Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2010, Article ID 876216, 15 pages
doi:10.1155/2010/876216

Research Article

Multiagent Q-Learning for Aloha-Like Spectrum Access in Cognitive Radio Systems

Husheng Li

Department of Electrical Engineering and Computer Science, The University of Tennessee, Knoxville, TN 37996, USA

Correspondence should be addressed to Husheng Li, husheng@eecs.utk.edu

Received 31 December 2009; Accepted 18 April 2010

Academic Editor: Vincent Lau

An Aloha-like spectrum access scheme without negotiation is considered for multiuser and multichannel cognitive radio systems. To avoid collisions incurred by the lack of coordination, each secondary user learns how to select channels according to its experience. Multiagent reinforcement leaning (MARL) is applied for the secondary users to learn good strategies of channel selection. Specifically, the framework of Q-learning is extended from single user case to multiagent case by considering other secondary users as a part of the environment. The dynamics of the Q-learning are illustrated using a Metrick-Polak plot, which shows the traces of Q-values in the two-user case. For both complete and partial observation cases, rigorous proofs of the convergence of multiagent Q-learning without communications, under certain conditions, are provided using the Robins-Monro algorithm and contraction mapping, respectively. The learning performance (speed and gain in utility) is evaluated by numerical simulations.

1. Introduction

In recent years, cognitive radio has attracted intensive studies in the community of wireless communications. It allows users without license (called secondary users) to access licensed frequency bands when licensed users (called primary users) are not present. Therefore, the cognitive radio technique can substantially alleviate the problem of underutilization of frequency spectrum [1, 2].

The following two problems are key to cognitive radio systems.

(i) Resource mining, that is, how to detect the available resource (the frequency bands that are not being used by primary users); usually it is done by spectrum sensing.

(ii) Resource allocation, that is, how to allocate the available resource to different secondary users.

Substantial work has been done for the resource mining problem. Many signal processing techniques have been applied to sense the frequency spectrum [3], for example, cyclostationary feature [4], quickest change detection [5], and collaborative spectrum sensing [6]. Meanwhile, a significant amount of research has been conducted for the resource allocation in cognitive radio systems [7, 8]. Typically, it is assumed that secondary users exchange information about available spectrum resources and then negotiate on the resource allocation according to their own requirements of traffic (since the same resource cannot be shared by different secondary users if orthogonal transmission is assumed). These studies typically apply theories in economics, for example, game theory, bargaining theory, or microeconomics.

However, in many applications of cognitive radio, such a negotiation-based resource allocation may incur significant overhead. In traditional wireless communication systems, the available resource is almost fixed (even if we consider the fluctuation of channel quality incurred by fading, the change of available resource is usually very slow and thus can be considered stationary). Therefore, the negotiation need not be carried out frequently, and the negotiation result can be applied for a long period of data communication, thus incurring only tolerable overhead. However, in many cognitive radio systems, the resource may change very rapidly since the activity of primary users may be highly dynamic.

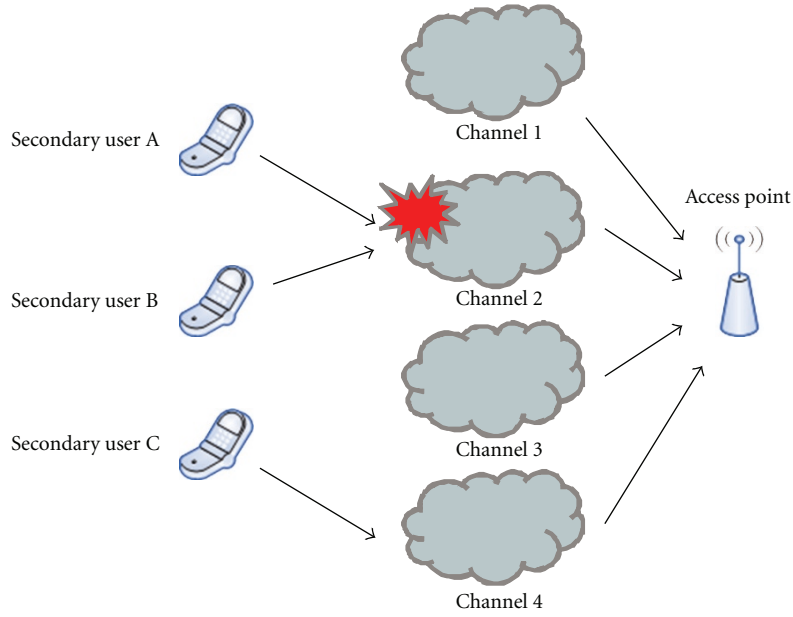

Figure 1: Illustration of competition and conflict in multiuser and multichannel cognitive radio systems.

Therefore, the available resource needs to be updated very frequently, and the data communication period between two spectrum sensing periods should be fairly short since minimum violation to primary users should be guaranteed. In such a situation, the negotiation of resource allocation may be highly inefficient since a substantial portion of time needs to be used for the negotiation. To alleviate such an inefficiency, high-speed transceivers need to be used to minimize the time consumed on negotiation. Particularly, the turn-around time that is, the time needed to switch from receiving (transmitting) to transmitting (receiving) should be very small, which is a substantial challenge to hardware design.

Motivated by the above discussion and observation, in this paper, we study the problem of spectrum access without negotiation in multiuser and multichannel cognitive radio systems. The spectrum access without negotiation is achieved by applying the framework of reinforcement learning. In such a scheme, each secondary user senses channels and then chooses an idle frequency channel to transmit data, as if no other secondary user exists. If two secondary users choose the same channel for data transmission, they will collide with each other and the corresponding data packets cannot be decoded. Such a procedure is illustrated in Figure 1, where three secondary users access an access point via four channels. Note that such a scheme is similar to Aloha [9] where no explicit collision avoidance is applied. We can also apply techniques similar to p-persistent Carrier Sensing Multiple Access (CSMA) that is, each secondary user transmits with probability p when it finds an available channel. However, it is beyond the scope of this paper. In the Aloha-like approach, since there is no mutual communication among these secondary users, collision is unavoidable. However, the secondary users can try to learn collision avoidance, as well as channel qualities (we assume that the secondary users have no *a priori* information about the channel qualities), according to their experience. In such a context, the learning procedure includes not only the available frequency spectrum but also the behavior of other secondary users.

To accomplish the learning of Aloha-like spectrum access, multiagent reinforcement learning (MARL) [10] is applied in this paper. One challenge of MARL in our context is that the secondary users do not know the payoffs (thus do not know the strategies) of other secondary users in each stage; thus the environment of each secondary user, including other secondary users, is nonstationary and may not assure the convergence of learning. Due to the assumption that there is no mutual communication between different secondary users, many traditional MARL techniques like fictitious play [11, 12] and Nash-Q learning [13] cannot be used since they need information exchange among players (e.g., exchanging their action information). To alleviate the lack of mutual communication, we extend the principle of single-agent Q-learning, that is, evaluating the values of different state-action pairs in an incremental way, to the multiagent situation without information exchange. By applying the theory of stochastic approximation [14], which has been used in many studies on wireless networks [15, 16], we will prove the main result of this paper, that is, the learning converges to a stationary point regardless of the initial strategies (Propositions 1 and 2).

Some studies on reinforcement learning in cognitive radio networks have been done [17–19]. In [17] and [19], the studies are focused on the resource competition in a spectrum auction system, where the channel allocation is determined by the spectrum regulator, which is different from this paper in which no regulator exists. Reference [18] discusses

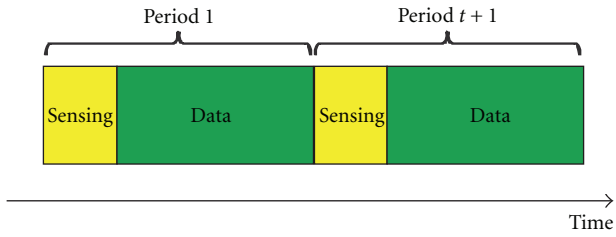

Figure 2: Timing structure of spectrum sensing and data transmission.

correlated equilibrium and achieves it by no-regret learning; that is, minimizing the gap between the current reward and optimal reward. In this approach, mutual communication is needed among the secondary users. However, in our study, no intersecondary-user communication is assumed.

Note that the study in this paper has subtle similarities to the evolutionary game theory [20], which has been successfully applied in the cooperative spectrum sensing in cognitive radio systems [21]. Both our study and the evolutionary game focus on the dynamics of strategy changes of users. However, there is a key difference between the two studies. The evolutionary game theory assumes pure strategies for the players (e.g., cooperate or free-ride in cooperative spectrum sensing [21]) and studies the proportions of players using different pure strategies. The key equation in the evolutionary game theory, called replicator equation, describes the dynamics of the corresponding proportions. In contrast to the evolutionary game, the players in our study use mixed strategies and the basic (16) describes the dynamics of the Q-values for different channels. Although the convergence is proved by studying ordinary different equations in both studies, the proof is significantly different since the equations have totally different expressions.

The remainder of this paper is organized as follows. In Section 2, the system model is introduced. Basic elements of the game and the proposed multiagent Q-learning for fully observable case (i.e., each secondary user can sense all channels) are introduced in Section 3. The corresponding convergence of Q-learning is proved in Section 4. The Q-learning for partially observable case (i.e., each secondary user can sense a subset of the channels) is discussed in Section 5. Numerical results are provided in Section 6, while conclusions are drawn in Section 7.

2. System Model

We consider N active secondary users accessing N licensed frequency channels. (When there are more than N channels, there is less competition; thus making the problem easier. We do not consider the case when the number of channels is less than the number of secondary users since a typical cognitive radio system can provide sufficient channels. Meanwhile, the proposed algorithm can also be applied to all possible cases of N.) We index the secondary users, as well as the channels, by integers $1, 2, \ldots, N$. For simplicity, we denote by i^- the set of users (channels) different from user (channel) i.

The following assumptions are made throughout this paper.

(i) The secondary users are sufficiently close to each other such that they share the same activity of primary users. There is no communication among these secondary users, thus excluding the possibility of negotiation.

(ii) We assume that the activity of primary users over each channel is a Markov chain (A more reasonable model for the activity of primary users is the semi-Markov chain. The corresponding analysis is more tedious but similar to that in this paper. Therefore, for simplicity of analysis, we consider only Markov chain in this paper) with states B (busy: the channel is occupied by primary users and cannot be used by secondary users) and I (idle: there is no primary user over this channel). We denote by $S_j(t)$ the state of channel j in the sensing period of the tth spectrum access period. For channel i, the transition probability from state B to state I (resp., from state I to state B) is denoted by P_{BI}^i (resp., P_{IB}^i). We assume that the N Markov chains for the N channels are mutually independent. We also assume perfect spectrum sensing and do not consider possible errors of spectrum sensing.

