [7]. Game theory is really useful in analyzing the mutual interactions among multiple users [8]. It was first used in the economic area, but nowadays, it has been widely used in many other scenarios such as in wireless communication. Previous work has successfully applied game theory to distributed spectrum access in wireless communication systems, and several different game models have been used to solve specific problems e.g., evolutionary game [9], coalition game [10]-[11], static game and repeated game [12] models. However, some of the existing work assumed that the wireless environment is static, in which the channel states remain unchanged during the decision period. In [13], the author formulated the channel selection as an evolutionary game and assumed that the spectrum environment is static. By comparing their own payoffs with the system average payoff, users adjust the channel selection strategies to select a channel with a larger payoff.

In fact, the above assumption is not realistic because the spectrum is always time-varying in cognitive radio networks. Hence, the approach proposed in [13] is not applicable to deal with the problem of OSA in a practical or dynamic environment. In this paper, the channel states are considered to be time-varying. Moreover, a dynamic CRN with multiple interactive users is considered where there is no information exchange among these users, i.e., each user does not require information about the other users' channel selection.

Based on previous work, some authors began to study the problem of distributed channel selection in dynamic environment. In [5], the authors investigated the distributed channel selection problem using a game-theoretic approach for an OSA system. Here, the dynamic means that the channel occupation state is time-varying. There are two channel states: occupied and idle. However, channel fading is not considered and the feedback of each transmission is assumed to be continuous. In this article, channel fading is considered, the received SNR at the receiver is time-varying and the feedback is discrete. The algorithm in [5] was originally designed for continuous feedback, and is not suitable for scenarios with discrete feedbacks. To deal with this problem, we proposed a novel multi-agent Q-learning algorithm in this work.

Some preliminary results on game-theoretic optimization with discrete feedback was reported in our recent work [14]. While only binary feedback (i.e., the feedback is either one or zero) was considered in [14], in this article the problem of OSA with multiple discrete feedbacks is also investigated. Furthermore, a realistic wireless communication system includeing several different kinds of applications, such as data, image, voice and video transmission is also investigated. These different applications have different SNR threshold requirements at the receiver to demodulate information successfully. In other words, the problem of OSA under heterogeneous thresholds is investigated in this article.

Compared with existing work that studies distributed channel selection in OSA systems,

our study has the following characteristics: (i) the wireless environment is dynamic and channels undergo fading, (ii) the feedback is discrete and only when the SNR at the receiver is larger than the threshold value can the user receive a positive feedback and (iii) different applications with different threshold values are considered.

3. System Model and Problem Formulation

3.1 System Model

Let us consider a distributed cognitive radio system that coexists with M licensed channels and N secondary users. Each secondary user competes for one of M channels. Denote the secondary users set as $N_u = \{1,...,N\}$ and the channels set as $M_c = \{1,...,M\}$, and each secondary user link consists of a transmitter and a receiver.

The transmission structure of a secondary user is shown in **Fig. 1** [5]. It is assumed that time is divided into slots (with equal length) and the value of SNR of each channel is block-fixed in a slot and changes randomly in the next slot. Each slot contains a channel selection and sensing, contention, data transmission and learning period. During the contention period, when more than one user chooses the same channel, they share the channel using some multiple access mechanism, e.g., CSMA [15]. It is also assumed that each secondary user has an equal probability of successful access, i.e., if there are n secondary users contending for the same channel, then each secondary user will contend successfully with a probability of $1/n$. Since the transmission structure is the same as that in [5], a detailed description is not provided here.

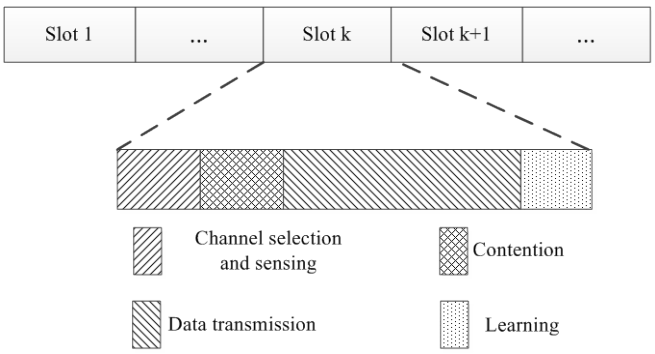


Fig. 1. Transmission structure of the system.

Let us suppose a user can receive a positive feedback only after a successful contention as well as when the received SNR is larger than the threshold value. Compared with previous work presented in [5] which considers continuous feedback, the feedback in this article is discrete. The key difference between the existing work and this article is shown diagrammatically in **Fig. 2**, in which, $a_n(t)$ is the action of user n, $a_{-n}(t)$ represents the actions of the other users except user n and $r_n(t)$ is the feedback of user n. Function $f(a_n(t), a_{-n}(t))$ represents the interaction among users, in another words, this function determines the value of feedback $r_n(t)$ based on $a_n(t)$ and $a_{-n}(t)$.

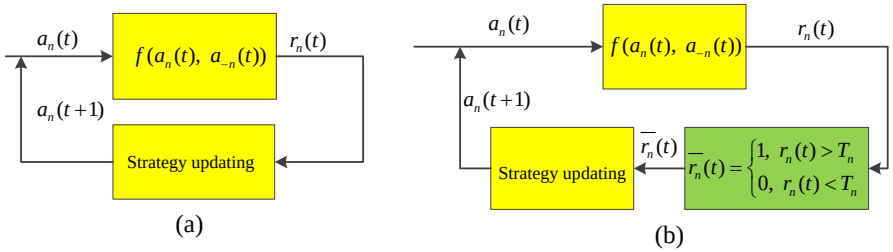


Fig. 2. (a) The illustrative diagram of learning procedure in most existing work; (b) the illustrative diagram of learnig procedure in our work.

3.2 Problem Formulation

In this work, it is assumed that all the channels undergo block fading. Due to channel fading, the SNR at the receiver may be very low and hence the receiver may not receive certain data packets successfully. In this case, the secondary user will get zero payoff. Let $s_m = \{n \in \{1,...,N\} : a_n = m\}$ and $c_m = |s_m|$, i.e., c_m is the number of users choosing channel m and $1/c_m$ is the probability that user n contends for channel m successfully. T_n is the threshold value of user n, i.e., the minimum value of SNR that the receiver n can demodulate information successfully. In the k th slot, the secondary user n chooses channel m and $C_{n,m}(k)$ denotes the instantaneous feedback for the secondary user n:

$$C_{n,m}(k) = \begin{cases} 1, & w.p. \quad \dfrac{1}{c_m} \Pr(\eta_m > T_n) \\ 0, & w.p. \quad 1 - \dfrac{1}{c_m} \Pr(\eta_m > T_n). \end{cases} \tag{1}$$

where η_m refers to the instantaneous received value of SNR when a user transmits on channel m and $\Pr(\eta_m > T_n)$ denotes the probability that the SNR at the receiver on channel m is larger than the threshold value.

