Reciprocity Inspired Learning for Opportunistic Spectrum Access in Cognitive Radio Networks

(Invited Paper)

Xianfu Chen[♮], Tao Chen[♮], Wei Cheng[♭], and Honggang Zhang[†‡]

[♮]VTT Technical Research Centre of Finland, P. O. Box 1100, FI-90571 Oulu, Finland
[♭]Department of Computer Science, The George Washington University, Washington DC, USA
[†]Université Européenne de Bretagne (UEB) and Supélec, CS 47601, 35576 Cesson-Sévigné Cedex, France
[‡]Department of ISEE, Zhejiang University, Zheda Road 38, Hangzhou 310027, China
Email: {xianfu.chen, tao.chen}@vtt.fi, {wcheng, cheng}@gwu.edu, honggang.zhang@supelec.fr

***Abstract*—This paper addresses opportunistic spectrum access (OSA) in non-cooperative cognitive radio networks (CRNs). The selfish behaviors of the secondary users (SUs) will cause a CRN to collapse. The SUs are thus enabled to build beliefs about how other SUs would respond to their decision makings. The interaction among the SUs is modeled as a stochastic learning process. In this way, each SU can independently learn the behaviors of the competitors, optimize the OSA strategies, and finally achieve the goal of reciprocity. Two learning algorithms are proposed to stabilize the stochastic CRNs, the convergence properties of which are also proven theoretically. Simulation results validate the performance of the proposed results, and show that the achieved system performance outperforms some existing protocols.**

I. Introduction

Cognitive radio (CR) has shown the effectiveness in bridging the enormous gulf in time and space between the regulation and the potential spectrum efficiency [1]. In this paper, we focus on designing opportunistic spectrum access (OSA) protocols, based on which each secondary user (SU) independently decides which channels to access in different time slots, in order to efficiently utilize the licensed spectrum when the primary users (PUs) are "sleeping". In the presence of multiple SUs, the OSA protocols must also account for the possibility of competition among users over the same channel. In general, game theoretic approaches have been exploited to determine the communication resources of multiple interacting users [2], [3].

Game theory bases its solution on the concept of equilibrium [4]. Users behaving within an equilibrium are often explained in terms of their beliefs about the strategies of the competitors. This paper is concerned with developing distributed learning algorithms for OSA in cognitive radio networks (CRNs) from not only a game-theoretic, but also a learning perspective. The distinction between learning and non-learning users is simply that the former change their beliefs, whereas the latter's beliefs are static. A variety of learning schemes have been applied to CR scenarios, as surveyed in [5]. Particularly, reinforcement learning (RL) [6] has been used to study intelligent decision makings in heterogeneous CRNs, where every SU has to adapt to both the PUs' behaviors and other SUs' decisions.

The main challenge of deploying distributed learning algorithms in non-cooperative CRNs is the problem of action coordination. Centralized schemes may be applied to ensure the optimal decision makings, however, they are not always applicable in dynamic CRNs. Hence, our goal is to explore the decentralized spectrum access strategies for the competitive SUs that generate near-optimal decisions. The conjectural variation model introduced by Bowley [7] is adopted to encourage potential cooperation among the SUs. This model enables the SUs to form beliefs about how other SUs react to their strategy changes. Specifically, by implementing such a strategic interaction mechanism, the SUs will no longer behave myopically.

The rest of this paper is organized as follows. In the following section, we formulate the problem of OSA in a CRN. In Section III, we first introduce the belief model and then propose two distributed learning algorithms to achieve optimal spectrum access strategies for the SUs. The case where the PUs' statistical behavior patterns are unknown is investigated in Section IV. Section V provides the numerical results to verify the validity, and efficiency of the proposed learning protocols. Finally, we present in Section VI a conclusion of this paper.

II. Problem Description

Consider a CRN consisting of a set $\mathcal{M} = \{1, \ldots, M\}$ of licensed channels with equal bandwidth B[1], a set $\mathcal{N} = \{1, \ldots, N\}$ of SUs opportunistically access these channels when they are not occupied by the PUs. All users in the network are operated in a time-slotted fashion. During each time slot, the PUs transmit over channel $m \in \mathcal{M}$ with probability $1 - \omega_m (0 \leq \omega_m \leq 1)$. The state for channel m at time t is denoted by $S_m(t) \in \{1(\text{idle}), 0(\text{occupied})\}$. For a

[1]The case for channels with different bandwidths B_m can be easily transformed to an equivalent problem, in which channel m is idle with probability $\omega_m^{equ} = \omega_m B_m / B_{\max}$. Herein, $B_{\max} = \max_{m \in \mathcal{M}} B_m$.

CROWNCOM 2013, July 08-10, Washington DC, United States

DOI 10.4108/icst.crowncom.2013.252895

given $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)$, $S_m(t)$ are independent for each m and t.