(iii) We assume that the channel state transition probabilities, as well as the channel rewards, are unknown with the secondary users at the beginning. They are fixed throughout the game, unless otherwise noted. Therefore, the secondary users need to learn the channel properties.

(iv) The timing structure of spectrum sensing and data transmission is illustrated in Figure 2, where data is transmitted after the spectrum sensing period. We assume that each secondary user is able to sense only one channel during the spectrum sensing period and transmit over only one channel during the data transmission period.

In Sections 3 and 4, we consider the case in which all secondary users have full knowledge of channel states in the previous spectrum access period (*complete observation*). Note that this does not contradict the assumption that a secondary user can sense only one channel during the spectrum sensing period since the secondary user can continue to sense other channels during the data transmission period (suppose that the signal from primary users can be well distinguished from that from secondary users, e.g., using different cyclostationary features [22]). If we consider the set of channel states in the previous spectrum access period as the system state, denoted by $S(t)$ at spectrum access period t, then the previous assumption implies a completely observable system state, which substantially simplifies the analysis. In Section 5, we will also study the case in which secondary users cannot continue to sense during the data transmission period (*partial observation*); thus each secondary user has only partial observations about the system state.

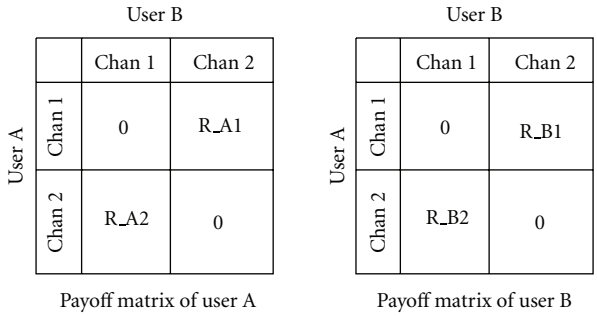

Figure 3: Examples of payoff matrices in a two-player and two-channel game of aloha-like spectrum access.

3. Game and Q-Learning

In this section, we introduce the game associated to the learning procedure and the application of Q-learning to the Aloha-like spectrum access problem. Note that in this section and Section 3, we assume that each secondary user knows all channel states in the previous time slot, that is, the completely observable case.

3.1. Game of Aloha-Like Spectrum Access. The Aloha-like spectrum access problem is essentially an $N \times N$ game. When secondary user i transmits over an idle channel j, it receives reward $R_{ij} > 0$ (e.g., channel capacity or successful transmission probability), if no other secondary user transmits over this channel, and reward 0, if one or more other secondary users are transmitting over this channel, since collision will happen. We assume that the reward R_{ij} does not change with time. When channels change slowly, the learning algorithm proposed in this paper can also be applied to track the change of channels. When channels change very fast, it is impossible for secondary users to learn. Since there is no explicit information exchange among secondary users, the collision avoidance is completely based on the received reward. The payoff matrices for the case of $N = 2$ are given in Figure 3. Note that the actions, denoted by $a_i(t)$ for user i at time t, in the game are the selections of channels. Obviously, the diagonal elements in the payoff matrices are all zero since collision yields zero reward.

It is well known that Nash equilibrium means the strategies such that unilaterally changing strategy incurs the degradation of its own performance. Mathematically, a Nash equilibrium means a set of strategies $\{\sigma_k^*\}_k$, where σ_k^* is the strategy of player k, which satisfy

$$r_k\left(\sigma_k^*, \sigma_{-k}^*\right) \geq r_k\left(\sigma_k, \sigma_{-k}^*\right), \quad \forall\, \sigma_k, \tag{1}$$

where r_k means the reward of player k and σ_{-k}^* means the strategies of all players except player k.

It is easy to verify that there are multiple Nash equilibrium points in the game. Obviously, orthogonal transmission strategies, that is, $a_i(t) \neq a_j(t)$, $\forall i \neq j$, are pure equilibria. The reason is the following. If a secondary user changes its strategy and transmits over other channels with nonzero probability, those transmission will collide with other secondary users (recall that, for the Nash equilibrium, all other secondary users do not change their strategies) and incurs performance degradation. The orthogonal channel assignment can be achieved in the following approach: *let all secondary users sense the channel randomly at the very beginning; once a secondary user finds an idle channel, it will access this channel forever; after a random number of rounds, all secondary users will find different channels, thus achieving the orthogonal transmission*. We call this scheme the *simple orthogonal channel assignment* since it is simple and fast. However, in this scheme, the different rewards of different channels are ignored. As will be seen in the numerical simulation results, the proposed learning procedure can significant outperform the simple orthogonal channel assignment.

3.2. Q-Value. We define the Q-function as the expected reward in one time slot (since the channel states are completely known to the secondary users and are not controlled by the secondary users, each secondary user needs to consider only the expected reward in one time slot, that is, a myopic strategy) of each action under different states; that is, for secondary user i and system state s, the Q-value of choosing channel j is given by

$$Q_{ij}^s = E[\mathrm{R}_i \mid a_i(t) = j,\ S(t) = s], \tag{2}$$

where R_i is the reward obtained by secondary user i, which is dependent on the action, as well as the system state, and the expectation is over the randomness of other users' actions, as well as the primary users' occupancies.

3.3. Exploration. In contrast to fictitious play [11], which is deterministic, the action in Q-learning is random. We assign nonzero probabilities for all channels such that all channels will be explored. Such an exploration guarantees that good channels will not be missed during the learning procedure. We consider Boltzmann distribution [23] for random exploration, that is,

$$P(\text{user } i \text{ chooses channel } j \mid \text{state } s) = \frac{e^{Q_{ij}^s/\gamma}}{\sum_{k=1}^N e^{Q_{ik}^s/\gamma}}, \tag{3}$$

where γ is called *temperature*, which controls the randomness of exploration. Obviously, the smaller γ is (the colder), the more focused the actions are. When $\gamma \to 0$, each user chooses only the channel having the largest Q-value.

When secondary user i selects channel j and the system state is s, the expected reward is given by

$$E[R_i(j) \mid s] = R_{ij} P\left(S_j(t) = I \mid s\right) \prod_{k=1, k\neq i}^{N} \left(1 - \frac{e^{Q_{kj}^s/\gamma}}{\sum_{l=1}^N e^{Q_{kl}^s/\gamma}}\right), \tag{4}$$

since secondary user k ($k \neq i$) chooses channel j with probability $e^{Q_{kj}^s/\gamma} / \sum_{l=1}^N e^{Q_{kl}^s/\gamma}$ (collision happens and secondary user i receives no reward) and channel j is idle with probability $P(S_j(t) = I \mid s)$; then the product in (4) is the probability that no other secondary user accesses channel j.

3.4. Updating Q-Values. In the procedure of Q-learning, the Q-functions are updated after each spectrum access using the following rule:

$$Q_{ij}^{s}(t+1) = \left(1-\alpha_{ij}(t)\right)Q_{ij}^{s}(t) + \alpha_{ij}(t)r_i(t)I\left(a_i(t)=j\right), \tag{5}$$

where $\alpha_{ij}(t)$ is a step factor (when channel j is not selected by user i, we set $\alpha_{ij}(t)=0$), $r_i(t)$ is the reward of secondary user i and I is the characteristic function of the event that channel j is selected by secondary user i at the tth spectrum access. Note that this is the standard Q-learning without considering the future states. An intuitive explanation for (5) is that, once channel j is accessed, the corresponding Q-value is updated by combining the old value and the new reward; if channel j is not chosen, we keep the old value by setting $\alpha_{ij}(t)=0$. Our study is focused on the dynamics of (5). To assure convergence, we assume that

$$\sum_{t=1}^{\infty}\alpha_{ij}(t)=\infty, \qquad \forall i=1,\ldots,N, j=1,\ldots,N, \tag{6}$$

as well as

$$\sum_{t=1}^{\infty}\alpha_{ij}^{2}(t)<\infty, \qquad \forall i=1,\ldots,N, j=1,\ldots,N. \tag{7}$$

Note that, in a typical stochastic game setting and Q-learning, the updating rule in (5) should consider the reward of the future and add a discounted term of the future reward to the right hand side of (5). However, in this paper, the optimal strategy is myopic since we assume that the system state is known, and thus the secondary users' actions do not affect the system state. For the case of partial observation (i.e., each secondary user knows only the state of a single channel), the action does change each secondary user's state (typically the belief of system state), and the future reward should be included in the right hand side of (5), which will be discussed in Section 5.

3.5. Stationary Point. The Q-values for different users are mutually coupled and all Q-values change if one Q-value is changed since the strategy of the corresponding user is changed, thus changing the expected rewards of other users. We define Q-values satisfying the following equations as a *stationary point*

$$\begin{aligned} Q_{ij}^{s} &= R_{ij}P\left(S_j(t)=I\mid s\right)\prod_{k\neq i}P\left(a_k(t)\neq j\right) \\ &= R_{ij}P\left(S_j(t)=I\mid s\right)\prod_{k\neq i}\left(1-\frac{e^{Q_{kj}^{s}/\gamma}}{\sum_{r=1}^{N}e^{Q_{kj}^{s}/\gamma}}\right), \\ &\qquad \forall i,j=1,\ldots,N. \end{aligned} \tag{8}$$

Note that the stationarity is only in the statistical sense since the Q-values can fluctuate around the stationary point due to the randomness of exploration. Obviously, as $\gamma\to 0$, the stationary point converges to a Nash equilibrium point. However, we are still not sure about the existence of such a stationary point. The following lemma assures the existence of stationary point. The proof is given in Appendix B.

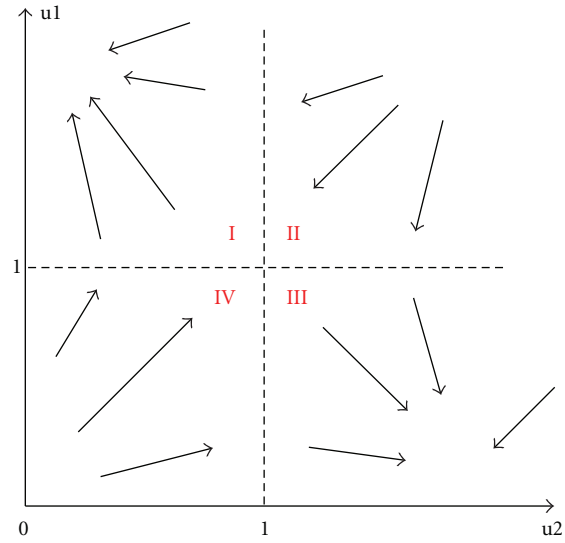

Figure 4: Illustration of the dynamics in the 2×2 Q-learning.

