Evolutionary Game and Learning for Dynamic Spectrum Access

Xu Chen and Jianwei Huang∗

Abstract

Efficient dynamic spectrum access mechanism is crucial for improving the spectrum utilization. In this paper, we consider the dynamic spectrum access mechanism design with both complete and incomplete network information. When the network information is available, we propose an evolutionary spectrum access mechanism. We use the replicator dynamics to study the dynamics of channel selections, and show that the mechanism achieves an equilibrium that is an evolutionarily stable strategy and is also max-min fair. With incomplete network information, we propose a distributed reinforcement learning mechanism for dynamic spectrum access. Each secondary user applies the maximum likelihood estimation method to estimate its expected payoff based on the local observations, and learns to adjust its mixed strategy for channel selections adaptively over time. We study the convergence of the learning mechanism based on the theory of stochastic approximation, and show that it globally converges to an approximate Nash equilibrium. Numerical results show that the proposed evolutionary spectrum access and distributed reinforcement learning mechanisms achieve up to 82% and 70% performance improvement than a random access mechanism, respectively, and are robust to random perturbations of channel selections.

I. INTRODUCTION

Wireless spectrum is scarce and yet severally under-utilized. Recent studies by Federal Communications Commission showed that the current utilization of some licensed spectrum can be as low as 15% even in major urban areas [1]. New dynamic spectrum access techniques enable unlicensed wireless users (secondary users) to opportunistically access the licensed channels owned by legacy spectrum holders (primary users), and thus can significantly improve the spectrum efficiency [2]. There are several key challenges of implementing dynamic spectrum access, e.g., the channel availabilities change (rapidly) over time and selfish secondary users compete for resources in a decentralized fashion without full network information. We will take the approaches of evolutionary game theory and reinforcement learning to address these challenges.

∗The authors are with the Department of Information Engineering, the Chinese University of Hong Kong; Email: {cx008,jwhuang}@ie.cuhk.edu.hk
The competitions among secondary users for common channels have often been studied using non-cooperative game theory (e.g., [3]–[7]). Liu and Wu in [4] modeled the interactions among spatially separated secondary users as congestion games with resource reuse. Elias et al. in [5] studied the competitive spectrum access by multiple interference-sensitive secondary users. Nie and Comniciu in [6] designed a self-enforcing distributed spectrum access mechanism based on potential games. Law et al. in [7] studied the price of anarchy of spectrum access game, and showed that users’ selfish choices may significantly degrade system performance. A common assumption of the above results is that each user knows the complete network information. This is, however, often expensive or infeasible to achieve due to significant signaling overhead and the competitors’ unwillingness to share information.

Another body of related work is to design cooperative spectrum access mechanisms using cooperative game theory. Attar, Nakhai and Aghvami in [8] achieved efficient and fair resource allocation between primary and secondary users using Nash bargaining solutions. Saad et al. in [9] designed a coalitional game framework for cooperative spectrum resource sharing.

Another common assumption of all the above work is that secondary users are fully rational and thus often adopt their channel selections based on best responses. To have full rationality, a user needs to have a high computational power to collect and analyze the network information in order to predict other users’ behaviors. This is often not feasible due to the limitations of today’s wireless devices.

When not knowing the channel information, secondary users need to learn the environment and adapt the channel selection decisions accordingly. Authors in [10], [11] used no-regret learning to solve this problem, assuming that the users’ channel selections are common information. The learning converges to a correlated equilibrium [12], wherein the common observed history serves as a signal to coordinate all users’ channel selections. When users’ channel selections are not observable, authors in [13]–[15] designed multi-agent multi-armed bandit learning algorithm to minimize the expected performance loss of distributed spectrum access. Li in [16] applied reinforcement learning to analyze Aloha-type spectrum access. Here we use reinforcement learning to study a different channel access mechanism, which leads to different user utility functions and significant different analysis.

In this paper, we consider a dynamic spectrum access mechanism design with and without complete network information (i.e., channel dynamics and user selections). The common assumption in both scenarios is bounded rationality, where secondary users choose channels only based on better responses instead of best responses. This requires much less computation power than the full rationality case, and
thus matches the reality of wireless communications better.

We first propose an evolutionary game approach for dynamic spectrum access with the complete network information, where each secondary user takes a better response to evolve its spectrum access decision over time. We then propose a distributed reinforcement learning mechanism for dynamic spectrum access with incomplete information, which does not require any prior knowledge of channel statistics or information exchange among users. In this case, each secondary user utilizes local observations including channel sensing and contention results to estimate its expected payoff, and learn to adjust its channel selection strategy adaptively. The main results and contributions of this paper are as follows:

- **Evolutionary spectrum accessing:** With complete network information, we formulate the dynamic spectrum access of multiple secondary users over stochastically heterogeneous primary channels as an evolutionary spectrum access game. The dynamics of the game corresponds to an evolutionary spectrum access mechanism. We characterize the evolutionary equilibrium in closed-form and show that it is an evolutionarily stable strategy (ESS) and is max-min fair.

- **Learning mechanism for dynamic spectrum access:** With incomplete network information, we propose a distributed reinforcement learning mechanism for dynamic spectrum access. By using the theory of stochastic approximation, we show that the learning algorithm globally converges to a $\xi$-approximate Nash equilibrium.

- **Superior performance:** Comparing with a congestion-unaware random access scheme, we show that the proposed evolutionary spectrum access mechanism and distributed reinforcement learning mechanism achieve up to 82% and 70% performance improvement, and are robust to random perturbations of channel selections.

The rest of the paper is organized as follows. We introduce the system model in Section II. After briefly review the evolutionary game theory in Section III, we present the evolutionary spectrum access mechanism with complete information in Section IV. Then we introduce distributed reinforcement learning mechanism with incomplete information in Section V. We illustrate the performance of the proposed mechanisms through numerical results in Section VI and finally conclude in Section VII.

### II. System Model

We consider a cognitive radio network with $M$ independent and stochastically heterogeneous (primary) channels. There are $N$ homogeneous secondary users trying to opportunistically access these channels.
using a slotted transmission structure (see Figure 1). We first assume that there exists a common control channel for exchanging the channel selection information among the secondary users. This common channel will be removed when we discuss the distributed reinforcement learning mechanism in Section V. The system model is described as follows:

- **Channel State**: the channel state for a channel \( m \) during time slot \( t \) is

\[
S_m(t) = \begin{cases} 
0, & \text{if the channel } m \text{ is occupied by primary transmissions,} \\
1, & \text{if the channel } m \text{ is idle.}
\end{cases}
\]

- **Channel State Changing**: for a channel \( m \), we assume that the channel state is an i.i.d. Bernoulli random variable, with an idle probability \( \theta_m \in (0, 1) \) and a busy probability \( 1 - \theta_m \). This model can be a good approximation of reality if the time slots for secondary transmissions are sufficiently long or the primary transmissions are highly bursty [13].