Assume that each SU is capable of accessing only one channel at each time slot. At the beginning of time t, SU $n \in \mathcal{N}$ selects a channel $a_n(t) \in \mathcal{M}$ according to its spectrum access strategy $\pi_n \in \Pi_n$ to sense (access). A strategy π_n is defined to be a probability vector $\pi_n = (\pi_n(1), \ldots, \pi_n(M))$, where $\pi_n(m)$ means the probability with which SU n accesses channel m. The outcome of spectrum sensing is supposed to be error-free. If the sensing result indicates that $S_{a_n(t)}(t) = 1$, the SUs selecting channel $a_n(t)$ compete to transmit; otherwise, SU n will wait until next time slot and choose possibly a different channel to access. The collision model is applied, under which if two or more SUs transmit over the same idle channel then none of the transmissions are successful. At the end of the same time slot, SU n receives from its receiver an acknowledgement $Z_{n,a_n(t)}(t)$ that equals 1 if the transmission went through and equals 0 otherwise. The number of bits that SU n is able to send in time t is[2]

$$W_n(t) = BS_{a_n(t)}(t)Z_{n,a_n(t)}(t).$$

Obviously, $W_n(t)$ is a random variable that depends on the PUs' behaviors and, more importantly for us, the spectrum access strategies implemented by the SUs.

The overarching target in the rest of this paper is to design the strategies $\{\pi_n \mid n \in \mathcal{N}\}$ that maximize the utility

$$\begin{aligned} u_n(\pi_n, \boldsymbol{\pi}_{-n}) &= \mathrm{E}\left\{BS_{a_n(t)}(t)Z_{n,a_n(t)}(t)\right\} \\ &= \sum_{m\in\mathcal{M}} B\omega_m\pi_n(m) \prod_{l\in\mathcal{N}\setminus\{n\}} (1-\pi_l(m)), \end{aligned}$$

for $\forall n \in \mathcal{N}$, where $-n$ denote all the other SUs in $\mathcal{N}$ except SU n. The OSA in non-cooperative CRNs can be formally defined by the tuple $\mathcal{G} = \langle \mathcal{N}, \{\Pi_n\}, \{u_n\}\rangle$, for which an important solution is the Nash equilibrium (NE).

Definition. *A strategy profile $(\pi_n, \boldsymbol{\pi}_{-n})$ constitutes an NE of $\mathcal{G}$ if $u_n(\pi_n, \boldsymbol{\pi}_{-n}) \le u_n(\pi_n', \boldsymbol{\pi}_{-n})$, for all $\pi_n' \in \Pi_n$ and $n \in \mathcal{N}$.*

It's shown in [8] that at the NE, each SU n selects channel m with probability $\pi_n(m) = \omega_m / \sum_{i\in\mathcal{M}} \omega_i$. But if one SU deviates from this symmetric strategy[3], it can achieve better performance. Such selfish behaviors may cause significant reduction in the overall network performance.

III. Learning with Dynamic Conjectures

To promote cooperation among the non-cooperative SUs, we propose to use the idea of dynamic conjectures [9]: each SU conjectures that its strategy changes will influence other SUs' contemporaneous access decisions.

A. The Belief Function

For the purpose of utility maximization, each SU $n \in \mathcal{N}$ forms conjectures about the contention measures $c_{n,m} = \prod_{l\in\mathcal{N}\setminus\{n\}} (1-\pi_l(m))$, for all $m \in \mathcal{M}$. Specifically, let

$$\tilde{c}_{n,m}(\pi_n(m)) = \bar{c}_{n,m} - \delta_{n,m}(\pi_n(m) - \bar{\pi}_n(m)), \quad (1)$$

be a belief of $c_{n,m}$. Herein, $\delta_{n,m} > 0$ is the belief factor, and $\bar{c}_{n,m}$ and $\bar{\pi}_n(m)$ are called the reference points [10]. The belief functions deployed by the SUs are based on the concept of reciprocity, which refers to the interaction mechanism that if the SUs realize the probabilities of interacting with each other in the future is high, they will consider their influence on other SUs' strategies. Otherwise, all SUs will act myopically, leading to terrible network performance reduction.

Taking into account the conjectures about the strategies of other competing SUs, SU n's utility function thus becomes

$$\begin{aligned} u_n(\pi_n, \tilde{c}_n(\pi_n)) = \sum_{m\in\mathcal{M}} & B\omega_m\pi_n(m)[\bar{c}_{n,m} \\ & -\delta_{n,m}(\pi_n(m) - \bar{\pi}_n(m))], \end{aligned}$$

where $\tilde{c}_n(\pi_n) = (\tilde{c}_{n,1}(\pi_n(1)), \ldots, \tilde{c}_{n,M}(\pi_n(M)))$. If the OSA is repeated over time, the SUs can learn from their prior observations. Let $c_{n,m}^t$, $\tilde{c}_{n,m}^t$, $\pi_n^t(m)$ be SU n's contention measure, belief function and spectrum access strategy at time t. We propose that each SU n sets its reference points at time slot t to be $c_{n,m}^{t-1}$ and $\pi_n^{t-1}(m)$. Therefore, SU n's utility function at time t is

$$\begin{aligned} u_n(\pi_n^t, \tilde{c}_n^t(\pi_n^t)) = \sum_{m\in\mathcal{M}} & B\omega_m\pi_n^t(m)\big[c_{n,m}^{t-1} \\ & -\delta_{n,m}\left(\pi_n^t(m) - \pi_n^{t-1}(m)\right)\big]. \end{aligned} \quad (2)$$

An intuitive explanation for (2) is that, each SU n believes a change of $\pi_n^t(m) - \pi_n^{t-1}(m)$ in its spectrum access strategy at time t will induce a change of $\delta_{n,m}\left(\pi_n^t(m) - \pi_n^{t-1}(m)\right)$ in the contention measure exactly corresponding to the strategies of other SUs.