4. Game-theoretic Distributed Channel Selection

4.1 Game Model

In this system, there is no central controller and no information exchange among users. Instead, users make their decisions autonomously and in a distributive and interactive manner. All these factors motivate us to formulate the problem of distributed channel selection as a non-cooperative game. The opportunistic spectrum access game is denoted as $\mathcal{G}_c = [N_u, \{A_n\}_{n \in N_u}, \{u_n\}_{n \in N_u}]$, where $N_u = \{1,...,N\}$ is the set of players (secondary users), $A_n = \{1,...,M\}$ is the set of available actions (channels) and u_n is the utility function of player n. The utility function, which is defined as the expected feedback of the secondary user n can be given as

$$u_n(a_n, a_{-n}) = \mathbf{E}[C_{n,m}] = \frac{1}{c_m} \Pr(\eta_{a_n} > T_n), \tag{2}$$

where a_{-n} is the channel selection profile of all the players except player n. Then the proposed channel selection game can be expressed as:

$$(\mathcal{G}_c) \max_{a_n \in A_n} u_n(a_n, a_{-n}). \tag{3}$$

The throughput of the system is defined as the total utility value of all the secondary users:

$$U = \sum_{n=1}^{N} u_n(a_n, a_{-n}). \tag{4}$$

The definition of utility originates from the term 'outage capacity', which is defined as the rate that a reliable transmission can be achieved [16]. In this article, the utility is the probability that a user can achieve reliable transmission. Choosing the expected amount of feedback as the utility can guarantee that a user will choose a channel achieving a higher probability of successful transmission. The outage capacity has been well considered in the literature [17]-[19]. However, the difference in this work is as follows: the optimization of outage capacity is through the users' online learning, while it is through central optimization in those reference articles.

4.2 Analysis of Nash equilibrium (NE)

In this subsection, we first put forward and then analyze the concept of a Nash equilibrium [20], which is the most well-known stable solution for a non-cooperative game model.

Definition 1(NE). A channel selection profile $a^* = (a_1^*, ..., a_N^*)$ is a pure strategy NE if and only if no player can improve their utility by deviating unilaterally, i.e.,

$$u_n(a_n^*, a_{-n}^*) \geq u_n(a_n, a_{-n}^*) \qquad \forall n \in N_u, \forall a_n \in A_n. \tag{5}$$

The properties of the proposed game $\mathcal{G}_c$ are characterized by the following theorem.

Theorem 1. $\mathcal{G}_c$ is an exact potential game which has at least one pure strategy NE point.

Proof: As shown in (2), the number of secondary users selecting each channel m is c_m, $\forall m \in \{1, ..., M\}$. The following potential function Φ for the channel selection game can be defined as:

$$\Phi(a_n, a_{-n}) = \sum_{m=1}^{M} \sum_{i=1}^{c_m} \varphi_m(i), \tag{6}$$

where $\varphi_m(i) = \frac{1}{i} \Pr(\eta_m > T)$. The above function is also known as Rosenthal's potential function [21].

Suppose that an arbitrary player n unilaterally changes its channel selection from a_n to $\bar{a}_n$, then the change in individual utility function caused by this unilateral change is given by:

$$u_n(\bar{a}_n, a_{-n}) - u_n(a_n, a_{-n}) = \varphi_{\bar{a}_n}(c_{\bar{a}_n} + 1) - \varphi_{a_n}(c_{a_n}). \tag{7}$$

In fact, player n's unilateral change only affects the users on the channel a_n and $\bar{a}_n$, with the change in the potential function given by:

$$\begin{aligned} \Phi(\bar{a}_n, a_{-n}) - \Phi(a_n, a_{-n}) &= \left(\sum_{i=1}^{c_{\bar{a}_n}+1} \varphi_{\bar{a}_n}(i) + \sum_{i=1}^{c_{a_n}-1} \varphi_{a_n}(i) \right) - \left(\sum_{i=1}^{c_{\bar{a}_n}} \varphi_{\bar{a}_n}(i) + \sum_{i=1}^{c_{a_n}} \varphi_{a_n}(i) \right) \\ &= \varphi_{\bar{a}_n}(c_{\bar{a}_n} + 1) - \varphi_{a_n}(c_{a_n}). \end{aligned} \tag{8}$$

Based on (7) and (8), we can get :

$$u_n(\bar{a}_n, a_{-n}) - u_n(a_n, a_{-n}) = \Phi(\bar{a}_n, a_{-n}) - \Phi(a_n, a_{-n}). \quad (9)$$

From (9) it is evident that the change in individual utility function caused by the unilateral change is identical to the change in the potential function. According to the definition given in [21], the channel selection game is an exact potential game with a potential function Φ. An exact potential game belongs to a potential game, and every potential game has at least one pure strategy NE point. Therefore, Theorem1 is proved.

5. Reinforcement Learning Solution for Achieving NE in Fading Environment

Using a Q-learning algorithm, the users are only concerned about the result caused by their specific actions and do not require the information of channel selections of other users [22]-[23]. In this section, a multi-agent Q-learning algorithm is proposed to achieve the NE of the formulated opportunistic spectrum access game in the presence of unknown, dynamic and incomplete information constraints.

Following the idea proposed in [24], a Q function is used to replace the utility function as shown in (10). The utility of (2) is the probability that a secondary user contends for the channel successfully as well as the SNR at the receiver is larger than the threshold value, i.e., communicate successfully. If we denote the frequency of successful communication of user n as $X_n(k)$, then we can get the frequency at the $(k+1)$ th slot as (11), where $C_n(k+1)$ is the feedback obtained by user n at the $(k+1)$ th slot. Since the value of $C_n(k+1)$ is either zero or one, the numerator of the equation is the number of successful transmissions, and the denominator is the total number of iterations. When k tends to infinity we can say the frequency is equal to the probability, i.e., $\lim_{k\to\infty} X_n(k) = u_n(a_n(k), a_{-n}(k))$.

$$Q_n(k) = u_n(a_n(k), a_{-n}(k)), \quad (10)$$

$$X_n(k+1) = \frac{kX_n(k) + C_n(k+1)}{k+1}. \quad (11)$$

Then, based on (10) and (11) we can deduce the Q value update function as:

$$Q_n(k+1) = (1 - \frac{1}{k+1})Q_n(k) + \frac{1}{k+1}C_n(k+1) \qquad \forall n \in N. \quad (12)$$

5.1 Algorithm Description

To characterize the proposed multi-agent Q-learning algorithm, the channel selection game $\mathcal{G}_c$ can be extended to a mixed strategy form and give the following definition. Let $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_N)$ denote the mixed strategy profile of the channel selection game. More specifically, $\mathbf{P}_n = (P_{n,1}, P_{n,2}, ..., P_{n,M}), \forall n \in N_u$ is the channel selection probability vector of a secondary user n, where $P_{n,m}$ denotes the probability with which user n selects channel m.