- **Homogeneous Users**: all users achieve the same maximum data rate on the same channel.

- **Heterogeneous Channel Throughput**: channel \( m \) offers a maximum data rate (when only one user transmits) \( B_m \), which is channel specific due to differences in bandwidth and technologies.

- **Time Slot Structure**: each secondary user \( n \) executes the following four stages synchronously during each time slot (see Figure 1):

  - **Channel Sensing**: sense one of the channels, based on the channel selection decision made at the end of last time slot.

  - **Channel Contention**: if the sensed channel is idle, compete for the channels with two steps:

    1) randomly generate a backoff timer value \( \tau_n \) according to a common uniform distribution on \((0, \tau_{max})\).

    2) monitor the channel until the timer expires, and grab the channel if there is no ongoing transmission.

For simplicity, we assume that the timer value is continuous (e.g., [13]), and thus the probability that multiple users generate the same counter value and cause a collision is zero.

- **Data Transmission**: transmit data packets if the user successfully grabs the channel.

- **Channel Selection**: in the complete information case, broadcast the chosen channel ID to other users through the common control channel, and then make the channel selection decision based

---

\[1\] Please refer to [17] for the details on how to set up and maintain a reliable common control channel in cognitive radio networks.
on the evolutionary spectrum access mechanism. In the incomplete information case, update the channel and user estimations based on the current access result, and then make the channel selection decision according to the distributed reinforcement learning algorithm.

We have shown the following result in our prior work:

**Lemma 1** ([19]). If $k_m$ secondary users sense the same idle channel $m$, then the expected throughput of each secondary user is $B_m \theta_m / k_m$.

Lemma 1 illustrates the congestion effect among competing users. Note that the total expected rate of the $k_m(t)$ secondary users is $B_m \theta_m$, which does not depend on the number of users. This means that there is no waste of social welfare due to users’ competition. This assumption helps us to focus the analysis on the users’ choices due to its own potential loss. For detailed discussions of more general channel access mechanisms (under complete information and full rationality), see [7]. Since our analysis is from the secondary users’ perspective, we will use the terms “secondary user” and “user” interchangeably.

### III. Overview of Evolutionary Game Theory

Evolutionary game theory was first used in biology to study the change of animal populations, and then later applied in economics to model human behaviors. It is also a good way to understand how a large number of users converge to Nash equilibria in a dynamic system [20]. A player in an evolutionary game has bounded rationality, which means it has only limited computational capability and knowledge, and tries to improve its decisions as it learns about the environment over time [21]. The replicator dynamics and the evolutionarily stable strategy concept are the keys to study game dynamics and equilibrium solution in evolutionary game theory [20].

#### A. Replicator Dynamics

The replicator dynamics are a set of differential equations that model how population shares corresponding to different strategies evolve over time in an evolutionary game. We consider a game with a large population of players, where each player chooses a strategy $i$ from a finite set of strategies $\mathcal{I} = \{1, \ldots, I\}$. The population state is $\mathbf{x}(t) = (x_i(t), \forall i \in \mathcal{I})$, where $x_i(t)$ denoting the proportion of players adopting strategy $i$ at time $t$. The replicator dynamics are given as

$$\dot{x}_i(t) = \beta x_i(t) \left[ U(i, \mathbf{x}(t)) - \bar{U}(t) \right], \forall i \in \mathcal{I},$$  (1)
where $U(i, x(t))$ is the payoff of the players choosing strategy $i$ in state $x(t)$, $\bar{U}(t)$ is the average payoff of the population, and $\beta > 0$ is the rate of strategy adaptation. Equation (1) shows that strategies with better payoffs than the average will be chosen by more players in the population [20].

B. Evolutionarily Stable Strategy

The evolutionarily stable strategy (ESS) is a key concept to describe the evolutionary equilibrium. An ESS ensures the stability such that the population is robust to perturbations by a small fraction of players.

Formally, suppose that a small share $\epsilon \in (0, 1)$ of players in the population deviate to choose a mutant strategy $j \in I$, all other players stick to the incumbent strategy $i$. Let $U(a, x_i)$ denote the payoff of choosing a strategy $a \in \{i, j\}$ given that all other players choosing the same strategy $i$. Similarly, $U(a, x_{\epsilon j + (1-\epsilon)i})$ denotes the payoff of choosing a strategy $a \in \{i, j\}$ if $\epsilon$ fraction of players playing the mutant strategy $j$ while the others playing strategy $i$.

Definition 1 ([20]). A strategy $i$ is an evolutionarily stable strategy if for every strategy $j \neq i$, there exists an $\epsilon \in (0, 1)$ such that $U(i, x_{\epsilon j + (1-\epsilon)i}) > U(j, x_{\epsilon j + (1-\epsilon)i})$ for any $j \neq i$ and $\epsilon \in (0, \bar{\epsilon})$.

Definition 1 means that the mutant strategy $j$ cannot invade the population when the perturbation is small enough, if the incumbent strategy $i$ is an ESS. The next equivalent definition connects ESS with the Nash equilibrium.

Definition 2 ([20]). A strategy $i$ is an evolutionarily stable strategy if for all $j \neq i$,

\[ U(i, x_i) > U(j, x_i), \]  

or

\[ U(i, x_i) = U(j, x_i) \text{ and } U(i, x_j) > U(j, x_j). \]

From a non-cooperative game theoretic point of view (i.e., regard $x_a$ as the strategy $a$ of the other players), an ESS must be a Nash equilibrium (NE) since $U(i, i) \geq U(j, i)$ based on (2) and (3). We also see that any strict NE (i.e., $U(i, i) > U(j, i), \forall j \neq i$) must be an ESS.

IV. EVOLUTIONARY SPECTRUM ACCESS

We now apply the evolutionary game theory to design an efficient dynamic spectrum access mechanism with complete network information. We use the replicator dynamics to analyze the dynamics of spectrum

\footnote{Following convention in the definition of ESS, we consider a symmetric game where all users adopt the same strategy $i$ at the ESS. The definition can be (and will be) easily extended to the case of asymmetric game, where we can think of the population’s collective behavior as a mixed strategy $i$ at the ESS.}
access and show that the spectrum access equilibrium is an ESS, which guarantees that the spectrum access mechanism is robust to the random perturbations of users’ channel selections.