B. Best-response Learning

Along with the previous discussions, the SUs maximize their utilities in the spirit of best-response to the dynamics of the learning process.

1) The Best-response Strategies: Treat $c_{n,m}^1$ and $\pi_n^1(m)$ $(n \in \mathcal{N}, m \in \mathcal{M})$ as initial parameters, we then find for each SU n an optimal strategy plan that consists of a sequence of single time slot strategy functions

$$\pi_n^t(m) = f_{n,m}\left(\pi_n^{t-1}(m), c_{n,m}^{t-1}\right), \text{ for } t = 2, 3, \ldots$$

which represent the best-response behavior for SU n at any time slot t given its belief factors $\delta_{n,m}$.

Theorem 1. *The best-response spectrum access strategy for each SU $n \in \mathcal{N}$ is given by (3), where $\lambda_n^t \ge 0$ is chosen such that $\sum_{m\in\mathcal{M}} \pi_n^t(m) = 1$. Here, $[x]_a^b$ with $b > a$, denotes the Euclidean projection of x onto the interval $[a, b]$.*

Proof: For each SU $n \in \mathcal{N}$, the best-response strategy π_n^t at time slot t maximizes its utility $u_n(\pi_n^t, \tilde{c}_n^t(\pi_n^t))$, that is,

$$\begin{aligned} &\underset{\pi_n^t\in\Pi_n}{\text{maximize}} \quad u_n(\pi_n^t, \tilde{c}_n^t(\pi_n^t)) \\ &\text{s.t.} \quad \text{C1: } \sum_{m\in\mathcal{M}} \pi_n^t(m) = 1, \text{ and C2: } \pi_n^t(m) \ge 0, \forall m \in \mathcal{M}. \end{aligned}$$

It's easy to check that the above optimization problem is convex with linear constraints C1-C2. Thus the Lagrangian

[2] Suppose B bits can be transmitted over a channel during one time slot.

[3] A common strategy π implemented by all SUs is defined to be a symmetric strategy, i.e., $\pi = \pi_1 \ldots = \pi_N$.

$$\pi_n^t(m) = f_{n,m}\left(\pi_n^{t-1}(m), c_{n,m}^{t-1}\right) = \begin{cases} \left[\frac{1}{2}\pi_n^{t-1}(m) + \frac{1}{2\delta_{n,m}}c_{n,m}^{t-1} + \frac{\lambda_n^t}{2B\delta_{n,m}\omega_m}\right]_0^1, & \text{if } \omega_m > 0, \\ 0, & \text{if } \omega_m = 0. \end{cases} \quad (3)$$

function for SU n can be written as

$$L_n\left(\pi_n^t, \tilde{c}_n^t\left(\pi_n^t\right), \lambda_n^t, \gamma_n^t\right) = u_n\left(\pi_n^t, \tilde{c}_n^t\left(\pi_n^t\right)\right) + \lambda_n^t\left(\sum_{m\in\mathcal{M}}\pi_n^t(m) - 1\right) + \sum_{m\in\mathcal{M}}\gamma_{n,m}^t\pi_n^t(m), \quad (4)$$

where λ_n^t and $\gamma_{n,m}^t$ are Lagrangian multipliers. The Karush-Kuhn-Tucker (K.K.T.) conditions [11] are given by

$$\begin{aligned} \frac{\partial L_n\left(\pi_n^t, \tilde{c}_n^t\left(\pi_n^t\right), \lambda_n^t, \gamma_n^t\right)}{\partial \pi_n^t(m)} &= B\omega_m[-2\delta_{n,m}\pi_n^t(m) + c_{n,m}^{t-1} \\ &\quad +\delta_{n,m}\pi_n^{t-1}(m)] + \lambda_n^t + \gamma_{n,m}^t \\ &= 0, \forall m \in \mathcal{M} \\ \pi_n^t(m) &\geq 0, \forall m \in \mathcal{M} \\ \gamma_{n,m}^t\pi_n^t(m) &= 0, \forall m \in \mathcal{M} \\ \sum_{m\in\mathcal{M}}\pi_n^t(m) &= 1. \end{aligned}$$

It's then straightforward to have the result in Theorem 2. This concludes the proof. ■

The detailed description of the best-response learning for OSA in CRNs is summarized in Algorithm 1.

Algorithm 1

Initialization:

(a) $t = 1$, initialize π_n^t and $\delta_{n,m} > 0$, for $\forall n \in \mathcal{N}$ and $\forall m \in \mathcal{M}$; choose channels according to π_n^t, for all SUs.