The proposed multi-agent Q-learning algorithm is described in Algorithm 1. The stop criterion can be one of the following: 1) the maximum iteration number is reached, 2) for each player n, where $\forall n \in N_u$, there is a component of the channel selection probability sufficiently approaching one (e.g. 0.99).

Algorithm 1. *Multi-agent Q-learning Algorithm for Distributed Channel Selection with Binary Feedback.*

Initialization: Set the iteration index $k=0$ and the initial channel selection probability vector $P_{n,m}(k)=1/M$, $Q_{n,m}(k)=0$, $\forall n \in N_u, \forall m \in \{1,...,M\}$.

Loop for $k=0,1,...,$

Channel access and get random feedback: At the beginning of the k th slot, each secondary user n selects a channel $a_n(k)$ according to its current channel selection probability vector $P_n(k)$. Then, the secondary user performs channel sensing and channel contention. At the end of the k th slot, each secondary user n receives the random feedback $C_{n,m}(k)$ specified by (1).

Update Q-value: All the secondary users update their Q-value according to the following rules:

$$Q_{n,m}(k+1)=(1-\frac{1}{k+1})Q_{n,m}(k)+\frac{1}{k+1}C_{n,m}(k+1), m=a_n(k+1) \tag{13}$$

$$Q_{n,m}(k+1)=Q_{n,m}(k),\ m \neq a_n(k+1) \tag{14}$$

Update channel selection probability: All the secondary users update their channel selection probability vectors according to the following rule:

$$P_{n,m}(k+1)=\frac{e^{Q_{n,m}(k)/\gamma}}{\sum_{m=1}^{M} e^{Q_{n,m}(k)/\gamma}}, \tag{15}$$

where γ is called temperature and controls the frequency of exploration. The smaller γ is, the more focused the actions are. Consequently, when $\gamma \to 0$, each secondary user tend to select the channel with the largest Q-value.

End loop

The equations (13)-(15) show that users can update their channel selection strategies according to their own feedback, i.e., they can choose their channel independently and there is no information exchange among users.

5.2 Convergence of Q-learning

The Q-values for different users are mutually coupled and all Q-values change if one Q-value is changed. Based on (2), (10) and (15), we have Q-values as follows:

$$Q_{n,m}=\Pr(\eta_m > T_n)\prod_{l\neq n}\Pr(a_l(k)\neq m)=\Pr(\eta_m > T_n)\prod_{l\neq n}(1-\frac{e^{Q_{n,m}(k)/\gamma}}{\sum_{m=1}^{M} e^{Q_{n,m}(k)/\gamma}}), \tag{16}$$

where $\Pr(\eta_m > T_n)$ is the probability that the SNR is larger than the threshold value of user n. We define the Q-values satisfying (16) as stationary points

Following the idea in [12], we can get the following theorem.

Theorem 2. The proposed multi-agent Q-learning algorithm converges to Nash equilibria with a probability of one.

Proof: First let us define the following equation:

$$\mathbf{q}=(Q_{1,1},...,Q_{1,M},Q_{2,1},...,Q_{2,M},...,Q_{N,1},...,Q_{N,M})^T. \tag{17}$$

Then (16) can be rewritten as:

$$g(\mathbf{q}) = \mathbf{A}(\mathbf{q}) - \mathbf{q} = 0, \tag{18}$$

$$\mathbf{A}_{n,m} = \Pr(\eta_m > T_n) \prod_{l \neq n} (1 - \frac{e^{Q_{n,m}(k)/\gamma}}{\sum_{m=1}^{M} e^{Q_{n,m}(k)/\gamma}}), \tag{19}$$

where $\mathbf{A}_{n,m}$ is the probability that user n achieves successful communication (including the two aspects considered in this article). Futhermore, $\mathbf{A}_{n,m}$ is also the expected feedback of user n since the instantaneous feedback is either one or zero.

Then, the updating rule in (13) is equivalent to solving equation (16) using the Robbins-Monro algorithm [25], i.e.,

$$\mathbf{q}(k+1) = (1 - \frac{1}{k+1})\mathbf{q}(k) + \frac{1}{k+1}\mathbf{c}(k+1) = \mathbf{q}(k) + \frac{1}{k+1}\mathbf{Y}(k), \tag{20}$$

where $\mathbf{c}(k+1)$ is the vector of feedback and $\mathbf{Y}(k)$ satisfies the following equation:

$$\begin{aligned} \mathbf{Y}(k) &= \mathbf{c}(k+1) - \mathbf{q}(k) = \bar{\mathbf{c}}(k+1) - \mathbf{q}(k) + \mathbf{c}(k+1) - \bar{\mathbf{c}}(k+1) \\ &= g(\mathbf{q}(k)) + \delta\mathbf{m}(k), \end{aligned} \tag{21}$$

where $g(\mathbf{q}(k)) = \bar{\mathbf{c}}(k+1) - \mathbf{q}(k), \delta\mathbf{m}(k) = \mathbf{c}(k+1) - \bar{\mathbf{c}}(k+1)$ is noise, and $\bar{\mathbf{c}}(k+1) = \mathbf{A}(\mathbf{q}(k))$. Obviously, $\mathrm{E}[\delta\mathbf{m}(k)]=0$ as the expectation of the difference between the feedback and the expected feedback is equal to zero. Therefore, the observation $\delta\mathbf{m}(k)$ is a Martingale difference.

The procedure of Robbins-Monro algorithm (i.e. the updating of the Q-value) is the stochastic approximation of the solution for the equation. It is well known that the convergence of such a procedure can be characterized by an ordinary differential equation (ODE).

According to Lemma 2 in [12], we know that with probability one, the sequence $\mathbf{q}(k)$ converges to some limit sets of the ODE:

$$\mathbf{q} = g(\mathbf{q}). \tag{22}$$

Applying Lyapunov function, the solution of the ODE (22) converges to the stationary point determined by (18). Combining Lemmas 1, 2 and 3 from [12], it can be proved that the proposed multi-agent Q-learning algorithm converges to a stationary point with a probability of one. Therefore, Theorem 2 is proved. □

5.3 Dynamic Spectrum Access with Multiple Discrete Feedbacks

The above sections all consider that the feedback is binary, i.e., the feedback is either one or zero. In this subsection, we consider a more realistic wireless environment in which there are multiple feedbacks. The difference between binary feedback and multiple discrete feedbacks is shown in **Fig. 3**, where r_i represents the instantaneous feedback, and $T_i\,(i = 1, 2, 3)$ is the different threshold values.