A. Evolutionary Game Formulation

The evolutionary spectrum access game is formulated as follows:

- Players: the set of users $N = \{1, 2, ..., N\}$.
- Strategies: each user can access any one of the set of channels $\mathcal{M} = \{1, 2, ..., M\}$.
- Population state: the user distribution over $M$ channels at time $t$, $x(t) = (x_m(t), \forall m \in \mathcal{M})$, where $x_m(t)$ is proportion of users selecting channel $m$ at time $t$. We have $\sum_{m \in \mathcal{M}} x_m(t) = 1$ for all $t$.
- Payoff: a user’s expected throughput $U_n(a_n, x(t))$ when choosing channel $a_n \in \mathcal{M}$, given that the population state is $x(t)$. From Lemma 1 we have $U_n(a_n, x(t)) = \frac{B_n \theta_{a_n}}{N x_{a_n}(t)}$.

For the ease of analysis, we assume that the total number of users $N$ is large enough such that $x_m(t)$ are continuous variables for all $m$ and $t$. In Section VI we will show that our algorithm works equally well when $N$ is small. Such assumption is not required for the learning mechanism in Section V.

B. Evolutionary Spectrum Access Mechanism and Replicator Dynamics

Based on the evolutionary game formulation above, we propose an evolutionary game spectrum access mechanism in Table I. Parameter $\alpha$ in the mechanism represents the strategy adaptation rate. The key idea is that users who have lower payoffs than the targeted average population payoff $\sum_{i=1}^M B_i \theta_i/N$ will adjust their channel selections. We show that the channel selection dynamics in the mechanism can be described with the replicator dynamics in (4).

**Theorem 1.** For the evolutionary spectrum access mechanism in Table II the replicator dynamics are

$$\dot{x}_m(t) = \beta \left( \frac{B_m \theta_m}{N} - \frac{\sum_{i=1}^M B_i \theta_i}{N} x_m(t) \right), \forall m \in \mathcal{M}, (4)$$

where the derivative is with respect to time and the rate of strategy adaptation $\beta = \frac{\alpha N}{\sum_{i=1}^M B_i \theta_i}$.

Theorem 1 is proved in Appendix A. Note that for a channel $m$ with $x_m(t) > 0$, the replicator dynamics in (4) can be written into the standard form from in (1) as

$$\dot{x}_m(t) = \beta x_m(t) \left( \frac{B_m \theta_m}{N x_m(t)} - \frac{\sum_{i=1}^M B_i \theta_i}{N} \right).$$
The replicator dynamics in (4) imply that the channels that offer better payoffs than the average payoff will be chosen by more users in the next time slot. Note that the mapping between the replicator dynamics and the proposed algorithm is not standard, and is the result of a careful algorithm design.

C. Evolutionary Equilibrium

We next investigate the equilibrium of the evolutionary spectrum access mechanism. At the equilibrium, \( \dot{x}_m(t) = 0 \) as no user can gain a better payoff by changing its channel selection at the equilibrium.

**Theorem 2.** The replicator dynamics in (4) has the evolutionary equilibrium \( x^* = (x^*_m, \forall m \in M) \), where

\[
x^*_m = \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i}, \forall m \in M.
\]

(5)

Next we characterize two properties of the evolutionary equilibrium.

1) **Evolutionarily Stable Strategy:** In general, the equilibrium of the replicator dynamics may not be an ESS [20]. For our model, we can prove the following.

**Theorem 3.** For the evolutionary spectrum access mechanism, the evolutionary equilibrium \( x^* \) is an ESS.

Actually we can prove a stronger result than Theorem 3. Typically, an ESS is only locally asymptotically stable (i.e., stable within a limited region around the ESS) [20]. For our case, we can show that the evolutionary equilibrium \( x^* \) is globally asymptotically stable (i.e., stable in the entire feasible region of a population state \( x \), \( \{x = (x_m, m \in M) \mid \sum_{m=1}^{M} x_m = 1 \text{ and } x_m \geq 0, \forall m \in M \} \)).

**Theorem 4.** For the evolutionary spectrum access mechanism, the evolutionary equilibrium \( x^* \) is globally asymptotically stable.

The proofs of both Theorems 3 and 4 are given in the Appendix. Since the ESS is globally asymptotically stable, the evolutionary spectrum access mechanism is robust to any degree of (not necessarily small) random perturbations of channel selections.

2) **Max-Min Fairness:** An important objective for the dynamic spectrum access is to ensure fair spectrum opportunities among the secondary users. We investigate the fairness of the evolutionary equilibrium \( x^* \) by introducing the following max-min optimization problem:

\[
\max_{x=(x_m, \forall m \in M)} \min_{n \in \mathcal{N}} \{U_n(a_n, x)\}
\]

subject to \( \sum_{m=1}^{M} x_m = 1 \text{ and } x_m > 0, \forall m \in M \).

(6)
The objective in (6) is to find a population state $x$ that maximizes the minimum expected payoff among $N$ users. The constraint is to ensure that the population state $x$ is feasible and the spectrum opportunities are fully utilized. By solving the problem in (6) (for details, see Appendix D), we have

**Theorem 5.** The evolutionary equilibrium $x^*$ is the unique global optimal solution of problem (6).

Theorem 5 shows that the evolutionary spectrum access mechanism achieves the max-min fair spectrum opportunity sharing among the secondary users.

**V. DISTRIBUTED REINFORCEMENT LEARNING FOR DYNAMIC SPECTRUM ACCESS**

For the evolutionary spectrum access mechanism in Section IV, we assume that each user has prior knowledge of spectrum dynamics (i.e., channel idle probability $\theta_m$ for each $m$) and can observe other users’ channel selections by explicit information exchange. Acquiring such complete information requires infrastructure (e.g., base station or coordinator) for spectrum dynamics measurement and a common control channel for information exchange. The signaling overhead and energy consumption can be quite significant and even infeasible in some network scenarios. To address this issue, we propose a distributed reinforcement learning mechanism for dynamic spectrum access without the knowledge of the spectrum dynamics. Users also do not need to exchange information and can make channel selections based their local observations only. We show that the learning mechanism converges to a $\xi$-approximate Nash equilibrium.

**A. Expected Payoff Estimation**

To achieve accurate estimation based on local observation, a user needs to gather a large number of observation samples. This motivates us to divide the learning time into a sequence of decision periods indexed by $T(=1,2,\ldots)$, where each decision period consists of $L$ time slots. During a single decision period, a user accesses the same channel in all $L$ time slots. Thus the total number of users accessing each channel does not change within a decision period, which allows users to better learn the environment.