End Initialization

Learning:

(b) At time slot t, each SU $n \in \mathcal{N}$

($b.1$) senses and competes for the selected channel, and transmits B bits if successfully occupying the channel;

($b.2$) broadcasts strategy information π_n^t to other SUs.

(c) Set $t \leftarrow t + 1$.

(d) For all $n \in \mathcal{N}$ and $m \in \mathcal{M}$, do (3).

(e) SU n accesses channel m at time t with probability $\pi_n^t(m)$.

End Learning

2) Network Stability: Next, we are concerned with the stability of this algorithm. The Theorem 2 shows that the stochastic network is stable if each function $f_{n,m}$ is a contraction mapping. Denote $\boldsymbol{\pi}^t = (\pi_1^t, \ldots, \pi_N^t)$ as the strategy profile of all SUs at time slot t.

Theorem 2. *Suppose that the belief factor in the belief function (1) satisfies $\delta_{n,m} \geq N - 1$, for $\forall n \in \mathcal{N}$ and $\forall m \in \mathcal{M}$, the Algorithm 1 converges to a unique stable state.*

Proof: Without loss of generality, we assume that $\omega_m > 0$, for $\forall m \in \mathcal{M}$. At the moment, the best-response strategy in (3) can be rewritten as

$$f_{n,m}\left(\pi_n^{t-1}(m), c_{n,m}^{t-1}\right) = \frac{\pi_n^{t-1}(m)}{2} + \frac{\prod_{l\in\mathcal{N}\setminus\{n\}}\left(1 - \pi_l^{t-1}(m)\right)}{2\delta_{n,m}} + \frac{\lambda_n^t}{2B\delta_{n,m}\omega_m}.$$

We now proceed to prove that $f_{n,m}$ is a contraction mapping if the condition in Theorem 2 occurs. Suppose that $\{\boldsymbol{\pi}^t\}_{t=1}^{\infty}$ and $\{\widetilde{\boldsymbol{\pi}}^t\}_{t=1}^{\infty}$ are two strategy profile sequences. Let $J_{p,q}(1 \leq p, q \leq N \times M)$ denote the element at row p and column q of the Jacobian matrix $\mathbf{J}$ of function $f_{n,m}$, then

$$J_{p,q} = \begin{cases} \frac{1}{2}, & \text{if } p = q; \\ -\frac{\prod_{k\in\mathcal{N}\setminus\{n,l\}}\left(1 - \pi_k^{t-1}(m)\right)}{2\delta_{n,m}}, & \text{if } m = j \text{ and } p \neq q; \\ 0, & \text{otherwise,} \end{cases}$$

where $n = \lceil p/M \rceil$, $m = p - M\lfloor p/M \rfloor$, $l = \lceil q/M \rceil$, and $j = q - M\lfloor q/M \rfloor$. Consider $\|\cdot\|_\infty$ for $\mathbf{J}$, we derive

$$\begin{aligned} \|\mathbf{J}\|_\infty &= \max_{p\in\{1,\ldots,N\times M\}}\sum_{q\in\{1,\ldots,N\times M\}}|J_{p,q}| \\ &= \frac{1}{2} + \max_{p\in\{1,\ldots,N\times M\}}\sum_{q\in\{q|m=j\}\setminus\{p\}}\frac{1}{2\delta_{n,m}} \\ &\quad \times \prod_{k\in\mathcal{N}\setminus\{n,l\}}\left(1 - \pi_k^{t-1}(m)\right) \\ &\leq \frac{1}{2} + \frac{1}{2}\max_{p\in\{1,\ldots,N\times M\}}\sum_{q\in\{q|m=j\}\setminus\{p\}}\frac{1}{\delta_{n,m}} \\ &= \frac{1}{2} + \frac{1}{2}\max_{p\in\{1,\ldots,N\times M\}}\frac{N-1}{\delta_{n,m}}. \end{aligned}$$

If $\delta_{n,m} \geq N - 1$, there $\exists\varepsilon \in (0, 1)$, such that $\|\mathbf{J}\|_\infty = 1 - \varepsilon < 1$, i.e., $\|\boldsymbol{\pi}^t - \widetilde{\boldsymbol{\pi}}^t\|_\infty \leq \|\mathbf{J}\|_\infty\left\|\boldsymbol{\pi}^{t-1} - \widetilde{\boldsymbol{\pi}}^{t-1}\right\|_\infty$. Thus $\{\boldsymbol{\pi}^t\}_{t=1}^{\infty}$ converges to the unique stable state by the contraction mapping theorem [12]. ■

The stability of the stochastic network requires joint condition on $\delta_{n,m} \geq N - 1$, for all $n \in \mathcal{N}$ and $m \in \mathcal{M}$. We may think that our belief model in (1) and the dynamics it generates are much less appealing if the condition does not hold. However, if the network converges to a stable state, the SUs' beliefs eventually cease to be falsified and our approach is justified.

C. Gradient Ascent Learning

A series of learning algorithms are derived based on the stochastic gradient ascent [13]. Since they are derived from first principles with function estimators in mind, they guarantee the convergence to local maximal. In addition to the convergence property, the gradient ascent technique has made it possible to convert an inelegant, inconvenient learning algorithm into a much simpler and more easily analyzed algorithm. These significant practical benefits motivate the development of gradient ascent learners [14].