Lemma 1. *For sufficiently small γ, there exists at least one stationary point satisfying* (8).

4. Convergence of Q-learning without Information Exchange

In this section, we study the convergence of the proposed Q-learning algorithm. First, we provide an intuitive explanation for the convergence in 2×2 case. Then, we apply the tools of stochastic approximation and ordinary differential equation (ODE) to prove the convergence rigorously.

4.1. Intuition on Convergence. As will be shown in Proposition 1, the updating rule of Q-values in (5) will converge to a stationary equilibrium point close to Nash equilibrium. Before the rigorous proof, we provide an intuitive explanation for the convergence using the geometric argument proposed in [24].

The intuitive explanation is provided in Figure 4 for the case of $N=2$ (we call it *Metrick-Polak plot* since it was originally proposed by A. Metrick and B. Polak in [24]). For simplicity, we ignore the indices of state and assume that both channels are idle. The axes are $\mu_1=Q_{11}/Q_{12}$ and $\mu_2=Q_{21}/Q_{22}$, respectively. As labeled in the figure, the plane is divided into four regions separated by two lines $\mu_1=1$ and $\mu_2=1$, in which the dynamics of Q-learning are different. We discuss these four regions separately.

(i) Region I: in this region, $Q_{11}>Q_{12}$; therefore, secondary user 1 prefers visiting channel 1; meanwhile, secondary user 2 prefers accessing channel 2 since $Q_{22}>Q_{21}$; then, with large probability, the strategies will converge to a stationary point in which secondary users 1 and 2 access channels 1 and 2, respectively.

(ii) Region II: in this region, both secondary users prefer accessing channel 1, thus causing many collisions.

Therefore, both Q_{11} and Q_{21} will be reduced until entering either region I or region III.

(iii) Region III: similar to region I.

(iv) Region IV: similar to region II.

Then, we observe that the points in Regions II and IV are unstable and will move into Region I or III with large probability. In Regions I and III, the strategy will move close to the stationary point with large probability. Therefore, regardless where the initial point is, the updating rule in (5) will converge to a stationary point with large probability.

4.2. Stochastic Approximation-Based Convergence. In this section, we prove the convergence of the Q-learning of the proposed Aloha-like spectrum access with Boltzman distributed exploration. First, we find the equivalence between the updating rule (5) and Robbins-Monro iteration [25] for solving an equation with unknown expression (a brief introduction is provided in Appendix A). Then, we apply a conclusion in stochastic approximation [14] to relate the dynamics of the updating rule to an ODE and prove the convergence of the ODE.

4.2.1. Robbins-Monro Iteration. At a stationary point, the expected values of Q-functions satisfy the equations in (8). For system state s, define

$$\mathbf{q}^s \triangleq \left(Q_{11}^s,\ldots,Q_{1N}^s,\ldots,Q_{21}^s,\ldots,Q_{2N}^s,Q_{N1}^s,\ldots,Q_{NN}^s\right)^T. \tag{9}$$

Then, (8) can be rewritten as

$$\mathbf{g}^s\left(\mathbf{q}^s\right) = \mathbf{A}^s\left(\mathbf{q}^s\right)\bar{\mathbf{r}} - \mathbf{q}^s = 0, \tag{10}$$

where

$$\bar{\mathbf{r}} \triangleq \left(R_{11},\ldots,R_{1N},\ldots,R_{21},\ldots,R_{2N},R_{N1},\ldots,R_{NN}\right)^T, \tag{11}$$

and (function $\mathrm{mod}(x,N)$ means the remainder of dividing integer x with integer N)

$$\mathbf{A}_{ij}^s \triangleq \begin{cases} P\left(S_{\mathrm{mod}\,(j,N)} = Is\right) \displaystyle\prod_{k \neq \mathrm{mod}\,(i,N)} \left(1 - \frac{e^{Q_{k\,\mathrm{mod}\,(j,N)}^s/\gamma}}{\sum_{p=1}^{N} e^{Q_{kp}^s/\gamma}}\right), \\ \qquad \text{if } \mathrm{mod}\,(i,N) = \mathrm{mod}\,(j,N), \\ 0, \\ \qquad \text{if } \mathrm{mod}\,(i,N) \neq \mathrm{mod}\,(j,N) \end{cases} \tag{12}$$

with convention $\mathrm{mod}(N,N) = N$. Obviously, $\mathbf{A}_{ij}^s$ is the probability that channel $\mathrm{mod}(j,N)$ can be used by secondary user $\mathrm{mod}(i,N)$ without collision with other secondary users, when the current system state is s.

Then, the updating rule in (5) is equivalent to solving (8) (the expression of the equation is unknown since the rewards, channel transition probabilities, as well as the strategies of other users, are all unknown) using Robbins-Monro algorithm [14], that is,

$$\begin{aligned} \mathbf{q}^s(t+1) &= (1-\alpha(t))\mathbf{q}^s(t) + \alpha(t)\mathbf{r}(t) \\ &= \mathbf{q}^s(t) + \alpha(t)\mathbf{Y}^s(t), \end{aligned} \tag{13}$$

where $\alpha(t)$ is the vector of all step factors, $\mathbf{r}(t)$ is the vector of rewards obtained at spectrum access period t and $\mathbf{Y}^s(t)$ is a random observation on function g^s contaminated by noise, that is,

$$\begin{aligned} \mathbf{Y}^s(t) &= \mathbf{r}(t) - \mathbf{q}^s(t) \\ &= \bar{\mathbf{r}}^s(t) - \mathbf{q}^s(t) + \mathbf{r}(t) - \bar{\mathbf{r}}^s(t) \\ &= \mathbf{g}^s\left(\mathbf{q}^s(t)\right) + \delta\mathbf{m}^s(t), \end{aligned} \tag{14}$$

where $\mathbf{g}^s(\mathbf{q}^s(t)) = \bar{\mathbf{r}}^s(t) - \mathbf{q}^s(t)$, $\delta\mathbf{m}^s(t) = \mathbf{r}(t) - \bar{\mathbf{r}}^s(t)$ is noise and (recall that $r_i(t)$ means the reward of secondary user i at time t)

$$\bar{\mathbf{r}}^s(t) = \mathbf{A}^s\left(\mathbf{q}^s(t)\right)\bar{\mathbf{r}}. \tag{15}$$

Obviously, $E[\delta\mathbf{m}^s(t)] = 0$ since the expectation of the difference between the reward and the expected reward is equal to 0. Therefore, the observation $\delta\mathbf{m}^s(t)$ is a Martingale difference.

4.2.2. ODE and Convergence. The procedure of Robbins-Monro algorithm (i.e., the updating of Q-value) is the stochastic approximation of the solution of the equation. It is well known that the convergence of such a procedure can be characterized by an ODE. Since the noise $\delta\mathbf{m}(t)$ in (14) is a Martingale difference, it is easy to verify the conditions in Theorem 1 in Appendix A and obtain the following lemma (the proof is given in Appendix C).

Lemma 2. *With probability 1, the sequence $\mathbf{q}^s(t)$, $\forall s$, converges to some limit set of the ODE*

$$\dot{\mathbf{q}}^s = \mathbf{g}^s\left(\mathbf{q}^s\right). \tag{16}$$

What remains to do is to analyze the convergence property of the ODE (16). We obtain the following lemma by applying Lyapunov function. The proof is given in Appendix D.

Lemma 3. *If a stationary point determined by* (10) *exists, the solution of ODE* (16) *converges to the stationary point for sufficiently large γ.*

Combining Lemmas 1, 2, and 3, we obtain the main result in this paper.

Proposition 1. *Suppose that a stationary point determined by* (10) *exists. For any system state s and sufficiently large γ, the Q-learning converges to a stationary point with probability 1.*

Note that a sufficiently small γ guarantees the existence of stationary point and a sufficiently large γ assures the convergence of the learning procedure. However, they do

not conflict since they are not necessary conditions. As we found in our simulations, we can always choose a suitable γ to guarantee the existence of the stationary point and the convergence.

5. Q-Learning with Partial Observations

In this section, we remove the assumption that all secondary users know all channel states in the previous spectrum access period and assume that each secondary user knows the state of only the channel sensed in the previous spectrum access period; thus making the system state partially observable. The difficulties of analyzing such a scenario are given below:

(i) The system state is partially observable.

(ii) The game is imperfectly monitored, that is, each player does not know other players' actions.

(iii) The game has incomplete information, that is, each player does not know the strategies of other players, as well as their beliefs on the system state.

Note that the latter two difficulties are common for both the complete and partial observation cases. However, the imperfect monitoring and incomplete information add much more difficulty in the partial observation case. In this section, we formulate the Q-learning algorithm and then prove the convergence under certain conditions.

5.1. State Definition. It is well known that, in partially observable Markov decision process (POMDP) problems, the belief on the system state, that is, $P(S(t)\mathcal{H}_{t-1})$ ($\mathcal{H}_{t-1}$ is the observation history before period t), can play the role of system state. Due to the special structure of the game, we can define the state of secondary user i at period t as

$$\Theta_i(t) = (\tau_{i1}(t), \ldots, \tau_{iN}(t), S_{i1}(t), \ldots, S_{iN}(t)), \tag{17}$$

where $\tau_{ij}(t)$, $j = 1, \ldots, N$, means the number of consecutive periods during which channel j has not been sensed before period t (e.g., if the last time that channel j was sensed by secondary user i is time slot 5, $\tau_{ij}(t) = t - 5$.) and $S_{ij}(t)$ is the state of channel j in the last time when it is sensed before period t.

5.2. Learning in the POMDP Game. For the purpose of learning, we define the objective function for user i as the discounted sum of rewards in each spectrum access period with discount factor β, that is,

$$J_i = (1-\beta)\sum_{t=0}^{\infty}\beta^t E[r_i(t)]. \tag{18}$$

Then, to maximize the objectively function, the corresponding Q-learning strategy is given by [23]

$$Q_{ij}^{\Theta}(t+1) = \left(1-\alpha_{ij}^{\Theta}(t)\right)Q_{ij}^{\Theta}(t) + \alpha_{ij}^{\Theta}(t)\left(r_i(t)I(a_i(t)=j) + \max_{k=1,\ldots,N}\beta Q_{ik}^{\Theta'}(t)\right), \tag{19}$$

where Θ' is uniquely determined by Θ and j, and $\alpha_{ij}^{\Theta}(t)$ is the step factor dependent on the time, channel, user, and belief state. Note that Θ' is the system state in the next time slot, which is random. Intuitively, the new Q-value is updated by combining the old value and the new estimation, which is the sum of the new reward and discounted old Q-value.