According to Lemma 1, a user’s expected payoff during decision period $T$ depends on the number of users $k_m(T)$ choosing the same channel during that decision period, the channel idle probability $\theta_m$, and the data rate $B_m$. The data rate $B_m$ can be obtained when a user first successfully transmits on the

Since the users are homogeneous, the population state uniquely determines the system performance and the identities of the users are not important.
channel \( m \). The remaining issue is how to get accurate estimations of \( k_m(T) \) and \( \theta_m \). We will apply the maximum likelihood estimation (MLE) \cite{22} due to its efficiency and ease of implementation.

1) Maximum Likelihood Estimation of \( k_m(T) \): At the beginning of each time slot \( l(=1, \ldots, L) \) of a decision period \( T \), a user \( n \) will sense the same channel \( m \). If the channel is idle, the user will compete to grab the channel according to the timer mechanism in Section \[III\]. At the end of each time slot \( l \), a user \( n \) observes \( S^n_m(T, l) \) and \( I^n_m(T, l) \), which denote the state of the channel \( m \) (i.e., whether occupied by the primary traffic) and whether the user has successfully grabs the channel \( m \), i.e.,

\[
I^n_m(T, l) = \begin{cases} 
1, & \text{if user } n \text{ successfully grabs channel } m, \\
0, & \text{otherwise.}
\end{cases}
\]

At the end of each decision period \( T \), each user \( n \) can collect a set of local observations \( \Omega_n(T) = \{S^n_m(T, l), I^n_m(T, l)\}_{l=1}^L \). Note that if \( S^n_m(T, l) = 0 \) (i.e., the channel is occupied by the primary traffic), we set \( I^n_m(T, l) \) to be NULL, which contain no valid information.

When the channel \( m \) is idle (i.e., no primary traffic), \( k_m(T) \) users compete for the channel according to timer mechanism in Section \[III\]. If a user \( n \) is among the \( k_m(T) \) users, it captures the channel \( m \) with probability

\[
Pr\{I^n_m(T, l) = 1|k_m(T)\} = \int_0^{\tau_{\max}} \frac{1}{\tau_{\max}} \left(1 - \frac{\tau_n}{\tau_{\max}}\right)^{k_m(T)-1} d\tau_n = \frac{1}{k_m(T)}.
\]

Since there are a total of \( \sum_{l=1}^L S^n_m(T, l) \) rounds of channel contentions in the period \( T \) and each round is independent, the total number of successful channel captures \( \sum_{l=1}^L I^n_m(T, l) \) follows the binomial distribution. A user \( n \) can compute the likelihood of \( k_m(T) \), i.e., the probability of the realized observations \( \Omega_n(T) \) given the parameter \( k_m(T) \) as

\[
L[\Omega_n(T)|k_m(T)] = \left( \frac{\sum_{l=1}^L S^n_m(T, l)}{\sum_{l=1}^L I^n_m(T, l)} \right) \left( \frac{1}{k_m(T)} \right)^{\sum_{l=1}^L I^n_m(T, l)} \left( 1 - \frac{1}{k_m(T)} \right)^{\sum_{l=1}^L S^n_m(T, l) - \sum_{l=1}^L I^n_m(T, l)}.
\]

The MLE of \( k_m(T) \) can be computed by maximizing the log-likelihood function \( \ln L[\Omega_n(T)|k_m(T)] \), i.e., \( \max_{k_m(T)} \ln L[\Omega_n(T)|k_m(T)] \). By solving the first order condition, we obtain

\[
\frac{1}{k_m(T)} = \frac{\sum_{l=1}^L I^n_m(T, l)}{\sum_{l=1}^L S^n_m(T, l)},
\]

which is the sample averaging estimation. By the law of large numbers, we know that the estimated value
\[
\frac{1}{k_m(T)} \text{ converges to the true value } \frac{1}{k_m(T)} \text{ as the length of decision period } L \text{ goes to infinity. When the length of decision period } L \text{ is large, according to the central limit theorem, we have }
\]
\[
\frac{1}{k_m(T)} - \frac{1}{k_m(T)} \sim \mathcal{N} \left( 0, \frac{1}{k_m(T)} \frac{1 - \frac{1}{k_m(T)}}{\sum_{l=1}^{L} S_{m,t}(T)} \right),
\]
(8)
where \( \mathcal{N} \left( 0, \frac{1}{k_m(T)} \frac{1 - \frac{1}{k_m(T)}}{\sum_{l=1}^{L} S_{m,t}(T)} \right) \) denotes the normal distribution with mean 0 and variance \( \frac{1}{k_m(T)} \frac{1 - \frac{1}{k_m(T)}}{\sum_{l=1}^{L} S_{m,t}(T)} \).

Note that the estimation of \( k_m(T) \) only replies on the local observations during decision period \( T \), as the number of users accessing channel \( m \) might change over different decision periods.

2) Maximum Likelihood Estimation of \( \theta_m \): We next apply the MLE to estimate the channel idle probability \( \theta_m \). Since the channel state \( S_{m,n}(T,l) \) is i.i.d. over different time slots and different decision periods, we can improve the estimation by averaging over multiple periods.

We first compute one-period MLE of \( \theta_m \). Given the observation set \( \Omega_{n}(T) \) of a decision period \( T \), a user \( n \) can compute the likelihood of \( \theta_m \) as
\[
\mathcal{L}[\Omega_{n}(T)|\theta_m] = \exp \left( \sum_{l=1}^{L} S_{m,t}(T,l) \theta_m (1 - \theta_m) \right).
\]
Similarly, by maximizing the log-likelihood function, we obtain \( \hat{\theta}_m = \frac{\sum_{l=1}^{L} S_{m,t}(T,l)}{L} \). The estimated value \( \hat{\theta}_m \) converges to the true value \( \theta_m \) as the length of decision period \( L \) goes to infinity. When the length of decision period \( L \) is large, we have
\[
\hat{\theta}_m - \theta_m \sim \mathcal{N} \left( 0, \frac{\theta_m (1 - \theta_m)}{L} \right).
\]
(9)

We then average the estimation over multiple decision periods. When a user \( n \) finishes accessing a channel \( m \) for a total of \( C \) periods, it updates the estimation of the channel idle probability \( \theta_m \) as
\[
\tilde{\theta}_m(C) = \frac{1}{C} \sum_{i=1}^{C} \hat{\theta}_m(i),
\]
(10)
where \( \tilde{\theta}_m(C) \) is the estimation of \( \theta_m \) based on the information of all \( C \) decision periods, and \( \hat{\theta}_m(i) \) is the one-period estimation. By doing so, we have
\[
\tilde{\theta}_m(C) - \theta_m \sim \mathcal{N} \left( 0, \frac{\theta_m (1 - \theta_m)}{CL} \right),
\]
(11)
\footnote{The total periods from the first visit to the \( C \)-th visit can be larger than \( C \), as the user may choose to access other channels during some decision periods in the middle.}
which reduces the variance of estimation in (9) by a factor of $C$.