At the beginning of each time slot, each SU updates its spectrum access strategies gradually in the ascent direction of its conjectural utility defined by (4). More specifically, at time t, SU $n \in \mathcal{N}$ updates its strategy according

$$\pi_n^t(m) = f_{n,m}\left(\pi_n^{t-1}(m), c_{n,m}^{t-1}\right) = \pi_n^{t-1}(m) + \kappa_n \frac{\partial L_n\left(\pi_n, \tilde{c}_n^t(\pi_n), \lambda_n^t, \gamma_n^t\right)}{\partial \pi_n(m)}\Big|_{\pi_n(m)=\pi_n^{t-1}(m),} \quad \text{for all } m \in \mathcal{M}. \tag{5}$$

to (5), where $\kappa_n > 0$ is the step size. So effectively, $\partial L\left(\pi_n, \tilde{c}_n^t, \lambda_n^t, \gamma_n^t\right)/\partial \pi_n(m)|_{\pi_n(m)=\pi_n^{t-1}(m)} > 0$ means the probability of choosing a good channel increases by a rate. Likewise, the probability of choosing a bad channel decreases by a rate. Substituting (4) into (5), we may have

$$\pi_n^t(m) = \left[\pi_n^{t-1}(m) + \kappa_n\left[B\omega_m\left(c_{n,m}^{t-1} - \delta_{n,m}\pi_n^{t-1}(m)\right) + \lambda_n^t\right]\right]_0^1, \tag{6}$$

where λ_n^t satisfies $\sum_{m\in\mathcal{M}} \pi_n^t(m) = 1$. The Algorithm 2 is thus proposed by replacing (3) in Algorithm 1 with (6).

Theorem 3. *Suppose that the belief factor $\delta_{n,m}$ in belief function (1) and the step size κ_n in (6) satisfy $\delta_{n,m} \geq N-1$, and $0 < \kappa_n \leq 1/(B\omega_m\delta_{n,m})$, for $\forall n \in \mathcal{N}$ and $\forall m \in \mathcal{M}$, the dynamics of the Algorithm 2 converge.*

Proof: The proof can be obtained similarly as in the proof of Theorem 2, and is thus omitted. ■

We may find that given the same belief factors, both Algorithms 1 and 2 exhibit similar convergence properties, if the step size in Algorithm 2 is small enough. In practice, the best-response strategies often lead to large fluctuations that may cause temporary system instability. On the other hand, by setting the step size sufficiently small, the gradient ascent learning experiences a more smoother trajectory.

IV. The Unknown $\boldsymbol{\omega}$ Case

For the OSA discussed in previous section, each SU is supposed to have the perfect knowledge of $\boldsymbol{\omega}$. However, under many realistic circumstances, the $\boldsymbol{\omega}$ is initially unknown to all the SUs and in addition to the competition among the users, is learned independently over time utilizing the past access decisions. Combining the results in Section III, we design the following access protocol in CRNs for the unknown $\boldsymbol{\omega}$ case.

Let $X_{n,m}(t)$ denote the number of times that channel $m \in \mathcal{M}$ is chosen for spectrum sensing by SU $n \in \mathcal{N}$ in t time slots. SU n records all these decisions in a vector $\mathbf{X}_n(t) \triangleq (X_{n,1}(t), \ldots, X_{n,M}(t))$. At the same time, SU n maintains another vector $\mathbf{Y}_n(t) \triangleq (Y_{n,1}(t), \ldots, Y_{n,M}(t))$ where the sensing results are kept. Herein, $Y_{n,m}(t)$ indicates the number of times that channel m is sensed to be idle by SU n until time t. After every step of spectrum sensing, the vectors $\mathbf{X}_n(t)$ and $\mathbf{Y}_n(t)$ are updated accordingly. Each SU n estimates the value of ω_m in time slot $t+1$ through

$$\tilde{\omega}_{n,m}^t = Y_{n,m}(t)/X_{n,m}(t). \tag{7}$$

Regardless the sensing outcomes, we set $\tilde{\omega}_{n,m}^t = 0$ when $X_{n,m}(t) = 0$, for all $n \in \mathcal{N}$ and $m \in \mathcal{M}$. The conjectural utility function in (2) is thus approximated as

$$\tilde{u}_n^t\left(\pi_n^t, \tilde{c}_n^t\left(\pi_n^t\right)\right) = \sum_{m\in\mathcal{M}} B\tilde{\omega}_{n,m}^{t-1}\pi_n^t(m)\left[c_{n,m}^{t-1} - \delta_{n,m}\left(\pi_n^t(m) - \pi_n^{t-1}(m)\right)\right]. \tag{8}$$

The intuition behind (8) is that as time goes by, the estimated $\tilde{\omega}_{n,m}^{t-1}$ will finally converge to ω_m in probability, which implies that the unknown $\boldsymbol{\omega}$ case will eventually reduce to the scenarios we discussed in Section III.