Similarly to the complete information situation, we have the following proposition which states the convergence of the learning procedure with partial information and large γ. The proof is given in Appendix E. Note that numerical simulation shows that small γ also results in convergence. However, we are still unable to prove it rigorously.

Proposition 2. *When γ is sufficiently large, the learning procedure in* (19) *converges.*

6. Numerical Results

In this section, we use numerical simulations to demonstrate the theoretical results obtained in previous sections. For the fully observable case, we use the following step factor:

$$\alpha_{ij}(t) = \frac{\alpha_0}{\#\text{ of times user } i \text{ selects channel } j \text{ before time } t}, \tag{20}$$

where α_0 is the initial learning factor. A similar step factor $\alpha_{ij}^{\Theta}(t)$ is used for the partially observable case. In Sections 6.1, 6.2, and 6.3, we consider the fully observable case and, in Section 6.4, we consider the partially observable case. Note that, in all simulations, we initialize the Q-values by choosing uniformly random variables in the interval $[0, 1]$.

6.1. Dynamics. Figures 5 and 6 show the dynamics of μ_1 versus μ_2 (recall that $\mu_1 = Q_{11}/Q_{12}$ and $\mu_2 = Q_{21}/Q_{22}$) of several typical trajectories for the state of both channels being idle when $N = 2$. We assume that $R_{ij} = 1$ for all i and j. Note that $\gamma = 0.1$ in Figure 5 and $\gamma = 0.01$ in Figure 6. We observe that the trajectories move from unstable regions (II and IV in Figure 4) to stable regions (I and III in Figure 4). We also observe that the trajectories for smaller temperature γ is smoother since less explorations are carried out.

Figure 7 shows the evolution of the probability of choosing channel 1 when $N = 2$, $\gamma = 0.1$ and both channels are idle. We observe that both secondary users prefer channel 1 at the beginning and soon secondary user 1 intends to choose channel 2, thus avoiding collisions.

6.2. CDF of Rewards. In this subsection, we consider the performance of reward averaged over all system states. When $N = 3$, we set $P_{II} = 0.9, 0.8, 0.7$ and $P_{BI} = 0.3, 0.4, 0.1$ for the three channels, respectively. When $N = 2$, we use the first two channels in the case of $N = 3$. The rewards of different channels for different secondary users are randomly generated with a uniform distribution between $[0.5, 1]$. The CDF curves of performance gain, defined as the difference of average rewards after and before the learning procedure, are plotted in Figure 8 for both $N = 2$ and $N = 3$. Note that the CDF curves are obtained from 100 realizations of

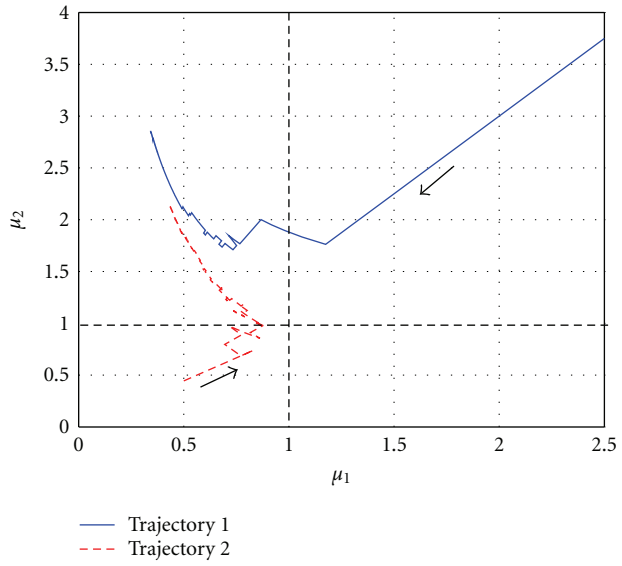

Figure 5: An example of dynamics of the Q-learning when $N = 2$, $\gamma = 0.1$, and system state is fully observable.

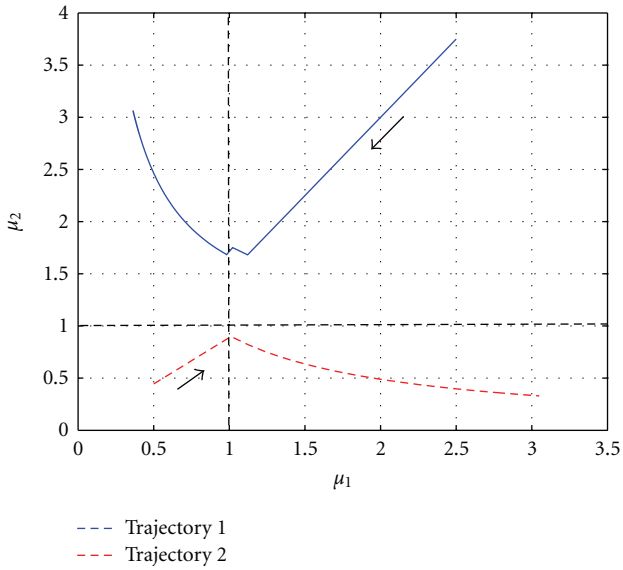

Figure 6: An example of dynamics of the Q-learning when $N = 2$, $\gamma = 0.01$, and system state is fully observable.

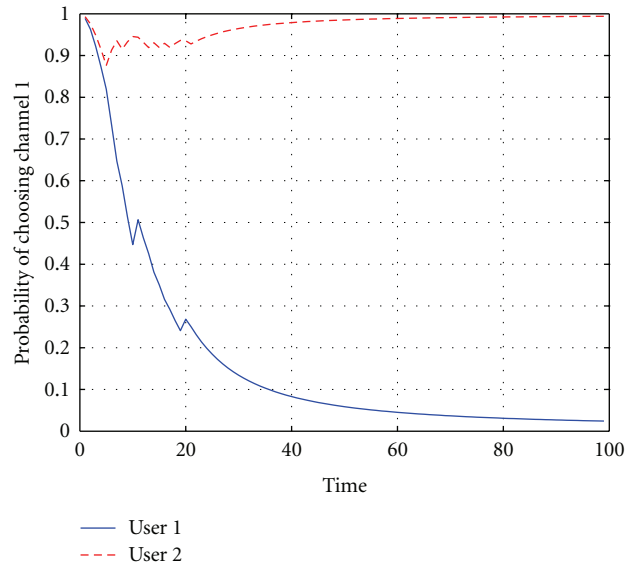

Figure 7: An example of the evolution of aloha-like spectrum access probability when $N = 2$ and system state is fully observable.

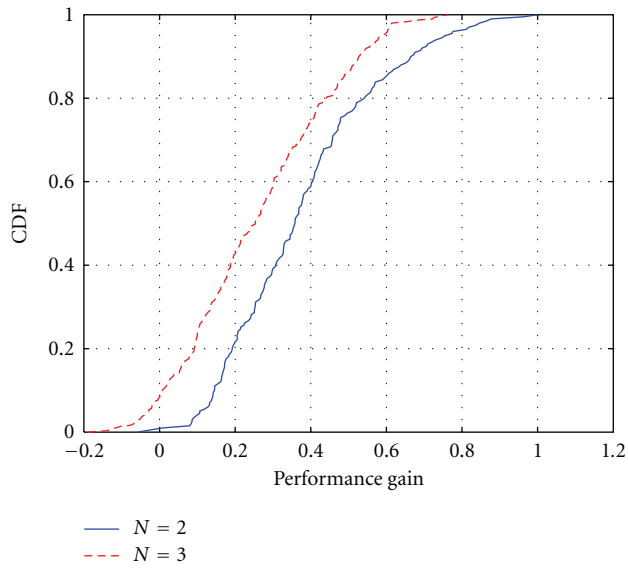

Figure 8: CDF of performance gain over the random Q values when system state is fully observable.

learning procedure. From a CDF curve, we can read the distribution of the performance gains. For example, for the curve $N = 2$, performance gain 0.4 in the horizontal axis corresponds to 0.6 in the vertical axis; this means that around 60% of the secondary users obtain performance gain less than 0.4. We observe that when $N = 2$, most performance gains are positive. However, when $N = 3$, a small portion of the performance gains are negative, that is, the performance is decreased after the learning procedure. Such a performance reduction is reasonable since Nash equilibrium may not be Pareto optimal. We also plotted the average performance gains versus different α_0 and γ in Figure 9. We observe that larger γ results in worse performance gain. When γ is small, smaller α_0 yields better performance, but decreases faster than larger α_0 when γ increases. The performance gain over the simple orthogonal channel assignment scheme is given in Figure 10. We observe that the learning procedure generates a much better performance than the simple orthogonal channel assignment.

6.3. Learning Speed. We define the stopping time of learning as the time that the relative fluctuation of average reward, which is obtained from 2000 spectrum access periods using the current Q-values, has been below 5 percent for successive

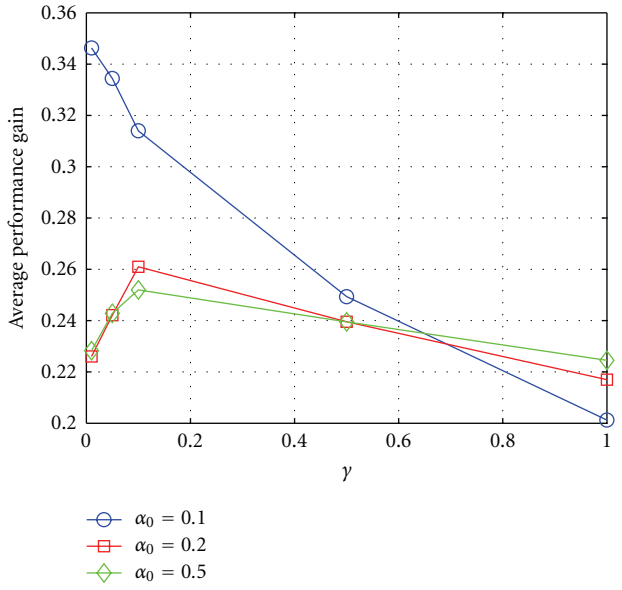

Figure 9: Average performance gains over the random Q values versus different γ and α when system state is fully observable.