By the MLE, we can obtain the estimation of $k_m(T)$ and $\theta_m$ as $\hat{k}_m(T)$ and $\hat{\theta}_m$, respectively, and then estimate the true expected payoff $U_n(T) = \frac{B_m \theta_m}{k_m(T)}$ as $\tilde{U}_n(T) = \frac{B_m \hat{\theta}_m}{\hat{k}_m(T)}$.

### B. Distributed Reinforcement Learning Mechanism

We now propose a distributed reinforcement learning mechanism as in Table II, which determines the channel selections of all users at the beginning of each decision period $T$. The idea is to extend the principle of single-agent reinforcement learning [23] to a multi-agent setting.

At the beginning of each period $T$, a user $n \in N$ chooses a channel $a_n(T) \in M$ to access according to its mixed strategy $\sigma_n(T) = (\sigma_m^n(T), \forall m \in M)$, where $\sigma_m^n(T)$ is the probability of choosing channel $m$. The mixed strategy is generated according to $P_n(T) = (P_m^n(T), \forall m \in M)$, which represents its perceptions of the payoff performance of choosing different channels based on local estimations. Perceptions are based on local observations in the past and may not accurately reflect the expected payoff. For example, if a user $n$ has not accessed a channel $m$ for many decision intervals, then perception $P_m^n(T)$ can be out of date. The key challenge for the learning algorithm is to update the perceptions with proper parameters such that perceptions equal to expected payoffs at the equilibrium.

Similarly with the single-agent reinforcement learning, we choose the Boltzmann distribution as the mapping from perceptions to mixed strategies, i.e.,

$$\sigma_m^n(T) = \frac{e^{\gamma P_m^n(T)}}{\sum_{i=1}^M e^{\gamma P_i^n(T)}}, \forall m \in M,$$

where $\gamma$ is the temperature that controls the randomness of channel selections. When $\gamma \to 0$, each user will choose to access channels uniformly at random. When $\gamma \to \infty$, user $n$ always chooses the channel with the largest perception value $P_m^n(T)$ among all channel $m \in M$. We will show later on that the choice of $\gamma$ trades off convergence and performance of the learning algorithm.

At the end of a decision period $T$, a user $n$ computes its estimated expected payoff $\tilde{U}_n(T)$ as in Section [V-A] (i.e., by using the MLE method based on the set of local observations $\Omega_n(T)$ during the period), and adjusts its perceptions as

$$P_m^n(T + 1) = \begin{cases} (1 - \mu_T)P_m^n(T) + \mu_T \tilde{U}_n(T), & \text{if } a_n(T) = m, \\ P_m^n(T), & \text{otherwise,} \end{cases}$$

(13)
where \((\mu_T \in (0, 1), \forall T)\) are the smoothing factors. A user only changes the perception of the channel just accessed in the current decision period, and keeps the perceptions of other channels unchanged.

Table II summarizes the distributed reinforcement learning mechanism. Next we study the dynamics and convergence of the learning mechanism based on the theory of stochastic approximation.

C. Mean Dynamics of Distributed Reinforcement Learning

We first study the dynamics of distributed reinforcement learning mechanism. Such study will provide useful insights for the limiting behavior of the learning mechanism.

First, the perception value update in (13) can be written in the following equivalent form,

\[
P_m^n(T + 1) - P_m^n(T) = \mu_T [Z_m^n(T) - P_m^n(T)], \forall n \in \mathcal{N}, m \in \mathcal{M},
\]

where \(Z_m^n(T)\) is the update value defined as

\[
Z_m^n(T) = \begin{cases} 
\bar{U}_n(T), & \text{if } a_n(T) = m, \\
P_m^n(T), & \text{otherwise}.
\end{cases}
\]

For the sake of brevity, we denote the perception values, update values, and mixed strategies of all the users as \(P(T) \triangleq (P_m^n(T), \forall m \in \mathcal{M}, n \in \mathcal{N}), Z(T) \triangleq (Z_m^n(T), \forall m \in \mathcal{M}, n \in \mathcal{N}), \) and \(\sigma(T) \triangleq (\sigma_m^n(T), \forall m \in \mathcal{M}, n \in \mathcal{N}),\) respectively.

Let \(Pr\{k_m(T) = i|\sigma(T)\}\) denote the probability of \(k_m(T)\) users accessing channel \(m\) given that the mixed strategies of all users are \(\sigma(T)\). Since each user independently chooses a channel according to its mixed strategy \(\sigma_n(T)\), thus the number of \(k_m(T)\) users accessing a channel \(m\) is a random variable, which follows the Binomial distribution of \(N\) independent non-homogeneous Bernoulli tries with the probability mass function as

\[
Pr\{k_m(T) = i|\sigma(T)\} = \sum_{\sum_{n=1}^{N} I(a_n(T) = m) = i} (\sigma_m^n(T))^{I(a_n(T) = m)} (1 - \sigma_m^n(T))^{1-I(a_n(T) = m)},
\]

where \(I(a_n(T) = m) \in \{0, 1\}\) is an indicator of whether user \(n\) chooses channel \(m\) at period \(T\).

Since update value \(Z_m^n(T)\) depends on user \(n\)'s estimated payoff \(\bar{U}_n(T)\) (which in turn depends on \(k_m(T)\)), thus \(Z_m^n(T)\) is also a random variable. The equations in (14) are hence stochastic difference equations, which are difficult to analyze directly. We thus focus on the analysis of its mean dynamics \([24]\). To proceed, we define the mapping from the perceptions \(P(T)\) to the expected payoff of user \(n\)
choosing channel \( m \) as \( G^n_m(P(T)) \triangleq E[U_n(T) | P(T), a_n(T) = m] \). Here the expectation \( E[\cdot] \) is taken with respective to the mixed strategies \( \sigma(T) \) of all users (i.e., the perceptions \( P(T) \) of all users due to (12)). We show that

**Lemma 2.** When the length of a decision period \( L \) is large enough, the mean dynamics of the distributed reinforcement learning mechanism are given as

\[
\dot{P}^n_m(T) = \sigma^n_m(T)[G^n_m(P(T)) - P^n_m(T)], \forall n \in \mathcal{N}, m \in \mathcal{M}.
\]

The proof is given in Appendix E. From (14) and Lemma 2 we see that when the gap between the expected payoff \( G^n_m(P(T)) \) and the perception value \( P^n_m(T) \) is large, the perception value will updated (i.e., added) with a larger expected value \( Z^n_m(T) - P^n_m(T) \). Thus, the gap between the expected payoff \( G^n_m(P(T)) \) and the perception value \( P^n_m(T) \) will be reduced significantly. This implies that the learning mechanism converges to an equilibrium where the perception value \( P^n_m(T) \) equals to the expected payoff \( G^n_m(P(T)) \).