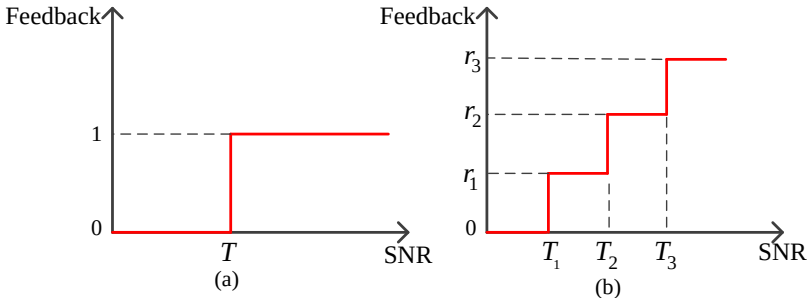


Fig. 3. (a) The sketch map of binary feedback; (b) the sketch map of multiple discrete feedbacks.

Due to channel fading, the transmission rate of each channel is always time-varying. With the help of adaptive modulation and coding, the channel transmission rate is classified into several states according to the instantaneous received SNR. The rate set of channel m is denoted as $S_m = \{s_{m,1}, s_{m,2}, ...s_{m,L}\}$. Practically, the channel rate set S_m can be obtained by the following procedure. First, partition the entire SNR region into L non-overlapping consecutive intervals with boundary points $\{T_i\}_0^{L-1}$. Here, we apply the SNR region partitioning scheme whose objective is to maintain a predefined packet error rate. And then $s_{m,l}$ is choosen if $\eta_m \in [T_{l-1}, T_l)$, where η_m is the instantaneous received SNR of channel m. In this subsection, we define the feedback of user n choosing channel m as:

$$C_{n,m} = \begin{cases} s_{m,l}, & w.p.\ \dfrac{1}{c_m} \\ 0, & w.p.\ 1-\dfrac{1}{c_m}, \end{cases} \tag{23}$$

where $s_{m,l}$ is determined by comparing the instantaneous received SNR with different threshold values and c_m is the total number of users selecting channel m.

Similar to binary feedback, we can define the utility as the expected feedback, which originates from the term outage capacity. The corresponding representation of utility is the same as that of binary feedback and hence is not repeated here.

The interaction among users is formulated as a non-cooperative game, which has the same property as shown in Theorem 1. The learning approach with multiple discrete feedbacks is similar to Algorithm 1, the only difference is that the feedback in (13) is multiple rather than binary. Since the game model and the learning algorithm is the same as the case for dynamic spectrum access with binary feedback, it is not described again in this subsection.

6. Simulation Results and Discussion

In this section, we investigate the convergence and throughput performance of the proposed multi-agent Q-learning algorithm with binary feedback and multiple discrete feedbacks. Since channels undergo fading, we consider the SNR at each time slot is a random value from 5 to 10 dB. As shown in (15), the value of γ reflects the frequency of exploration. After a number of iterations, to ensure that the system reaches a stationary point, the value of γ should tend to

zero. In the simulation, we set the value of γ as $1/k$, where k is the total number of iterations.

6.1 Simulation Results under Binary Feedback

A. Convergence Behavior

1) Convergence Behavior under Homogeneous Threshold

For this case, there are five secondary users and three channels in the system, and the threshold value of all the users is 9 dB. For an arbitrarily chosen user, the evolution of channel selection probabilities is shown in **Fig. 4**. Through this simulation result, it is evident that the channel selection probabilities converge to a pure strategy ({0,0,1}) in about 40 iterations.

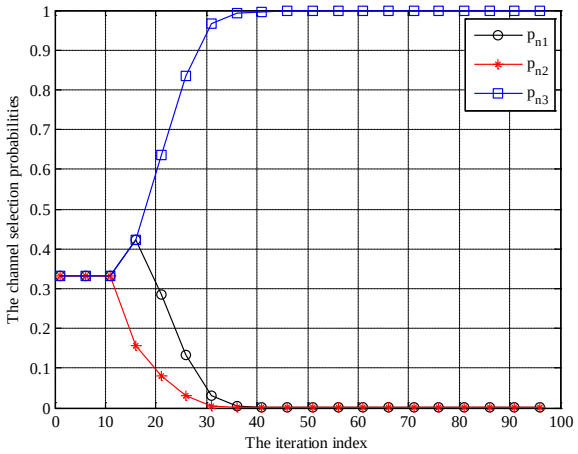


Fig. 4. Evolution of the channel selection probability of an arbitrary secondary user under homogeneous threshold.

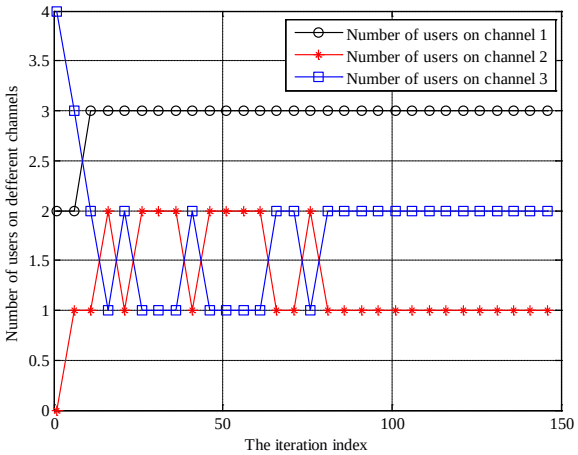


Fig. 5. Evolution of the number of secondary users selecting each channel under homogeneous threshold.

Moreover, the evolution of the number of secondary users selecting each channel is shown in **Fig. 5**. There are six users and three available channels in the system, and the threshold value of all the users is 9 dB. It is noted that after convergence, $S_1^* = 3$, $S_2^* = 1$, $S_3^* = 2$ (S_i is

the number of secondary users choosing channel i), and in other words the system achieves the NE.

The convergence speed versus different threshold values of SNR is studied in **Fig. 6.** The results are obtained by taking 10000 independent trials and then taking the corresponding expectation. It is noted that as the threshold value increases, the greater iteration times needed to achieve convergence. The reason is as the threshold value increases, the SNR at the receiver is more likely to be lower than the threshold, and hence the user cannot receive positive feedback to select a better channel in the next slot. Thus more time is needed to achieve convergence.