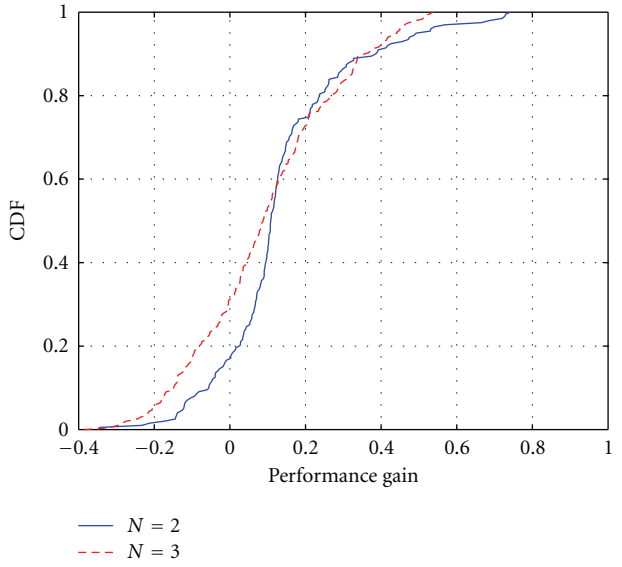

Figure 10: CDF of performance gain over the simple orthogonal channel assignment when system state is fully observable.

5 time slots. That is, compute the relative fluctuation at time slot t using

$$\delta(t) = \frac{\|\mathbf{Q}(t) - \mathbf{Q}(t-1)\|}{\|\mathbf{Q}(t)\|}, \tag{21}$$

where $\mathbf{Q}$ is the vector containing all Q-values and the norm is 2-norm. Then, when $\delta(t)$ is smaller than 0.05 for 5 consecutive times, we claim that the learning is completed. Then, the learning delay is the time spent before the stopping time. Obviously, the smaller the learning delay is, the faster the learning is. Figures 11 and 12 show the delays of learning,

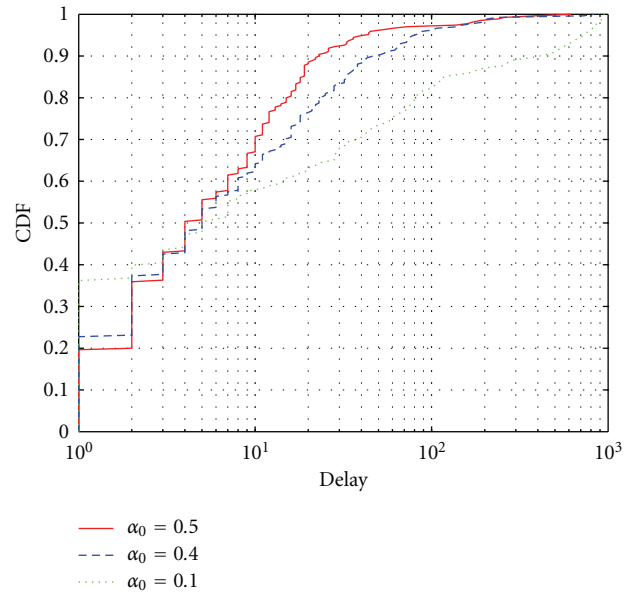

Figure 11: CDF of learning delay with different learning factor α_0 when $N = 2$ and system state is fully observable.

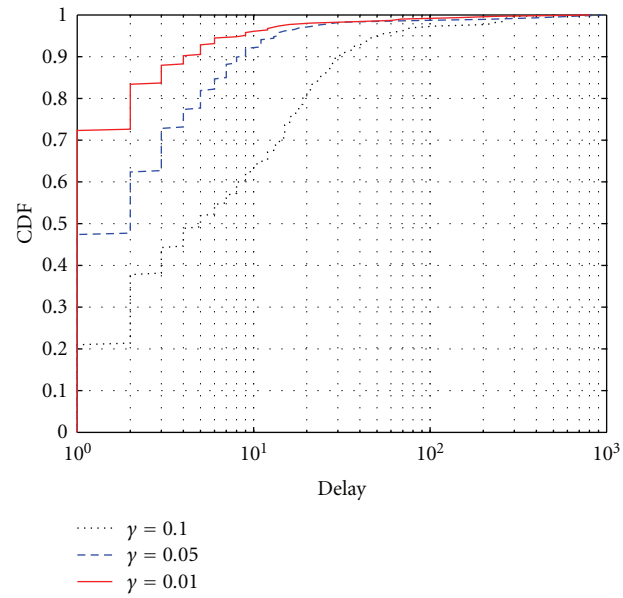

Figure 12: CDF of learning delay with different temperature γ when $N = 2$ and system state is fully observable.

which characterizes the learning speed, for different learning factor α_0 and different temperature γ, respectively, when $N = 2$. The original Q values are randomly selected. When the probabilities of choosing channel 1 are larger than 0.95 for one secondary user and smaller than 0.05 for the other secondary user, we claim that the learning procedure is completed. We observe that larger learning factor α_0 results in smaller delay while smaller γ yields faster learning procedure.

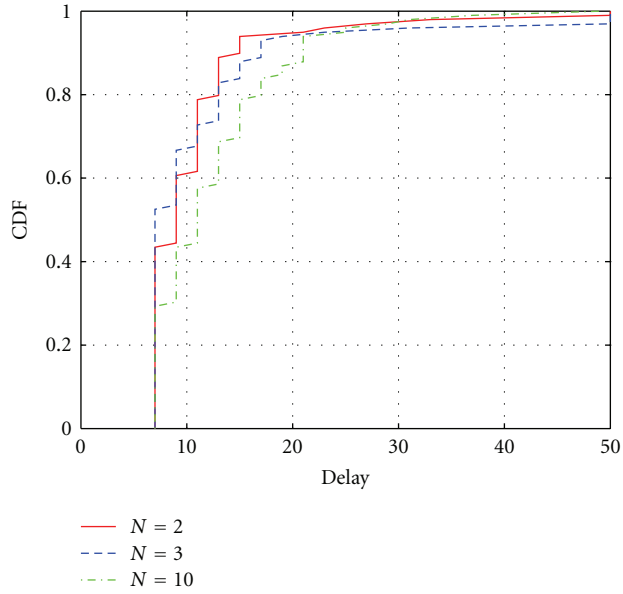

Figure 13: CDF of learning delay when the system state is fully observable.

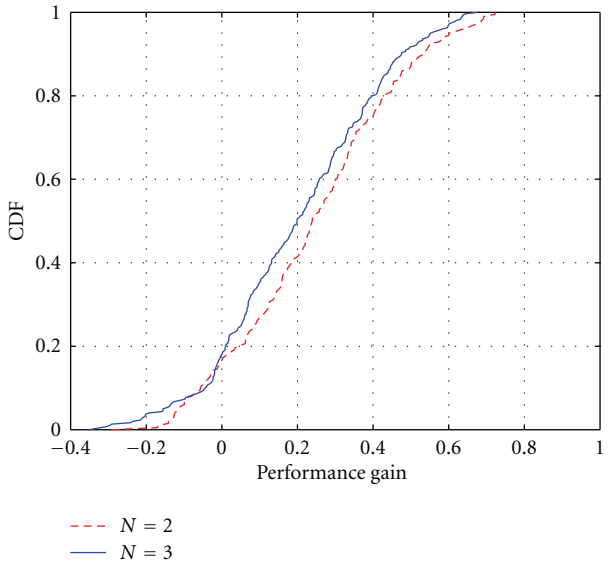

Figure 14: CDF of performance gain when the system state is fully observable and the channel rewards change with time.

The speed of learning is compared for $N = 2$, $N = 3$ and $N = 10$ in Figure 13 (both α and γ are fixed). We observe that, for more than 90% of the realizations, the learning can be completed within 20 spectrum access periods. However, the learning procedure may last for a long period of time for some situations. We can notice that the learning speeds are similar for cases $N = 2$ and $N = 3$. We also observe that, when N is much larger ($N = 10$), the increase of delay is not significant.

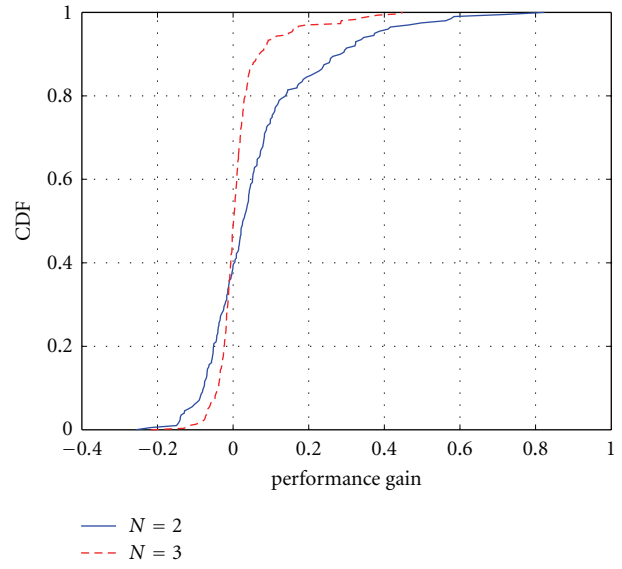

Figure 15: CDF of performance gain over the random Q values in partial observation case.

6.4. Time-Varying Channel Rewards. In previous simulations, the rewards of successfully accessing channels, R_{ij} are assumed to be constant. In practical systems, they may change with time since wireless channels are usually dynamic. In Figure 14, we show the CDF of performance gains (the configuration is the same as that in Figure 8) when channel changes slowly. We used a simple model for channel reward, which is given by

$$R_{ij}(t+1) = \max\left(0, R_{ij}(t) + 0.05 * \theta\right), \tag{22}$$

where θ is a random variable uniformly distributed between 0 and 1. From Figure 14, we observe that the learning algorithm still improves the performance significantly.

6.5. Partial Observation Case. Figure 15 shows the performance gain of learning in the case of partial observations. We adopt the Q-learning mechanism introduced in Section 5. Note that there are infinitely many belief states since a channel could be unsensed for an infinite period of time. For computational simplicity, we set all $\tau_{ij}(t) > 5$ to $\tau_{ij}(t) = 5$ (recall that $\tau_{ij}(t)$ is the period of time that channel j has not been sensed by user i before time t). From Figure 15, we observe that the performance is actually degraded for around 40% ($N = 2$) or 50% ($N = 3$) cases. However, the amplitude of performance degradation is averagely less than the amplitude of performance gain. We also observe that the performance gain is decreased when N is increased from 2 to 3.