The following Algorithm 3 is designed to achieve the optimal spectrum access strategies for all SUs.

Algorithm 3

Initialization:

(a) $t = 1$, initialize $\kappa_n > 0$, π_n^t, and $\delta_{n,m} > 0$, for all $n \in \mathcal{N}$ and $m \in \mathcal{M}$; choose a channel according to π_n^t for each SU.

End Initialization

Learning:

(b) At time slot t, each SU $n \in \mathcal{N}$

- $(b.1)$ senses and competes for the selected channel, and transmits B bits if successfully occupying the channel;
- $(b.2)$ records the sensing decision and sensing result in vectors $\mathbf{X}_n(t)$ and $\mathbf{Y}_n(t)$;
- $(b.3)$ computes $\tilde{\omega}_{n,m}^t$ according to (7) for all $m \in \mathcal{M}$;
- $(b.4)$ broadcasts strategy information π_n^t to other SUs.

(c) Set $t \leftarrow t+1$.

(d) For all $n \in \mathcal{N}$ and $m \in \mathcal{M}$, do $\pi_n^t(m) \leftarrow$

$$\begin{cases} \left[\dfrac{\pi_n^{t-1}(m)}{2} + \dfrac{c_{n,m}^{t-1}}{2\delta_{n,m}} + \dfrac{\lambda_n^t}{2B\delta_{n,m}\tilde{\omega}_{n,m}^{t-1}}\right]_0^1, & \text{if } \omega_m > 0; \\ 0, & \text{if } \omega_m = 0, \end{cases}$$

or $\pi_n^t(m) \leftarrow$

$$\left[\pi_n^{t-1}(m) + \kappa_n\left[B\tilde{\omega}_{n,m}^{t-1}\left(c_{n,m}^{t-1} - \delta_{n,m}\pi_n^{t-1}(m)\right) + \lambda_n^t\right]\right]_0^1.$$

(e) SU n accesses channel m at time t with probability $\pi_n^t(m)$.

End Learning

V. Numerical Results

This section presents experiments to evaluate the performance of the algorithms developed in this paper. In the following simulations, we set $B = 1$Hz.

A. The Known $\boldsymbol{\omega}$ Case

We first consider a simple CRN where $N = 2$ SUs compete for accessing $M = 2$ licensed channels with idle probabilities $\omega_1 = 0.5$ and $\omega_2 = 0.9$. Denote the probability of SU 1 selecting channel 1 at time slot t by $\alpha(t)$ and selecting channel 2 by $1-\alpha(t)$. Similarly, SU 2 selects channel 1 at time t with probability $\beta(t)$, then selects channel 2 with probability $1-\beta(t)$. The strategies are initialized to be $(\alpha(1), \beta(1)) = (0.6, 0.6)$, and the belief factors $\delta_{n,m}$ are uniformly distributed between 4 and 8. For simplicity, the step size in Algorithm 2 is set to be $\kappa_n = 0.06$ for both of the two SUs.

Fig. 1 and Fig. 2 compare the learning trajectories of both Algorithm 1 and Algorithm 2. We can see from the curves that the best-response learning approach converges in around 10 iterations, while the gradient ascent learning approach

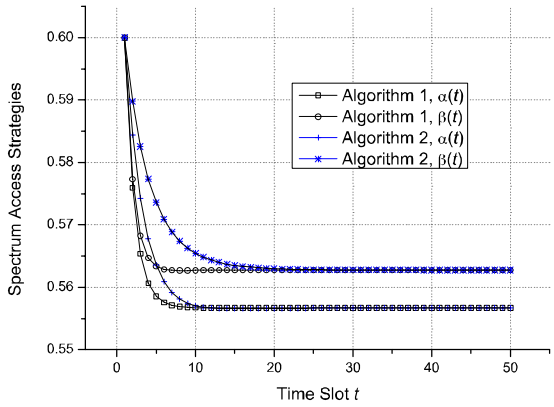

Fig. 1. The strategy dynamics of Algorithms 1 and 2 with initial strategies $(\alpha(1), \beta(1)) = (0.6, 0.6)$.

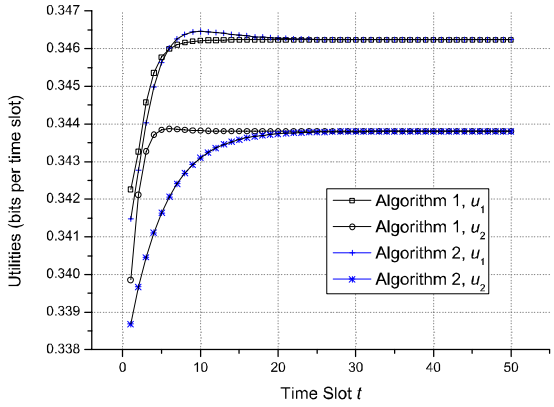

Fig. 2. The utility dynamics of Algorithms 1 and 2 with initial strategies $(\alpha(1), \beta(1)) = (0.6, 0.6)$.

experiences a more smoother trajectory and derives the same optimal access strategies after about 22 iterations. Another observation is that, whenever we generate the initial spectrum access strategies, the network state achieved by the proposed algorithms is independent of these initial values.