**D. Global Convergence and Stability of Distributed Reinforcement Learning**

We now study the convergence of distributed reinforcement learning mechanism. Recall that users’ mixed strategies are mapped from the perceptions based on the Boltzmann distribution as in (12). This mapping enables the learning to admit a contraction structure with a proper choice of temperature \( \gamma \).

**Lemma 3.** For the distributed reinforcement learning mechanism, if the temperature satisfies

\[
\gamma < \frac{1}{\max_{m \in \mathcal{M}} \{B_m \theta_m\} (N-1)^2 (\ln N+1)^2},
\]

then the mapping from the perceptions to the expected payoffs \( G(P(T)) = [G^n_m(P(T)), \forall m \in \mathcal{M}, n \in \mathcal{N}] \) is a maximum-norm contraction.

The proof is given in Appendix F. Note that \( \gamma < \frac{1}{\max_{m \in \mathcal{M}} \{B_m \theta_m\} (N-1)^2 (\ln N+1)^2} \) is a sufficient condition to form a contraction mapping, which is in turn is a sufficient condition for convergence. Simulation results show that a slightly larger \( \gamma \) may also lead to the convergence of the mapping. Based on the property of contraction mapping, there exists a fixed point \( P^* \) such that \( G(P^*) = P^* \). By the theory of stochastic approximations [24], we show that the distributed reinforcement learning mechanism also converges to the same limit point \( P^* \). The proof is given in Appendix G.

**Theorem 6.** For the distributed reinforcement learning mechanism with the mean dynamics in (17), if the length of a decision period \( L \) is large enough, the temperature \( \gamma < \frac{1}{\max_{m \in \mathcal{M}} \{B_m \theta_m\} (N-1)^2 (\ln N+1)^2} \),
\[ \sum_T \mu_T = \infty, \text{ and } \sum_T \mu_T^2 < \infty, \text{ then the sequence } \{P(T), \forall T \geq 0\} \text{ converges to the unique limit point } P^* \triangleq (P^*_{m}, \forall m \in \mathcal{M}, n \in \mathcal{N}) \text{ of the differential equations} \]
\[
\dot{P}_m^n(T) = \sigma_m^n(T)[G_m^n(P(T)) - P_m^n(T)], \forall n \in \mathcal{N}, m \in \mathcal{M} \tag{18}
\]

with probability one. Further, the limit point \( P^* \) satisfies
\[
G_m^n(P^*) = P_m^{n*}, \forall n \in \mathcal{N}, m \in \mathcal{M}. \tag{19}
\]

We next study the stability of distributed reinforcement learning mechanism. We show that the distributed reinforcement learning mechanism is globally asymptotically stable by proving that \( V(P(T)) = \max_{n,m} \{|P_m^n(T) - P_m^{n*}|\} \) is a strict Lyapunov function (please refer to Appendix [H] for the proof).

**Theorem 7.** For the distributed reinforcement learning mechanism, if the length of a decision period \( L \) is large enough, the temperature \( \gamma < \frac{1}{\max_{m \in \mathcal{M}} \{B_m b_m\} (N-1)^2 \ln(N+1)} \), \( \sum_T \mu_T = \infty, \text{ and } \sum_T \mu_T^2 < \infty, \) the limit point \( P^* \) is globally asymptotically stable.

Since the limit point \( P^* \) is globally asymptotically stable, the learning mechanism for dynamic spectrum access is robust to random perturbations of users’ channel selections.

**E. Approximate Nash Equilibrium**

We now explore the property of the equilibrium \( P^* \) of the distributed reinforcement learning mechanism. From Theorem 6 we see that
\[
G_m^n(P^*) = E[U_n(T)|P^*, a_n(T) = m] = P_m^{n*}. \tag{20}
\]

It means that the perception value \( P_m^{n*} \) is an accurate estimation of the expected payoff in the equilibrium. Moreover, we show that the mixed strategy \( \sigma^* = (\sigma_m^{n*} = \frac{e^{x^n P_m^{n*}}}{\sum_{i=1}^{M} e^{x^n P_m^{i*}}}, \forall n \in \mathcal{N}, m \in \mathcal{M}) \) is an approximate Nash equilibrium.

**Definition 3 ([26]).** A strategy \( A^* = (a_1^*, ..., a_N^*) \) is a \( \xi \)-approximate Nash equilibrium if
\[
U_n(a_n^*, a_{-n}^*) \geq \max_{a_n} U_n(a_n, a_{-n}^*) - \xi, \forall n \in \mathcal{N},
\]

where \( a_{-n}^* \) denotes the strategies of other players except player \( n \).
Here $\xi$ is the gap from a (precise) Nash equilibrium. For the distributed reinforcement learning, we show that

**Theorem 8.** For the distributed reinforcement learning mechanism, the mixed strategy $\sigma^*$ in the equilibrium $P^*$ is a $\xi$-approximate Nash equilibrium, with $\xi = \max_{n \in N} \{-\frac{1}{\gamma} \sum_{m=1}^{M} \sigma^*_m \ln \sigma^*_m\}$.

The proof is given in Appendix I. The gap $\xi$ can be interpreted as the weighted entropy, which describes the randomness of the learning exploration. A larger $\xi$ means worse learning performance. When each user adopts the uniform random access, the gap $\xi$ reaches the maximum value and results in the worst learning performance. Theorems 6 and 8 together illustrate the trade-off between the convergence and performance through the choice of $\gamma$. A small enough $\gamma$ is required to explore the environment (so that users are not getting stuck in channels with the current best payoffs) and guarantee the convergence of distributed reinforcement learning. If $\gamma$ is too small, however, then the performance gap $\xi$ is large due to over-exploration.

VI. Simulation Results

In this section, we evaluate the proposed evolutionary spectrum access and distributed reinforcement learning mechanisms by extensive simulations. Numerical results show that the proposed evolutionary spectrum access (with complete network information) and distributed reinforcement learning mechanisms (without network information) achieve superior performance over random access mechanism, and are robust to random perturbations of channel selections.