The learning delay for the partial observation case is shown in Figure 16, where the simulation setup is similar to that of Figure 13. Again, we observe that the learning speeds of $N = 2$ and $N = 3$ are similar to each other.

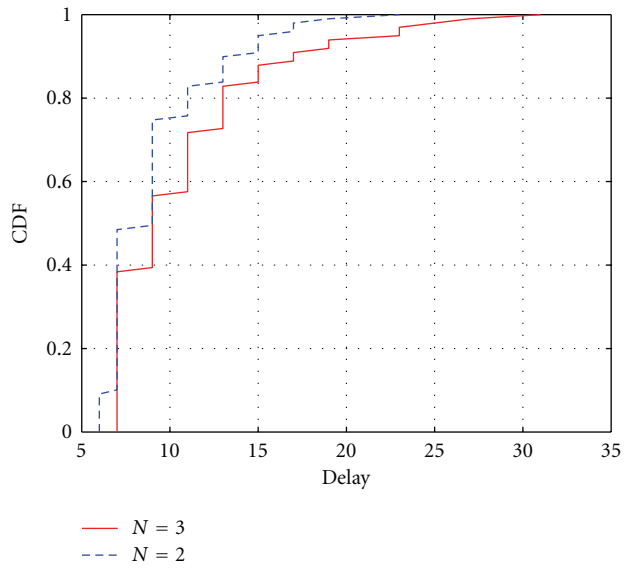

Figure 16: CDF of learning delay when the system state is partially observable.

7. Conclusions

We have discussed a learning procedure for Aloha-like spectrum access without negotiation in cognitive radio systems. During the learning, each secondary user considers the channel and other secondary users as its environment, updates its Q-values, and takes the best action. An intuitive explanation for the convergence of learning is provided using Metrick-Polak plot. By applying the theory of stochastic approximation and ODE, we have shown the convergence of learning under certain conditions. We also extended the case of full observations to the case of partial observations. Numerical results show that secondary users can learn to avoid collision quickly. The performance after the learning is significantly better than that before the learning and that using a simple scheme to achieve a Nash equilibrium. Note that our study is one extreme of the resource allocation problem since no negotiation is considered, while the other extreme is full negotiation to achieve optimal performance. Our future work will be the intermediate case; that is, limited negotiation for resource allocation.

Appendices

A. Stochastic Approximation

For being self-contained, we briefly introduce the theory of stochastic approximation and cite the conclusion used for proving Lemma 2.

Essentially, stochastic approximation is used to solve an equation with unknown expression and noisy observations. Consider equation

$$g(\theta) = 0, \tag{A.1}$$

where θ is the unknown variable and the expression of function g is unknown. Denote by θ^* the solution to this equation (we assume that there is only one solution to the equation). Suppose that $g(\theta) < 0$ for $\theta > \theta^*$ and $g(\theta) > 0$ for $\theta < \theta^*$. We have a series of noisy observations of $g(\theta)$, denoted by $\{Y_n\}$. Then, we can approximate the solution iteratively in the following way (called Robbins-Monro algorithm).

$$\theta_{n+1} = \theta_n + \alpha_n Y_n, \tag{A.2}$$

where $\alpha_n > 0$ is the step for the nth iteration.

The convergence of (A.2) is deeply related to a "mean" ODE, which is given by

$$\dot{\theta}(t) = g(\theta). \tag{A.3}$$

The following Theorem (part of Theorem 2.1 in [14]) discloses the relationship between the convergence in (A.2) and the mean ODE in (A.3).

Theorem 1. *If the following assumptions are satisfied, θ_n converges to a limit set A in which all points satisfy $g(\theta) = 0$ with probability 1:*

(A) $\sup_n E|Y_n|^2 < \infty$.

(B) *Noise is a martingale difference, that is,*

$$E[Y_n \mid \theta_0, Y_i, i < n] = g(\theta_n). \tag{A.4}$$

(C) *$g(\cdot)$ is continuous.*

(D) $\sum_i \alpha_i^2 < \infty$.

(E) *There exists a continuously differentiable function f such that $g(\cdot) = f'(\cdot)$ and f is a constant on the limit set A.*

B. Proof of Lemma 1

Proof. For simplicity, we fix one system state s since the Q-learning procedures for different state s are mutually independent when the system state is fully observable and the action of each secondary user does not affect the system state. Consider a Nash equilibrium point, at which there is no collision. Without loss of generality, we assume that secondary users $1, 2, \ldots, N$ use channels $1, 2, \ldots, N$, respectively.

Now, we choose a set of $\{Q_{ij}^s\}_{ij}$ such that

$$0 < Q_{ij}^s < Q_{ii}^s < P(S_i(t) = I \mid s)R_{ii}, \quad \forall i = 1, \ldots, N, j \neq i. \tag{B.1}$$

Then, we can always choose a sufficiently small γ such that

$$Q_{ii}^s < R_{ii}P(S_i(t) = I \mid s)\prod_{k \neq i}\left(1 - \frac{e^{Q_{ki}^s/\gamma}}{\sum_{r=1}^{N} e^{Q_{kr}^s/\gamma}}\right), \tag{B.2}$$

$$Q_{ij}^s > R_{ij}P\left(S_j(t) = I \mid s\right)\prod_{k \neq i}\left(1 - \frac{e^{Q_{kj}^s/\gamma}}{\sum_{r=1}^{N} e^{Q_{kr}^s/\gamma}}\right), \tag{B.3}$$

since the right hand sides of (B.2) and (B.3) converge to $R_{ii}P(S_t^j = I \mid s)$ and 0, respectively, as $\gamma \to 0$.

Then, we carry out the following iterations, that is, the Q-values of the mth iteration is given by, $\forall i, j = 1, 2, \ldots, N$,

$$Q_{ij}^s(m) = R_{ij}P\left(S_j(t) = I \mid s\right)\prod_{k\neq i}\left(1 - \frac{e^{Q_{kj}^s(m-1)/\gamma}}{\sum_{r=1}^N e^{Q_{kr}^s(m-1)/\gamma}}\right). \tag{B.4}$$

Next, we show that, $\forall i = 1, \ldots, N$, Q_{ii}^s increases while Q_{ij}^s $(i \neq j)$ decreases during the iterations by carrying out inductions on m. For the first iteration, Q_{ii}^s is increased while Q_{ij}^s $(i \neq j)$ is decreased due to the conditions (B.2) and (B.3). Suppose that, in the mth iteration, the conclusion holds. Then, in the $m + 1$th iteration, Q_{ii}^s is increased due to the expression of the right hand side of (B.4) and the assumptions $Q_{ii}^s(m) > Q_{ii}^s(m-1)$ and $Q_{ij}^s(m) < Q_{ij}^s(m-1)$ $(i \neq j)$. For the same reason, Q_{ij}^s is decreased in the $m + 1$th iteration. This concludes the induction.

Now, we have shown that $Q_{ii}(m)$ is a monotonically increasing sequence while $Q_{ij}(m)$ $(i \neq j)$ is a monotonically decreasing sequence. Since all sequences are bounded ($Q_{ii}(m) < R_{ii}$ and $Q_{ij}(m) > 0$ $(i \neq j)$), all sequences converge to their limits, which is the stationary point. This concludes the proof. □

C. Proof of Lemma 2

Proof. We verify the conditions in Theorem 1 one by one.

(i) Condition (A): This is obvious since Y_n is upper bounded by $\max_{i,j} R_{ij}$ (recall that Y_n is the difference between the instantaneous reward and the Q-value).

(ii) Condition (B): The martingale difference noise has been proved right after (14).

(iii) Condition (C): The function $g(x)$ is given by

$$\mathbf{g}(\mathbf{x}) = \mathbf{A}(\mathbf{x})\mathbf{r} - \mathbf{x}, \tag{C.1}$$

where $\mathbf{A}$ is defined in (12). We only need to verify the continuity of $\mathbf{A}(\mathbf{x})\mathbf{r}$. Obviously, each element in $\mathbf{A}(\mathbf{x})\mathbf{r}$ is differentiable with respect to $\mathbf{x}$. Therefore, $\mathbf{g}(\mathbf{x})$ is not only continuous but also differentiable.

(iv) Condition (D): It is guaranteed by (7).

(v) Condition (E): The function f can be defined as the integral of $\mathbf{g}$. It is continuously differentiable since $\mathbf{g}$ is continuous. It is a constant on the limit set since there is only one point at the limit set. □

D. Proof of Lemma 3

Proof. We apply Lyapunov's method to analyze the convergence of the ODE in (16). We define the Lyapunov function as

$$\begin{aligned} V^s(t) &= \left|\left|\mathbf{g}^s(t)\right|\right|^2 \\ &= \sum_{i,j}\left(\overline{r}_{ij}^s(t) - Q_{ij}^s(t)\right)^2, \end{aligned} \tag{D.1}$$

where $\overline{r}_{ij}^s(t)$ is the expected reward of secondary user i at period t.

Then, we examine the derivative of the Lyapunov function with respect to time t, that is,

$$\begin{aligned} \frac{dV^s(t)}{dt} &= 2\sum_{i,j}\frac{d\left(\overline{r}_{ij}^s(t) - Q_{ij}^s(t)\right)}{dt}\left(\overline{r}_{ij}^s(t) - Q_{ij}^s(t)\right) \\ &= 2\sum_{i,j}\frac{d\epsilon_{ij}^s(t)}{dt}\epsilon_{ij}^s(t), \end{aligned} \tag{D.2}$$

where $\epsilon_{ij}^s(t) \triangleq \overline{r}_{ij}^s(t) - Q_{ij}^s(t)$.

We have

$$\begin{aligned} \frac{d\epsilon_{ij}^s(t)}{dt} &= \frac{d\overline{r}_{ij}^s(t)}{dt} - \frac{dQ_{ij}^s(t)}{dt} \\ &= \frac{d\overline{r}_{ij}^s(t)}{dt} - \epsilon_{ij}^s(t), \end{aligned} \tag{D.3}$$

where we applied the ODE (16).

Then, we focus on the computation of $d\overline{r}_{ij}^s(t)/dt$.