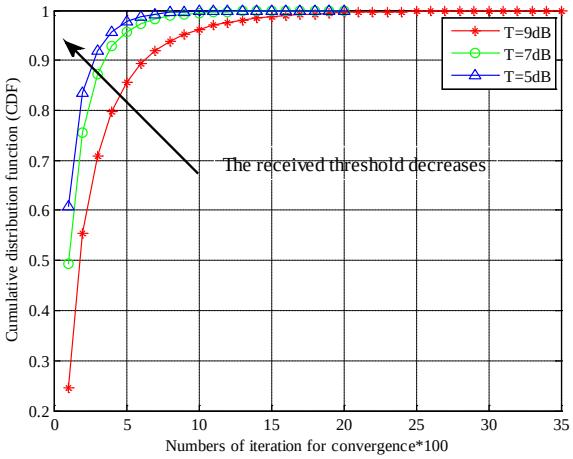


Fig. 6. The convergence behavior versus different threshold value (N =5, M =3).

2) Convergence Behavior under Heterogeneous Thresholds

In the following subsection, we will consider different applications existing in the CRNs supposing there are five users and three available channels. The threshold value of each user is 5,7,9,10 and 12 *dB* respectively. For an arbitrarily chosen user, the evolution of channel selection probabilities is shown in **Fig. 7**. The simulation result shows that the channel selection probabilities converge to a pure strategy ({1,0,0}) in about 30 iterations.

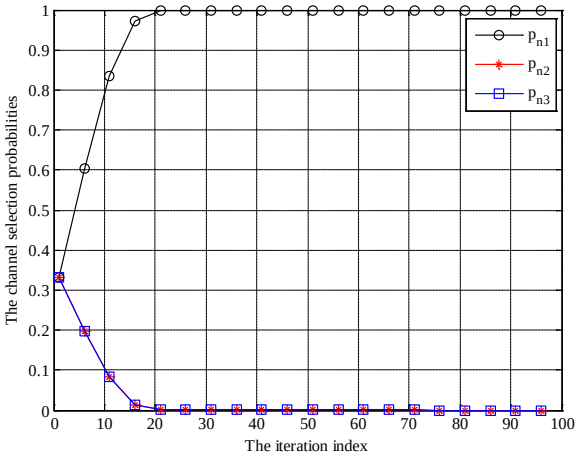


Fig. 7. Evolution of the channel selection probability of an arbitrarily secondary user under heterogeneous threshold.

The evolution of the number of secondary users selecting each channel is shown in **Fig. 8**. There are six users and three available channels in the system, and the threshold value of each user is 5,7,9,10,12 and $8\,dB$. It is noted that after convergence, $S_1^* = 1$, $S_2^* = 2$, $S_3^* = 3$, the system achieves the NE.

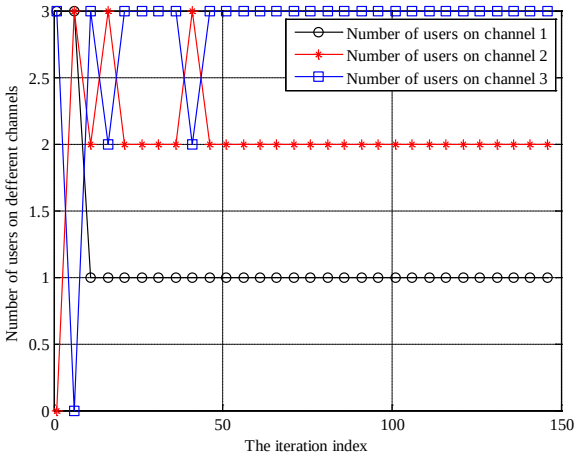


Fig. 8. Evolution of the number of secondary users selecting each channel under heterogeneous threshold.

The simulation results **Fig. 4**-**Fig. 8** show that the proposed multi-agent Q-learning algorithm is applicable to the following two scenarios: 1) users in the CRNs have the same threshold value, 2) there are different applications in the system and different users have different SNR threshold values.

B. Throughput Performance

The throughput performance of the system versus the number of users is shown in **Fig. 9** assuming there are five available channels. The number of users is increased from four to fifteen and the threshold value of all the users is 9 dB. The results are obtained by taking 100000 independent trials and then taking the expectation. Furthermore, the throughput performance of the proposed multi-agent Q-learning algorithm and random selection method are also compared in the figure.

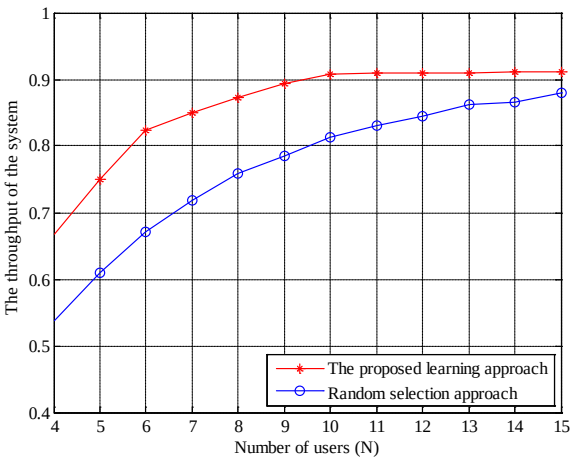


Fig. 9. The throughput performance of the system for different number of users.

According to the figure, two conclusions can be drawn: (i) the achievable performance of both approaches increase rapidly as N increases when the number of users is small, while it becomes moderate when the number of users is large. (ii) the proposed multi-agent Q-learning algorithm outperforms the random selection method, and what is more, the performance gap between the proposed multi-agent Q-learning algorithm and the random selection approach decreases as the number of users increases.

The reasons are as follows: 1) the access opportunities are abundant when the number of users is small, which means that adding a user to the system leads to relatively significant performance improvement. While the access opportunities are saturated when the number of users is large and consequently, the performance improvement decreases. 2) when the proposed multi-agent Q-learning algorithm converges to a pure strategy, users are spread over all the channels. However, for the random selection method, some channels may be crowded while other channels may be not occupied by any users as users select channels randomly. 3) when the number of users becomes sufficiently large, the users are uniformly spread over the channels and hence the performance gap between the multi-agent Q-learning algorithm and random selection method is negligible.