A. Evolutionary Spectrum Access

We first study the proposed evolutionary spectrum access mechanism (in Table I) with complete information. We consider a cognitive radio network with $M = 5$ primary channels, with the data rates $\{B_m\}_{m=1}^{M} = \{15, 70, 90, 40, 100\}$ Mbps, and the channel idle probabilities $\{\theta_m\}_{m=1}^{M} = \{\frac{2}{3}, \frac{4}{7}, \frac{5}{9}, \frac{1}{2}, \frac{4}{5}\}$. Figure 3 shows that the convergence speed of the evolutionary spectrum access mechanism decreases in the strategy adaptation factor $\alpha$. In the rest of the simulations, we choose a moderate strategy adaptation factor $\alpha = 0.5$ so that readers can clearly observe the dynamics in the figures.

1) Large population simulation: We first consider the case that the number of users is large, e.g., $N = 20$ and 40 respectively. The results are shown in Figures 6 to 9. From these figures, we see that

• **Fast convergence:** the mechanism takes less than 20 iterations to converge in all figures.
• **Convergence to ESS**: for both $N = 20$ and $40$, the mechanism converges to the same ESS $x^* = \left( \frac{B_1 \theta_1}{\sum_{i=1}^M B_i \theta_i}, \ldots, \frac{B_M \theta_M}{\sum_{i=1}^M B_i \theta_i} \right) = (0.05, 0.2, 0.25, 0.1, 0.4)$ (Figures 4 and 5), since the equilibrium characterize the population fractions and does not depend on the number of users (see Theorem 2). At the ESS $x^*$, each user achieves the same expected payoff $U_n(a_n^*, x^*) = \frac{\sum_{i=1}^M B_i \theta_i}{N} = 10$ (for $N = 20$) and 5 (for $N = 40$) (Figures 6 and 7). This illustrates the max-min property of the equilibrium.

• **Robustness to random perturbations**: we let a fraction of users randomly play mutation strategies when the system is at the ESS $x^*$. At the time slot $t = 30$, 50% and 90% of users will randomly choose a new channel in Figures 8 and 9, respectively. We see that the mechanism is capable to recover the ESS $x^*$ quickly after the mutation occurs. This demonstrates that the evolutionary spectrum access mechanism is robust to the perturbations in the network.

2) **Small population simulation**: We next consider the case that the number of users is small. In this case, the fraction of users on a channel $x_m(t)$ can on longer be treated as a continuous variable and hence the ESS state $x_m^* = \frac{B_m \theta_m}{\sum_{i=1}^M B_i \theta_i}$ may not be achievable (actually there may not exist an ESS). From an algorithmic point of view, each user $n$ still chooses a channel $a_n(t)$ follows the algorithm in Table I, i.e., choosing a new candidate channel $m$ with probability $p_m$ and switches to the new channel if it is a better response (i.e., $U_n(m, x(t)) > U_n(a_n(t), x(t))$). The results of the evolutionary spectrum access with the number of users $N = 3$ and $N = 5$ are shown in Figures ?? and ???, respectively. We see that the mechanism converges to a Nash equilibrium in both cases. The key reason is that the spectrum access game is a potential game (see Appendix [I] for details), which has the finite improvement property and iterative better responses lead to a Nash equilibrium [4].

**B. Distributed Reinforcement Learning**

We evaluate the distributed reinforcement learning mechanism with local observations in this part.

1) **Maximum Likelihood Estimation**: We first study the performance of the maximum likelihood estimation (MLE) of the number of secondary users simultaneously accessing on a channel during a decision period ($k_m(T)$) and the channel idle probability ($\theta_m$). We implement the MLE with different period length $L$ (and thus different number of observations). For each fixed $L$, we repeat the MLE for 20 times and calculate the mean and variance of the estimated values. The results are shown in Figures 10 and 11. As the number of observations increases, the estimated mean converges to the true value and the variance of the estimated value decreases.
2) Distributed Reinforcement Learning: We set the length of a decision period $L = 200$ time slots, which provides a good estimation. We further set the smoothing factor $\mu_T = \frac{1}{100+T}$ for all $T$, which satisfies that the condition $\sum_T \mu_T = \infty \text{ and } \sum_T \mu_T^2 < \infty$.

We first evaluate the distributed reinforcement learning mechanism with different choices of temperature $\gamma$. We run the learning mechanism sufficiently long until the time average system payoff does not change. Figure 12 verifies the trade-off between the convergence and performance, and demonstrates that a proper temperature $\gamma$ can offer the best performance. When $\gamma$ is small, the gap $\xi$ in Theorem 8 can be large. When $\gamma$ is very large, the algorithm may not converge and the performance is again negatively affected.

We next implement the distributed reinforcement learning mechanism with the number of users $N = 3$ and 20, respectively. We set $\gamma = 8.0$ since it achieves good system performance under a wide range choice of $N$ as in Figure 12. The results are shown in Figures 13 to 17. From these figures, we see that

- **Convergence to a Nash equilibrium on time average when the number of users is small (Figure 13):** when $N = 3$, a user’s time average payoff equals to the expected payoff at a pure strategy Nash equilibrium, wherein Users 1, 2, and 3 choose Channel 5, 5, and 3, respectively.

- **Convergence to a $\xi$-approximate ESS on time average when the number of users is large (Figure 14):** when $N = 20$, users achieve the same time average payoff as the expected payoff $U_n(a_n^*, x^*) = \frac{\sum_{i=1}^{N} B_i \theta_i}{N} = 10$ at ESS $x^*$ approximately. Note that an ESS is also a Nash equilibrium.

- **Robustness to random perturbations (Figures 15 and 16):** We let 2 (10, respectively) users reset their perception values and learning history, and randomly select channels at $T = 1000$ when the number of users $N = 3$ (20, respectively). The learning mechanism recover the equilibrium in both cases.

- **Superior performance over random access:** as a benchmark, we also simulate the congestion-unaware random access mechanism, wherein each user uniformly randomly chooses a channel to access at each period. Figure 17 shows that the maximum performance gain of the distributed learning algorithm over the random access algorithm can be as high as 70% when the number of users is small. This is because that the random access mechanism is static, while the learning mechanism adaptively adjusts user’ channel selection decision by responding to other users’s selections. When the number of users is much greater than the number of channels (e.g., $N \geq 40$), both mechanisms achieve the maximum system payoff as the network is saturated and all channel opportunities are fully utilized. In the same figure, we observe that the evolutionary spectrum access mechanism achieves a performance gain up

\[ \text{If we need a smaller convergence time with a less precision, we can choose } L = 100. \]
to 82% over the random access scheme.

Figure [17] also shows that the evolutionary spectrum access mechanism (with complete network information) always achieves a better total system payoff than the distributed learning mechanism (without network information). The performance gap characterizes the value of network information, and implies that we should only use the learning algorithm when the network information is not available.