For secondary user i and channel j, we have

$$\begin{aligned} \frac{d\overline{r}_{ij}^s(t)}{dt} &= R_{ij}P\left(S_j(t) = I \mid s\right)\frac{d}{dt}\left(\prod_{k\neq i}\left(1 - \frac{e^{Q_{kj}^s/\gamma}}{\sum_{r=1}^N e^{Q_{kr}^s/\gamma}}\right)\right) \\ &= R_{ij}P\left(S_j(t) = I \mid s\right)\sum_{k\neq i}\prod_{l\neq k, l\neq i}\left(1 - \frac{e^{Q_{lj}^s/\gamma}}{\sum_{r=1}^N e^{Q_{lr}^s/\gamma}}\right)\frac{d}{dt} \\ &\quad \times\left(1 - \frac{e^{Q_{kj}^s/\gamma}}{\sum_{r=1}^N e^{Q_{kr}^s/\gamma}}\right), \end{aligned} \tag{D.4}$$

where the first equation is due to the definition of $\overline{r}_{ij}^s(t)$ and the second equation is due to the rule of the derivative of products.

We consider the derivative in (D.4), that is,

$$\begin{aligned}
&\frac{d}{dt}\left(1-\frac{e^{Q_{kj}^s/\gamma}}{\sum_{r=1}^N e^{Q_{kr}^s/\gamma}}\right)\\
&\quad=-\frac{e^{Q_{kj}^s/\gamma}\left(\sum_{r=1}^N e^{Q_{kr}^s/\gamma}\right)}{\gamma\left(\sum_{r=1}^N e^{Q_{kr}^s/\gamma}\right)^2}\frac{dQ_{kj}^s}{dt}\\
&\qquad+\sum_{r=1}^N\frac{e^{Q_{kj}^s/\gamma}e^{Q_{kr}^s/\gamma}}{\gamma\left(\sum_{r=1}^N e^{Q_{kr}^s/\gamma}\right)^2}\frac{dQ_{kr}^s}{dt}\\
&\quad=\sum_{r=1,r\neq j}^N\frac{e^{Q_{kj}^s/\gamma}e^{Q_{kr}^s/\gamma}}{\gamma\left(\sum_{r=1}^N e^{Q_{kr}^s/\gamma}\right)^2}\left(\frac{dQ_{kr}^s}{dt}-\frac{dQ_{kj}^s}{dt}\right)\\
&\quad=\sum_{r=1,r\neq j}^N\frac{e^{Q_{kj}^s/\gamma}e^{Q_{kr}^s/\gamma}}{\gamma\left(\sum_{r=1}^N e^{Q_{kr}^s/\gamma}\right)^2}\left(\epsilon_{kr}^s(t)-\epsilon_{kj}^s(t)\right),
\end{aligned}\tag{D.5}$$

where the last equation is obtained from ODE (16).

Substituting (D.5) into (D.4), we obtain

$$\begin{aligned}
\frac{d\overline{r}_{ij}^s(t)}{dt}&=R_{ij}P\left(S_j(t)=I\mid s\right)\sum_{k\neq i\,l\neq k,l\neq i}\prod\left(1-\frac{e^{Q_{lj}^s/\gamma}}{\sum_{r=1}^N e^{Q_{lr}^s/\gamma}}\right)\\
&\quad\times\sum_{r=1,r\neq j}^N\frac{e^{Q_{kj}^s/\gamma}e^{Q_{kr}^s/\gamma}}{\gamma\left(\sum_{r=1}^N e^{Q_{kr}^s/\gamma}\right)^2}\left(\epsilon_{kr}^s(t)-\epsilon_{kj}^s(t)\right).
\end{aligned}\tag{D.6}$$

Combining (D.2) and (D.6), we have

$$\begin{aligned}
\frac{dV^s(t)}{dt}&=-2\sum_{i=1,j=1}^N\left(\epsilon_{ij}^s(t)\right)^2\\
&\quad+\sum_{k,r,i,j=1,(k,r)\neq(i,j)}^N c_{krij}\epsilon_{ij}^s(t)\epsilon_{kr}^s(t),
\end{aligned}\tag{D.7}$$

where the coefficient c_{krij} is given by

$$\begin{aligned}
&\frac{R_{ij}P\left(S_j(t)=I\mid s\right)e^{Q_{kj}^s/\gamma}e^{Q_{kr}^s/\gamma}}{\gamma\left(\sum_{q=1}^N e^{Q_{kq}^s/\gamma}\right)^2}\prod_{l\neq k,l\neq i}\left(1-\frac{e^{Q_{lj}^s/\gamma}}{\sum_{q=1}^N e^{Q_{lq}^s/\gamma}}\right)\\
&+\frac{R_{kr}P(S_r(t)=I\mid s)e^{Q_{ij}^s/\gamma}e^{Q_{ir}^s/\gamma}}{\gamma\left(\sum_{q=1}^N e^{Q_{iq}^s/\gamma}\right)^2}\prod_{l\neq k,l\neq i}\left(1-\frac{e^{Q_{lr}^s/\gamma}}{\sum_{q=1}^N e^{Q_{lq}^s/\gamma}}\right),
\end{aligned}\tag{D.8}$$

if $k\neq i$ and $r\neq j$, and

$$\begin{aligned}
&-\sum_{r=1,r\neq j}^N\frac{R_{ij}P\left(S_i^j=I\mid s\right)e^{Q_{kj}^s/\gamma}e^{Q_{kr}^s/\gamma}}{\gamma\left(\sum_{q=1}^N e^{Q_{kq}^s/\gamma}\right)^2}\prod_{l\neq k,l\neq i}\\
&\qquad\times\left(1-\frac{e^{Q_{lj}^s/\gamma}}{\sum_{q=1}^N e^{Q_{lq}^s/\gamma}}\right)\\
&-\sum_{r=1,r\neq j}^N\frac{R_{kr}P\left(S_k^r=I\mid s\right)e^{Q_{ij}^s/\gamma}e^{Q_{ir}^s/\gamma}}{\gamma\left(\sum_{q=1}^N e^{Q_{ij}^s/\gamma}\right)^2}\prod_{l\neq k,l\neq i}\\
&\qquad\times\left(1-\frac{e^{Q_{lj}^s/\gamma}}{\sum_{q=1}^N e^{Q_{lq}^s/\gamma}}\right),
\end{aligned}\tag{D.9}$$

if $k\neq i$ and $r=j$. When $k=i$, $c_{krij}=0$.

It is easy to verify

$$\frac{P(S_r(t)=Is)e^{Q_{ij}^s/\gamma}e^{Q_{ir}^s/\gamma}}{\left(\sum_{q=1}^N e^{Q_{iq}^s/\gamma}\right)^2}<1,\tag{D.10}$$

$$\prod_{l\neq k,l\neq i}\left(1-\frac{e^{Q_{lr}^s/\gamma}}{\sum_{q=1}^N e^{Q_{lq}^s/\gamma}}\right)<1.\tag{D.11}$$

Therefore, when γ is sufficiently large, we have

$$c_{krij}<\frac{2}{N^2-1}.\tag{D.12}$$

Then, we have

$$\begin{aligned}
\frac{1}{2}\frac{dV^s(t)}{dt}&<-\sum_{i,j=1}^N\epsilon_{ij}^2(t)\\
&\quad+\frac{2}{N^2-1}\sum_{k,r,i,j=1,(k,r)\neq(i,j)}^N\left|\epsilon_{ij}(t)\right|\left|\epsilon_{kr}(t)\right|\\
&=-\frac{1}{N^2-1}\sum_{k,r,i,j=1,(k,r)\neq(i,j)}^N\left(\left|\epsilon_{ij}(t)\right|+\left|\epsilon_{kr}(t)\right|\right)^2\\
&<0.
\end{aligned}\tag{D.13}$$

Therefore, when γ is sufficiently large, the derivative of the Lyapunov function is strictly negative, which implies that the ODE (16) converges to a stationary point. This concludes the proof. □

Remark 1. We can actually obtain a stronger conclusion from the last part of the proof, that is, the convergence can be assured if

$$\gamma\geq\frac{(N-1)(N^2-1)\max R_{ij}}{2}.\tag{D.14}$$

E. Proof of Proposition 2

Proof. We define a mapping from all Q-values to another set of Q-values, which is given by

$$T\left(Q_{ij}^{\Theta}\right) = E\left[r_{ij}\right] + \beta \min_{k} Q_{ik}^{\Theta'}, \quad \forall i, j, \Theta, \tag{E.1}$$

where Θ' is determined by Θ and j and r_{ij} is the average reward when secondary user i chooses channel j. Note that $E[r_{ij}]$ is a function of all Q-values.

What we need to prove is that T is a contraction mapping. Once this is proved, the remainder part is exactly the same as the proof of the convergence of Q-learning in [26]. Therefore, we focus on the analysis on the mapping T.

We consider two sets of Q-values, denoted by $\{Q_{ij}^{\Theta}\}_{i,j,\Theta}$ and $\{\widetilde{Q}_{ij}^{\Theta}\}_{i,j,\Theta}$, respectively. Considering the difference after the mapping T between the two sets of Q-values, we have

$$\begin{aligned} T\left(Q_{ij}^{\Theta}\right) - T\left(\widetilde{Q}_{ij}^{\Theta}\right) &= E\left[r_{ij}\right] - E\left[\widetilde{r}_{ij}\right] \\ &\quad + \beta \min_{k} Q_{ik}^{\Theta'} - \beta \min_{k} \widetilde{Q}_{ik}^{\Theta'}, \end{aligned} \tag{E.2}$$

where $E[\widetilde{r}_{ij}]$ means the average reward when the Q-values are $\{\widetilde{Q}_{ij}^{\Theta}\}_{i,j,\Theta}$. Then, we have

$$\begin{aligned} \left| T\left(Q_{ij}^{\Theta}\right) - T\left(\widetilde{Q}_{ij}^{\Theta}\right) \right| &\le \left| E\left[r_{ij}\right] - E\left[\widetilde{r}_{ij}\right] \right| \\ &\quad + \beta \left| \min_{k} Q_{ik}^{\Theta'} - \min_{k} \widetilde{Q}_{ik}^{\Theta'} \right|. \end{aligned} \tag{E.3}$$

We discuss the two terms in (E.3) separately. For the first term, we have

$$\begin{aligned} E\left[r_{ij}\right] &= \sum_{\{\Theta_k\}_{k\neq i}} P\left(\{\Theta_k\}_{k\neq i} \mid \Theta\right) R_{ij} P\left(S_j = I \Theta\right) \prod_{k\neq i} \\ &\quad \times \left(1 - \frac{e^{Q_{kj}^{\Theta_k}/\gamma}}{\sum_{p=1}^{N} e^{Q_{kp}^{\Theta_k}/\gamma}}\right), \end{aligned} \tag{E.4}$$

where $\{\Theta_k\}_{k\neq i}$ is the set of states of secondary users except user i and $P(\{\Theta_k\}_{k\neq i} \mid \Theta)$ is the probability of the set of states $\{\Theta_k\}_{k\neq i}$ conditioned on the state of secondary user i, Θ.