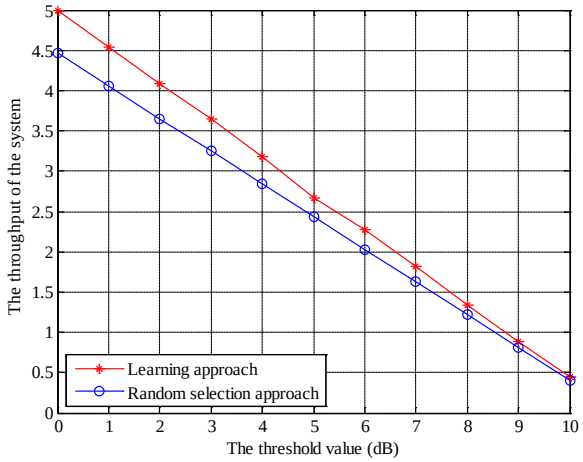


Fig. 10. The throughput performance of the system under different threshold value.

In **Fig. 10**, we compare the throughput performance of different channel selection approaches under different threshold values. It is assumed that there are five available channels and ten users in the system. The results are obtained by taking 100000 independent trials and then finding the expectation. Through the simulation results, it can be concluded that the proposed learning method achieves better performance when compared to the random selection approach. In addition, the throughput of the system decreases as the threshold value increases. As the SNR threshold value increases, it is more likely that the received SNR is lower than the threshold, which will lead to zero feedback and consequently the throughput will decrease.

6.2 Simulation Results under Multiple Discrete Feedbacks

A. Convergence Behavior

With the help of adaptive modulation and coding, the channel transmission rate is classified into several states according to the instantaneous received SNR. The state classification is determined by the average received SNR and the target packet error rate. Applying

HIPERLAN/2 standard [26], the channel rate set is given by $S_m = \{0,1,2,3,6\}$ when the average received SNR is 5 dB and the packet error rate is 10^{-3}. The rate is defined as the transmitted packets in a slot, and the threshold value is 1.303 dB and 2.687 dB, 5.496 dB, 26.890 dB.

We assume there are six users and three available channels in the CRNs. The evolution of the number of secondary users selecting each channel is shown in **Fig. 11**. After convergence, $S_1^* = 1$, $S_2^* = 3$ and $S_3^* = 2$. Through the simulation result, it can be proved that the proposed multi-agent Q-learning algorithm is also applicable to dynamic spectrum access with multiple discrete feedbacks.

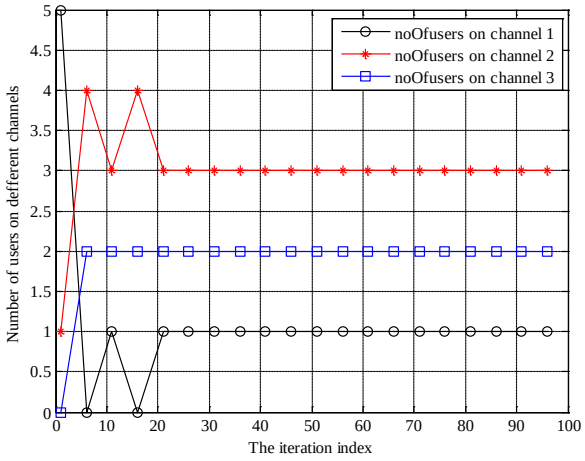


Fig. 11. Evolution of the number of secondary users selecting each channel under multipl discrete feedbacks.

B. Throughput Performance

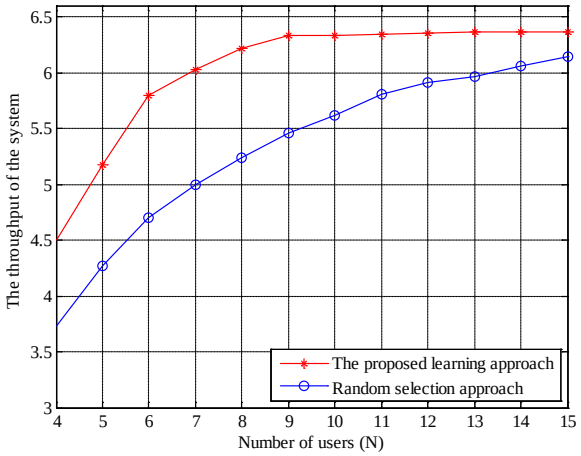


Fig. 12. The throughput performance of the system for different number of users under multipl discrete feedbacks.

The throughput performance versus the number of users is presented in **Fig. 12**. It is assumed that there are 5 available channels and the number of users increases from four to fifteen. The average received SNR is 5 dB and the packet error rate is 10^{-3}. The results are obtained by

taking 50000 independent trials and then finding the expectation. The simulation result shows that the proposed multi-agent Q-learning algorithm achieves better throughput performance compared with the random selection. Hence it is possible to arrive at the same conclusions as those drawn from the result in **Fig. 9**.

In **Fig. 13**, the throughput performance of the two different channel selection approaches under different average SNR is compared. It is assumed that there are five available channels and ten users. The error packet is 10^{-3} and the average received SNR increases from 5 to 15 dB. The simulation result shows that the proposed multi-agent Q-learning algorithm achieves better throughput performance. In addition, the average utility increases as the average received SNR increases, since a higher received SNR means higher feedback.

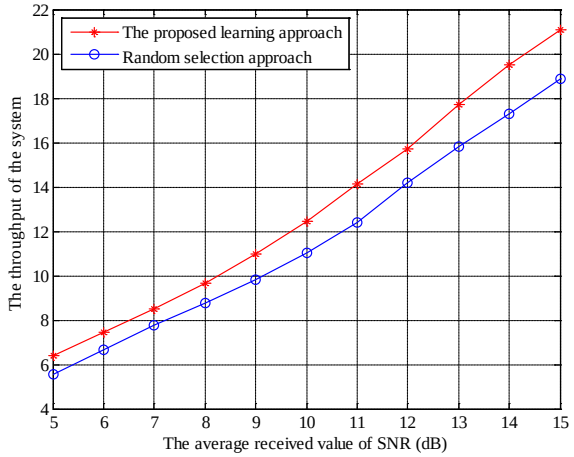


Fig. 13. The throughput performance of the system for different average received SNR.

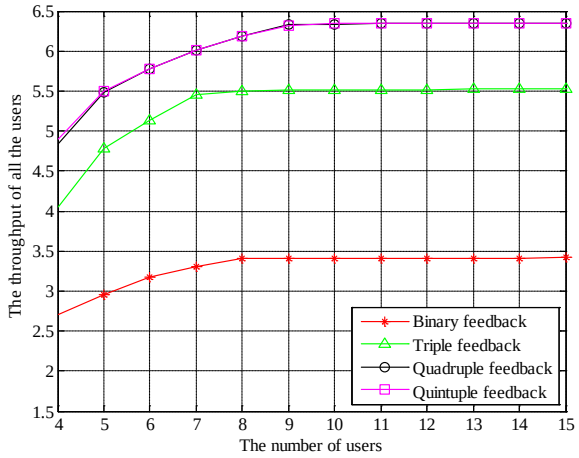


Fig. 14. The throughput performance under binary feedback and multiple discrete feedbacks.