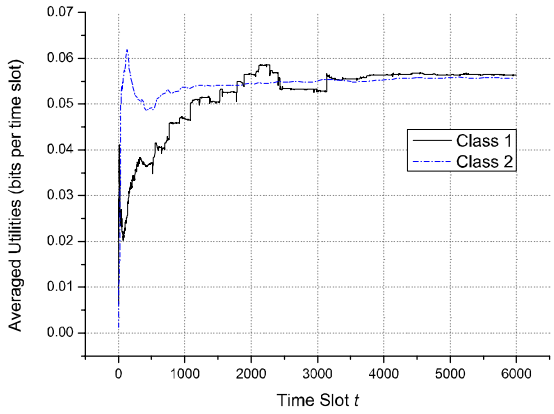

Fig. 3. The averaged utility dynamics of the SUs of each class.

B. The Unknown $\boldsymbol{\omega}$ Case

Next we show some numerical results to consider the impact of the learning duration on the learning protocol developed in Section IV. Fig. 3 dictates the averaged utility performance of the SUs, implementing the Algorithm 3. In this experiment, we suppose that there are $N = 30$ SUs and $M = 10$ licensed channels with randomly generated idle probabilities $\boldsymbol{\omega} = (0.4106, 0.5306, 0.9660, 0.9200, 0.1832, 0.3151, 0.3253, 0.3768, 0.0250, 0.6720)$. The belief factors $\delta_{n,m}$ are uniformly distributed in $[30, 36]$, and the step size is chosen to be $\kappa_n = 0.008$ for all SUs. Additionally, to better illustrate how SUs adaptively alter their learning protocols, it is assumed that there are two user classes, each of which consists of 15 SUs. Class 1 and class 2 choose the best-response learning approach and the gradient ascent learning approach, respectively. The results verify that the averaged utilities of each class converge and the utilities are close to each other after a period of learning. And the relatively smoother convergence of the Algorithm 2 is also justified in Fig. 3.

Spectrum sensing errors are inevitable in practical applications, which means that the design of spectrum access strategies for the purpose of optimal spectrum utilization should take into account the maximum collision probability chosen to protect the PUs' performances. Once a collision occurs, the PUs transmitting over the channel should send out a warning tone to the SUs. Let $P_{n,m}$ be the collision probability observed by SU n over channel m, we have (9). To ensure PUs' QoS requirements, we set $P_{n,m} \leq \zeta$, for all $n \in \mathcal{N}$ and $m \in \mathcal{M}$, where ζ is the QoS threshold that the PUs can tolerate. Using similar simulation environment as in Fig. 3, Fig. 4 shows the results for SUs of the two classes with parameters $\zeta = 0.1$ and $\boldsymbol{\omega} = (0.0960, 0.6003, 0.4323, 0.0570, 0.3696, 0.6253, 0.7085, 0.8735, 0.2521, 0.6628)$. We can find that the averaged utility performance still converges to the optimum even though there exist sensing errors.

C. Performance Comparison

Finally, to further verify the performance of the proposed algorithms in this paper, we compare them with two existing OSA protocols:

1) Multi-agent Q-learning protocol: In [15], a distributed multi-agent Q-learning OSA protocol was developed by generalizing single-agent Q-learning to the multi-agent scenarios. More specifically, each SU $n \in \mathcal{N}$ maintains a Q-value $Q_n(m)$, representing the expected bits that can be transmitted by accessing channel $m \in \mathcal{M}$. Once channel m is chosen by SU n at time t, the Q-value is updated according to

$$Q_n^{t+1}(m) = (1 - \alpha^t) Q_n^t(m) + \alpha^t W_n(t) I(a_n(t) = m),$$

where $\alpha^t \in [0, 1)$ is the learning rate and I is a characteristic function for the event that channel m is selected at time t. The strategy is updated based on the Boltzmann distribution, i.e.,

$$\pi_n^{t+1}(m) = \frac{\exp(Q_n^{t+1}(m)/\tau)}{\sum_{i \in \mathcal{M}} \exp(Q_n^{t+1}(i)/\tau)},$$

$$P_{n,m} = \lim_{t\to\infty} \frac{\text{No. of warning tones received by SU } n \text{ over channel } m \text{ in } t \text{ time slots}}{X_{n,m}(t)}. \quad (9)$$

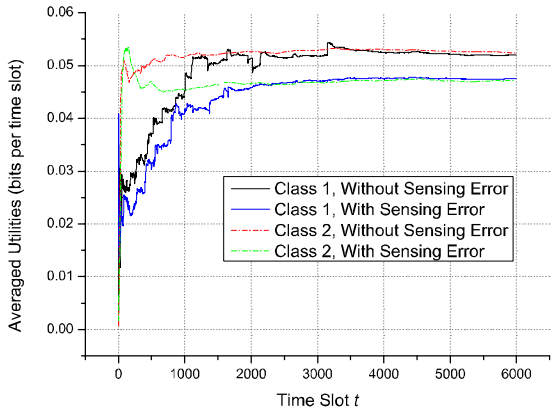

Fig. 4. The averaged utility dynamics of SUs of each class in the presence of sensing errors.