When γ is sufficiently large, we have

$$e^{Q_{kj}^{\Theta_k}/\gamma} = 1 + \frac{Q_{kj}^{\Theta_k}}{\gamma} + \vartheta\left(\frac{Q_{kj}^{\Theta_k}}{\gamma}\right), \tag{E.5}$$

where $\vartheta(Q_{kj}^{\Theta_k}/\gamma)$ is a polynomial of order $O((Q_{kj}^{\Theta_k}/\gamma)^2)$.

Then, it is easy to verify that (E.4) can be rewritten as

$$\begin{aligned} E\left[r_{ij}\right] &= \sum_{\{\Theta_k\}_{k\neq i}} P\left(\{\Theta_k\}_{k\neq i} \mid \Theta\right) R_{ij} P\left(S_j = I \mid \Theta\right) \\ &\quad \times \prod_{k\neq i} \left(\frac{n-1}{n} - \frac{Q_{kj}^{\Theta_k}}{n\gamma} + \varrho\left(\left\{Q_{kp}^{\Theta}\right\}_{p,\Theta}\right)\right), \\ &= \sum_{\{\Theta_k\}_{k\neq i}} P\left(\{\Theta_k\}_{k\neq i} \mid \Theta\right) R_{ij} P\left(S_j = I \mid \Theta\right) \\ &\quad \times \left(\left(\frac{n-1}{n}\right)^{n-1} - \sum_{k\neq i} \frac{(n-1)Q_{kj}^{\Theta_k}}{n^2\gamma} \right. \\ &\qquad \left. + \varepsilon\left(\left\{Q_{rp}^{\Theta}\right\}_{r,p,\Theta}\right)\right), \end{aligned} \tag{E.6}$$

where $\varrho(\{Q_{kp}^{\Theta}\}_{p,\Theta})$ and $\varepsilon(\{Q_{rp}^{\Theta}\}_{r,p,\Theta})$ are both polynomials of smaller order than $O(\{\widetilde{Q}_{rp}^{\Theta}\}_{r,p,\Theta})$. Note that the coefficients of both polynomials are independent of the Q-values.

Then, we have

$$\begin{aligned} \left| E\left[r_{ij}\right] - E\left[\widetilde{r}_{ij}\right] \right| &= \sum_{\{\Theta_k\}_{k\neq i}} C_{ij}^{\Theta_k} \left(\frac{Q_{kj}^{\Theta_k}}{\gamma} - \frac{\widetilde{Q}_{kj}^{\Theta_k}}{\gamma}\right) \\ &\quad + \varepsilon\left(\left\{Q_{rp}^{\Theta}\right\}_{r,p,\Theta}\right) - \varepsilon\left(\left\{\widetilde{Q}_{rp}^{\Theta}\right\}_{r,p,\Theta}\right). \end{aligned} \tag{E.7}$$

Then, we can always take a sufficiently large γ such that

$$\left| E\left[r_{ij}\right] - E\left[\widetilde{r}_{ij}\right] \right| \le \frac{1-\beta}{2} \max_{p,q,\theta} \left| Q_{pq}^{\theta} - \widetilde{Q}_{pq}^{\theta} \right|. \tag{E.8}$$

Now, we turn to the second term in (E.3). Without loss of generality, we assume $\beta \min_k Q_{ik}^{\Theta'} \ge \beta \min_k \widetilde{Q}_{ik}^{\Theta'}$. We have

$$\begin{aligned} &\beta \min_{k} Q_{ik}^{\Theta'} - \beta \min_{k} \widetilde{Q}_{ik}^{\Theta'} \\ &\quad \le \beta Q_{iq}^{\Theta'} - \beta \widetilde{Q}_{iq}^{\Theta'} \\ &\quad \le \beta \left(\max_{r,s,\theta} Q_{rs}^{\theta} - \widetilde{Q}_{rs}^{\theta}\right), \end{aligned} \tag{E.9}$$

where, in the first inequality, we define q as

$$q = \arg\min_{k} \widetilde{Q}_{ik}^{\Theta'}. \tag{E.10}$$

Due to symmetry, we have

$$\left| \beta \min_{k} Q_{ik}^{\Theta'} - \beta \min_{k} \widetilde{Q}_{ik}^{\Theta'} \right| \le \beta \left| \max_{r,s,\theta} \beta Q_{rs}^{\theta} - \widetilde{Q}_{rs}^{\theta} \right|. \tag{E.11}$$

Combining (E.8) and (E.11), we have

$$\begin{aligned}\left|T\left(Q_{ij}^{\Theta}\right)-T\left(\tilde{Q}_{ij}^{\Theta}\right)\right| &\leq \frac{1-\beta}{2}\max_{p,q,\theta}\left|Q_{pq}^{\theta}-\tilde{Q}_{pq}^{\theta}\right| \\ &\quad+\beta\left|\max_{r,s,\theta}\beta Q_{rs}^{\theta}-\tilde{Q}_{rs}^{\theta}\right| \qquad \text{(E.12)} \\ &\leq \frac{1+\beta}{2}\max_{p,q,\theta}\left|Q_{pq}^{\theta}-\tilde{Q}_{pq}^{\theta}\right|,\end{aligned}$$

which implies

$$\left\|T\left(Q_{ij}^{\Theta}\right)-T\left(\tilde{Q}_{ij}^{\Theta}\right)\right\|_{\infty} \leq \frac{1+\beta}{2}\left\|Q_{ij}^{\Theta}-\tilde{Q}_{ij}^{\Theta}\right\|_{\infty}. \qquad \text{(E.13)}$$

Therefore, T is a contraction mapping under the norm $\|\cdot\|_{\infty}$. This concludes the proof. □

Remark 2. Note that, in contrast to the stochastic approximation approach for the proof of the convergence in the complete observation case, we used a different approach to prove the convergence of the learning with partial observations since it is difficult to apply the stochastic approximation in the partial observation case. Although the stochastic approximation approach is slightly more complicated, we can find a finite value for γ in (D.14) to assure the convergence. For the contraction mapping approach, we are still unable to find such a finite value for γ.

Acknowledgment

This work was supported by the National Science Foundation under grant CCF-0830451.

References

[1] J. Mitola III, "Cognitive radio for flexible mobile multimedia communications," *Mobile Networks and Applications*, vol. 6, no. 5, pp. 435–441, 2001.

[2] J. Mitola III, *Cognitive Radio*, Licentiate Proposal, KTH, Stockholm, Sweden, 1998.

[3] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 79–89, 2007.

[4] K. Kim, I. A. Akbar, K. K. Bae, J.-S. Um, C. M. Spooner, and J. H. Reed, "Cyclostationary approaches to signal detection and classification in cognitive radio," in *Proceedings of the 2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pp. 212–215, April 2007.

[5] H. Li, C. Li, and H. Dai, "Quickest spectrum sensing in cognitive radio," in *Proceedings of the 42nd Annual Conference on Information Sciences and Systems (CISS '08)*, pp. 203–208, Princeton, NJ, USA, March 2008.

[6] A. Ghasemi and E. S. Sousa, "Collaborative spectrum sensing for opportunistic access in fading environments," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 131–136, November 2005.

[7] C. Kloeck, H. Jaekel, and F. Jondral, "Multi-agent radio resource allocation," *Mobile Networks and Applications*, vol. 11, no. 6, pp. 813–824, 2006.

[8] D. Niyato, E. Hossain, and Z. Han, "Dynamics of multiple-seller and multiple-buyer spectrum trading in cognitive radio networks: a game-theoretic modeling approach," *IEEE Transactions on Mobile Computing*, vol. 8, no. 8, pp. 1009–1022, 2009.

[9] F. F. Kuo, "The ALOHA system," in *Computer Networks*, pp. 501–518, Prentice-Hall, Englewood Cliffs, NJ, USA, 1973.

[10] L. Buşoniu, R. Babuška, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man and Cybernetics Part C*, vol. 38, no. 2, pp. 156–172, 2008.

[11] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, The MIT Press, Cambridge, Mass, USA, 1998.

[12] J. Robinson, "An iterative method of solving a game," *The Annals of Mathematics*, vol. 54, no. 2, pp. 296–301, 1969.

[13] J. Hu and M. P. Wellman, "Multiagent reinforcement learning: theoretical framework and an algorithm," in *Proceedings of the 15th International Conference on Machine Learning (ICML '98)*, pp. 242–250, July 1998.

[14] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, New York, NY, USA, 2003.

[15] J. Liu, Y. Yi, A. Proutiere, M. Chiang, and H. V. Poor, "Towards utility-optimal random access without message passing," *Wireless Communications and Mobile Computing*, vol. 10, no. 1, pp. 115–128, 2010.

[16] Y. Yi, G. de Veciana, and S. Shakkottai, "MAC scheduling with low overheads by learning neighborhood contention patterns," submitted to *IEEE/ACM Transactions on Networking*.

[17] F. Fu and M. van der Schaar, "Learning to compete for resources in wireless stochastic games," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 1904–1919, 2009.

[18] Z. Han, C. Pandana, and K. J. K. Liu, "Distributive opportunistic spectrum access for cognitive radio using correlated equilibrium and no-regret learning," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '07)*, pp. 11–15, March 2007.

[19] M. van der Schaar and F. Fu, "Spectrum access games and strategic learning in cognitive radio networks for delay-critical applications," *Proceedings of the IEEE*, vol. 97, no. 4, pp. 720–739, 2009.

[20] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.

[21] B. Wang, K. J. R. Liu, and T. C. Clancy, "Evolutionary game framework for behavior dynamics in cooperative spectrum sensing," in *Proceedings of IEEE Conference on Global Communications (Globecom '08)*, pp. 3123–3127, New Orleans, La, USA, November-December 2008.

[22] P. D. Sutton, K. E. Nolan, and L. E. Doyle, "Cyclostationary signatures in practical cognitive radio applications," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 13–24, 2008.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: A Introduction*, The MIT Press, Cambridge, Mass, USA, 1998.

[24] A. Metrick and B. Polak, "Fictitious play in 2×2 games: a geometric proof of convergence," *Economic Theory*, vol. 4, no. 6, pp. 923–933, 1994.

[25] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 2, pp. 400–407, 1951.

[26] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Machine Learning*, vol. 16, no. 3, pp. 185–202, 1994.