Furthermore, we can compare the throughput performance among the following cases in **Fig. 14**: binary feedback, triple feedback, quadruple feedback and quintuple feedback. The threshold value of the four cases is ($1.203\,dB$), ($1.303\,dB$ and $2.587\,dB$), ($1.303\,dB$ and $2.587\,dB$, $5.496\,dB$) and ($1.303\,dB$ and $2.687\,dB$, $5.496\,dB$, $26.890\,dB$) respectively. The threshold value is determined by the average received SNR and the target packet error rate.

The simulation result illustrates that the throughput performance under multiple discrete feedbacks is higher than that under binary feedback. Moreover, the throughput performance under quadruple feedback is the same as that under quintuple feedback. The reason is that the probability that the feedback equal to six is very low under the quintuple feedback as the threshold value is 26.890 dB. Consequently, the feedback under the quintuple feedback is the same as that under quadruple feedback.

6.3 Comparison with Existing Schemes with Continuous Feedback

In this subsection, we compare our proposed multi-agent Q-learning algorithm and the algorithm proposed in [5] which was originally designed for distributed channel selection with continuous feedback.

We found that the gap of the throughput performance between our proposed multi-agent Q-learning algorithm and approach proposed in [5] is small. However, the proposed multi-agent Q-learning algorithm outperforms the approach in [5] in terms of convergence speed. **Fig. 15** and **Fig. 16** show the comparison of cumulative distribution function (CDF) between the two algorithms under binary feedback and quintuple discrete feedback respectively. The results are obtained by taking 10000 independent trials.

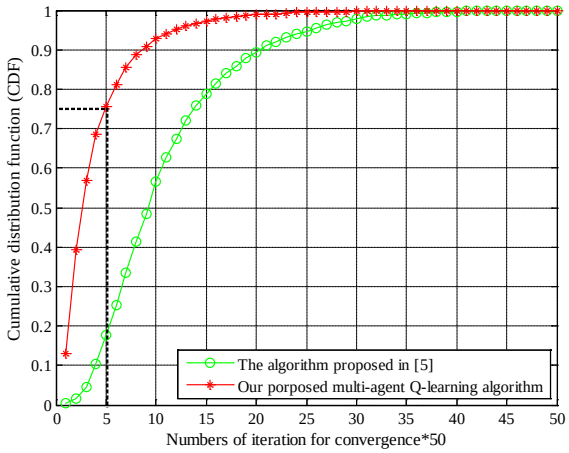


Fig. 15. The comparison of convergence speed under binary feedback.

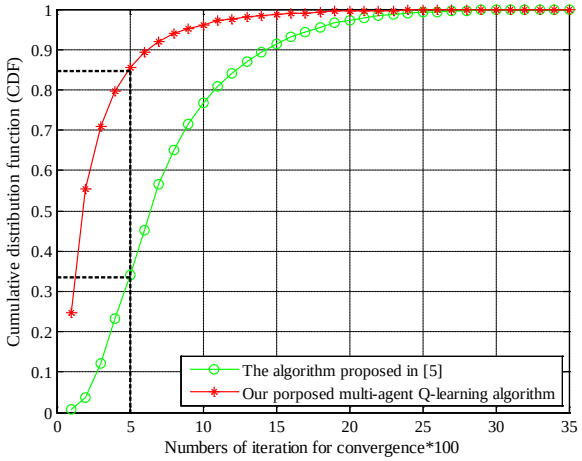


Fig. 16. The comparison of convergence speed under quintuple discrete feedback.

In **Fig. 15**, when the number of iterations is 250 about 75% of trials achieve convergence in our proposed multi-agent Q-learning algorithm, while only 19% of trials achieve convergence using the approach outlined in [5]. In **Fig. 16**, when the number of iteration is 500, about 85% of trials achieve convergence in our proposed multi-agent Q-learning algorithm, while only 35% of trials achieve convergence using the approach given in [5]. Comparing **Fig. 15** and **Fig. 16**, for both binary feedback and quintuple discrete feedback, we can conclude that the convergence speed slows down for both algorithms, while the decrease trend of approach in [5] is more apparent than our algorithm. In other words, the proposed multi-agent Q-learning algorithm is applicable to both binary feedback and multiple discrete feedbacks.

From the simulation results given in **Fig. 15** and **Fig. 16**, while the algorithm presented in [5] can achieve convergence, there is no theoretic proof of the convergence for this algorithm based on discrete feedback. However in this article, rigorous proof of the convergence is given in Theorem 2. Furthermore, the Q-value is updated based on the discrete feedback which accelerates the speed of convergence.

Some previous work has also investigated the problem of OSA using Q-learning algorithm, e.g., [6]. While, their work formulated OSA as a finite Markov decision process (MDP), whose Q-value update function is related to the state set and transition function. Since the MDP does not correspond to our work, it is not analyzed in this subsection.

7. Conclusion

In this article, distributed channel selection in opportunistic spectrum access systems with discrete feedback was investigated. In addition, this work also considered a realistic scenario with different thresholds. The interactions among the users in the time-varying environment was formulated as a non-cooperative game which was later proved to be a potential game. Then a multi-agent Q-learning algorithm was proposed in which users learn to adjust their channel selection strategies according to the instantaneous feedbacks. It was also proved that the proposed multi-agent Q-learning algorithm can converge to a Nash equilibrium with discrete feedback. Based on the binary feedback, multiple discrete feedbacks were also investigated considering both adaptive modulation and coding. The simulation results verified that users can adjust their channel selection strategies to pure strategy Nash equilibria according to the proposed multi-agent Q-learning algorithm with both binary feedback and multiple discrete feedbacks. Future work focusing on the theoretical analysis of the learning algorithm considering multiple discrete feedbacks is on-going.