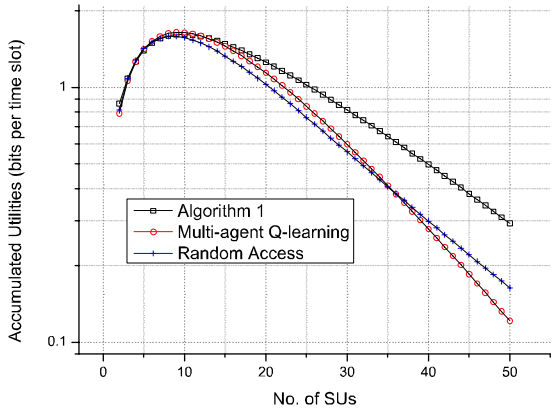

Fig. 5. Comparison of the proposed algorithms with the multi-agent Q-learning protocol and random access protocol.

where τ is the temperature [6]. In simulations, we set $\alpha^t = 0.1/t$ and $\tau = 0.1$.

2) Random access protocol: At each time, SU n accesses channel m according to the NE strategy [8], i.e., $\pi_n(m) = \omega_m / \sum_{i\in\mathcal{M}} \omega_i$, for all $n \in \mathcal{N}$ and $\forall m \in \mathcal{M}$.

We simulate the overall network performance in terms of accumulated utilities. In the experiment, the belief factors $\delta_{n,m}$ are adaptively randomized according to the number of SUs, N, and there are $M = 10$ licensed channels. The idle probabilities of the 10 channels are randomly selected from $(0, 1)$ to be $\boldsymbol{\omega} = (0.2110, 0.2877, 0.5663, 0.5030, 0.4521, 0.2825, 0.4176, 0.5849, 0.5717, 0.3597)$. Fig. 5 depicts the simulation outcomes. From the previous analysis and experimental results, we know that the proposed algorithms converge to the same optimal strategies. Therefore, only Algorithm 1 is examined. It can be found from the curves that when $N \leq M$, the achieved performance of the three protocols are comparable, and increase versus the number of SUs. The reason is that with more SUs, the spectrum opportunities will be better exploited. When $N > M$, the performances decrease as N increases. This is because the contention among the SUs can not be avoided in this case, and the collisions become even severer if more SUs compete to access those channels. Overall, the Algorithm 1 outperforms the multi-agent Q-learning protocol and the random access protocol.

VI. Conclusion

This paper investigates the problem of OSA in non-cooperative CRNs. To prevent the network collapse from the SUs' myopic behaviors, the SUs are enabled to form internal beliefs about how other SUs respond to their strategy variations. Such beliefs reflect an incentive among the SUs to cooperate. Based on the belief model, two learning algorithms are proposed for the SUs to achieve the optimal spectrum access strategies. We also derive the sufficient conditions under which the stochastic network converges to a stable state. The simulation results demonstrate that the proposed algorithms achieve significantly better performance, compared with the multi-agent Q-learning protocol and the random access protocol.

References

[1] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201-220, Feb. 2005.

[2] J. Pang and G. Scutari, "Joint sensing and power allocation in nonconvex cognitive radio games: Quasi-Nash equilibria," *IEEE Trans. Signal Process.*, Early Access Articles, DOI: 10.1109/TSP.2013.2239993.

[3] D. N. Nguyen and M. Krunz, "Spectrum management and power allocation in MIMO cognitive networks," in *Proc. INFOCOM*, Orlando, USA, Mar. 2012.

[4] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA: MIT Press, 1992.

[5] M. Bkassiny, Y. Li, and S. Jayaweera, "A survey on machine-learning techniques in cognitive radios," *IEEE Commun. Surveys Tuts.*, Early Access Articles, DOI: 10.1109/SURV.2012.100412.00017, 2012.

[6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[7] A. L. Bowley, *The Mathematical Groundwork of Economics*. Oxford: Oxford University Press, 1924.

[8] L. Lai, H. E. Gamal, H. Jiang, and H. V. Poor, "Cognitive medium access: Exploration, exploitation and competition," *IEEE Trans. Mobile Comput.*, vol. 10, no. 2, pp. 239-253, Feb. 2011.

[9] J. W. Friedman and C. Mezzetti, "Bounded rationality, dynamic oligopoly, and conjectural variations," *J. Econ. Behav. Organ.*, vol. 49, no. 3, pp. 287-306, Nov. 2002.

[10] A. Jean-Marie and M. Tidball, "Adapting behaviors through a learning process," *J. Econ. Behav. Organ.*, vol. 60, no. 3, pp. 399-422, Jul. 2006.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.

[12] A. Granas and J. Dugundji, *Fixed Point Theory*. New York: Springer-Verlag, 2003.

[13] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley & Sons, 2002.

[14] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proc. ICML*, Washington, USA, Aug. 2003.

[15] H. Li, "Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: A two by two case," in *Proc. SMC*, San Antonio, TX, Oct. 2009.