References

[1] S. Haykin, “Cognitive radio: Brain-empowered wireless communications,” *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201-220, February, 2005. Article (CrossRef Link)

[2] R. I. C. Chiang, G. B. Rowe and K. W. Sowwerby, “A quantitative analysis of spectral occupancy measurements for cognitive radio,” in *Proc. of IEEE Conf. on Vehicular Technology*, pp. 3016-3020, April 22-25, 2007. Article (CrossRef Link)

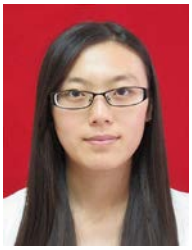
[3] X. Chen, T. Chen, W. Cheng and H. Zhang, “Reciprocity inspired learning for opportunistic spectrum access in cognitive radio networks,” in *Proc. of 2013 8th International Conference on Cognitive Radio Oriented Wireless Networks*, pp. 202-207, July 8-10, 2013. Article (CrossRef Link)

[4] M. Derakhshani and T. Le-Ngoc, “Learning-based opportunistic spectrum access with adaptive hopping transmission strategy,” *IEEE Transactions on Wireless Communications*, vol. 11, no. 11, pp. 3957-3967, September, 2012. Article (CrossRef Link)

[5] Y. Xu, J. Wang, Q. Wu, A. Anpalagan and Y. Yao, "Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution," *IEEE Transaction on Wireless Communication*, vol. 11, no. 4, pp. 1380-1391, February, 2012. Article (CrossRef Link)
[6] P. Venkatraman, B. Hamdaoui and M. Guizani, "Opportunistic bandwidth sharing through reinforcement learning," *IEEE Transaction on Vehicular Technology*, vol. 59, no. 6, pp. 3148-3153, April, 2010. Article (CrossRef Link)
[7] Y. Xu, A. Anpalagan, Q. Wu, L. Shen, Z. Gao and J. Wang, " Decision-theoretic distributed channel selection for opportunistic spectrum access: Strategies, challenges and solutions," *IEEE Communication Surveys & Tutorials*, vol. 15, no. 4, pp. 1689-1713, April, 2013. Article (CrossRef Link)
[8] R. Myerson, *Game theory: Analysis of conflict*, Cambridge, MA: Harvard Univ. Press, 1991. Article (CrossRef Link)
[9] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 2008-2017, August, 2009. Article (CrossRef Link)
[10] W. Saad, Z. Han, T. Basar, M. Debbah and A. Hjorungnes, "Coalition formation games for collaborative spectrum sensing," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 1, pp. 276-297, October, 2011. Article (CrossRef Link)
[11] W. Saad, Z. Han, R. Zheng, and A. Hjorungnes, "Coalitional games in partition form for joint spectrum sensing and access in cognitive radio networks," *IEEE Journal of Topics in Signal Processing*, vol. 6, no. 2, pp. 195-209, November, 2012. Article (CrossRef Link)
[12] H.Li, "Multi-agent q-learning for competitive spectrum access in cognitive radio systems," in *Proc. of IEEE Fifth Workshop on Networking Technologies for Software Defined Radio Networks*, pp. 1-6, June 21, 2010. Article (CrossRef Link)
[13] X. Chen and J. Huang, "Evolutionary stable spectrum access," *IEEE Transaction on Mobile Computing*, vol. 12, no. 7, pp. 1281-1293, 2013. Article (CrossRef Link)
[14] J. Chen, Z. Gao and Y. Xu, "Opportunistic spectrum access with limited feedback in unknown dynamic environment: a multi-agent learning method," in *Proc. of 2014 5th International Conference on Game Theory for Networks*, pp. 1-6, November 25-27, 2014. Article (CrossRef Link)
[15] Q. Zhao, L. Tong, A. Swami, et al., "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589-600, April, 2007. Article (CrossRef Link)
[16] G. Levin and S. Loyka, "On the outage capacity distribution of correlated keyhole MIMO channels," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3232-3245, July, 2008. Article (CrossRef Link)
[17] Y. Ko and K. Moessner, "Maximum outage capacity in dense indoor femtocell networks with joint energy and spectrum utilization," *IEEE Transactions on Wireless Communications*, vol. 11, no. 12, pp. 4416-4425, December, 2012. Article (CrossRef Link)
[18] X. Qu, X. Xu, H. Li and X. Tao, "Analysis of outage capacity for DF dual-hop relay and optimal power allocation," in *Proc. of 2010. IET-WSN. IET International Conference on Wireless Sensor Network*, pp. 194-197, November, 2010. Article (CrossRef Link)
[19] T. Samarasinghe, H. Inaltekin and J. S. Evans, "On the outage capacity of opportunistic beamforming with random user locations," *IEEE Transactions on Communications*, vol. 62, no. 8, pp. 3015-3026, August, 2014. Article (CrossRef Link)
[20] D. Monderer and L. S. Shapley, "Potential games," *games and economic behavior*, vol. 14, pp. 124-143, 1996. Article (CrossRef Link)
[21] B. Vcking and R. Aachen, "Congestion games: Optimization in competition," in *Proc. of 2006 Algorithms and Complexity in Durham Workshop*, pp. 9-20, 2006. Article (CrossRef Link)
[22] A. Galindo-Serrano and L. Giupponi, "Distributed q-learning for aggregated interference control in cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1823-1834, February, 2010. Article (CrossRef Link)

[23] P. Venkatraman, B. Hamdaoui and M. Guizani, "Opportunistic bandwidth sharing through reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 6, pp. 3148-3153, April, 2010. Article (CrossRef Link)
[24] H.Tembine, *Distributed strategic learning for wireless engineers, CRC Press*, 2012. Article (CrossRef Link)
[25] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, 2003. Article (CrossRef Link)
[26] A. Doufexi, S. Armour, M. Butler, el al., "A comparison of the HIPERLAN/2 and IEEE 802.11 a wireless LAN standards," *IEEE Communications Magazine*, vol. 40, no. 5, pp. 172-180, 2002. Article (CrossRef Link)

Zhan Gao received his B.S. degree in Communications Engineering, M.S. and Ph.D. degrees in Communications and Information System from the Institute of Communications Engineering, Nanjing, China, in 1999, 2001 and 2004, respectively. He is currently an associate professor at the PLA University of Science and Technology, China. His current research interests are cognitive radio networks, distributed optimization algorithms and digital signal processing.



Junhong Chen received her B.S. degree in Communications Engineering from Hohai University, Nanjing, China, in 2013. Now she is studying for the M.S. degree in Communications and Information System from College of Communications Engineering, PLA University of Science and Technology. Her research interests focus on cognitive radio, opportunistic spectrum access, leaning, and game theory.

Yuhua Xu received his B.S. degree in Communications Engineering, and Ph.D. degree in Communications and Information Systems from College of Communications Engineering, PLA University of Science and Technology, in 2006 and 2014 respectively. He has been with College of Communications Engineering, PLA University of Science and Technology since 2012, and currently as an Assistant Professor. His research interests focus on opportunistic spectrum access, learning theory, game theory, and distributed optimization techniques for wireless communications. He has published several papers in international conferences and reputed journals in his research area. He is an Editor for the *KSII Transactions on Internet and Information Systems*. In 2011 and 2012, he was awarded Certificate of Appreciation as Exemplary Reviewer for the IEEE Communications Letters.