VII. CONCLUSION

In cognitive radio networks, multiple secondary users will compete for limited and time-varying channel resources in a non-cooperative manner. In this paper, we consider a practical scenario where users have only limited computational power and thus bounded rationality. Evolutionary game theory provides an excellent theoretical framework for modeling and algorithm design for this scenario.

With complete network information, we propose an evolutionary channel access scheme, which converges to a evolutionarily stable strategy and achieves max-min fairness. With incomplete network information, we propose a distributed reinforcement learning mechanism that converges to an approximate Nash equilibrium based on only local observations. Numerical results demonstrate that the proposed mechanisms achieve superior performance over the random access mechanism, and are robust to random perturbations of channel selections.

In the future, we plan to extend the results to a non-i.i.d. primary channel environment. Another direction is to consider the co-existence of heterogeneous secondary users, where users achieve different transmission rates on the same channels.

APPENDIX

A. Proof of Theorem [7]

Given a population state $\mathbf{x}(t) = (x_1(t), ..., x_M(t))$, we can divide the set of channels $\mathcal{M}$ into the following three complete and mutually exclusive subsets: $\mathcal{M}_1 = \{m \in \mathcal{M} | \frac{B_m}{\sum_{i=1}^M B_i} \theta_m < x_m(t)\}$, $\mathcal{M}_2 = \{m \in \mathcal{M} | \frac{B_m}{\sum_{i=1}^M B_i} \theta_m = x_m(t)\}$, and $\mathcal{M}_3 = \{m \in \mathcal{M} | \frac{B_m}{\sum_{i=1}^M B_i} \theta_m > x_m(t)\}$. Then we show the dynamics of the algorithm can be described by equations (4).

For a channel $m \in \mathcal{M}_1$, there must exist at least one user on the channel since $x_m(t) > 0$. However, a user $n$ on this channel achieves an expected payoff less than the system average, i.e., $U_n(m, \mathbf{x}(t)) = \ldots$
\[ \frac{B_m \theta_m}{N x_m(t)} < \frac{\sum_{i=1}^{M} B_i \theta_i}{N} . \] According to the evolutionary spectrum access mechanism, each user has the probability of \( \alpha \left( 1 - \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} \right) \) to move out of the channel \( m \). Since \( \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} < x_m(t) \), it follows that \( p_m = 0 \). That is, no other users will move into this channel. Thus, the replicator dynamics is given as

\[ \dot{x}_m(t) = \alpha \left( 1 - \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} \right) x_m(t) = \frac{\alpha N}{\sum_{i=1}^{M} B_i \theta_i} \left( \frac{B_m \theta_m}{N} - \frac{\sum_{i=1}^{M} B_i \theta_i}{N} x_m(t) \right), \forall m \in \mathcal{M}_1. \]

For a channel \( m \in \mathcal{M}_2 \), there must also exist at least one user on the channel since \( x_m(t) > 0 \). For a user \( n \) on this channel, its expected payoff \( U_n(m, x(t)) = \frac{\sum_{i=1}^{M} B_i \theta_i}{N} \). According to the evolutionary spectrum access mechanism, no user will move out of the channel. Since \( p_m = 0 \), no other users will move into this channel. Thus, the replicator dynamics is given as

\[ \dot{x}_m(t) = \frac{\alpha N}{\sum_{i=1}^{M} B_i \theta_i} \left( \frac{B_m \theta_m}{N} - \frac{\sum_{i=1}^{M} B_i \theta_i}{N} x_m(t) \right) = 0, \forall m \in \mathcal{M}_2. \]

For a channel \( m \in \mathcal{M}_3 \), we first show that no users will move out of the channel. When \( x_m(t) = 0 \), this is trivially true. When \( x_m(t) > 0 \), a user \( n \) on the channel \( m \) has the expected payoff \( U_n(m, x(t)) > \frac{\sum_{i=1}^{M} B_i \theta_i}{N} \). According to the evolutionary spectrum access mechanism, no users will move out of this channel. Since \( p_m > 0 \), there will be some other users from the channel \( m' \in \mathcal{M}_1 \) moving into this channel. Let \( \Delta x(t) \) denote the fraction of the population that want to carry out the movement. We have

\[ \Delta x(t) = \sum_{m' \in \mathcal{M}_1} \frac{\alpha N}{\sum_{i=1}^{M} B_i \theta_i} \left( \frac{\sum_{i=1}^{M} B_i \theta_i}{N} x_m(t) - \frac{B_m \theta_m}{N} \right) = \alpha \sum_{m' \in \mathcal{M}_1} \left( x_m(t) - \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} \right). \]

Since

\[ \sum_{m' \in \mathcal{M}} \left( x_m(t) - \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} \right) = \sum_{m' \in \mathcal{M}} x_m(t) - \sum_{m' \in \mathcal{M}} \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} = 1 - 1 = 0, \]

and \( x_m(t) = \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} \), \forall m' \in \mathcal{M}_2 \), we then obtain

\[ \sum_{m' \in \mathcal{M}_1} \left( x_{m'}(t) - \frac{B_{m'} \theta_{m'}}{\sum_{i=1}^{M} B_i \theta_i} \right) = \sum_{m' \in \mathcal{M}_3} \left( \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} - x_m(t) \right) = \sum_{m'=1}^{M} \left[ \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} - x_m(t) \right]^+. \]
Then, the fraction of the population moving into a channel \( m \in \mathcal{M}_3 \) (replicator dynamics) is

\[
\dot{x}_m(t) = p_m \Delta x(t) = \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} - x_m(t) = \frac{\alpha}{\sum_{i=1}^{M} B_i \theta_i} \left( \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} - x_m(t) \right) + \alpha \sum_{m'=1}^{M} \left[ \frac{B_{m'} \theta_{m'}}{\sum_{i=1}^{M} B_i \theta_i} - x_{m'}(t) \right].
\]

Substituting (4) into (21), we have

\[
\dot{V}(x(t)) = \sum_{m=1}^{M} \frac{\partial V(x(t))}{\partial x_m(t)} \dot{x}_m(t) = 2 \sum_{m=1}^{M} (x_m(t) - x^*) \dot{x}_m(t).
\]

Substituting (4) into (21), we have

\[
\frac{dV(x(t))}{dt} = 2 \sum_{m=1}^{M} (x_m(t) - x^*) \beta \left( \frac{B_m \theta_m}{N} - \sum_{i=1}^{M} B_i \theta_i x_m(t) \right) = -2 \beta \sum_{m=1}^{M} B_i \theta_i \sum_{m=1}^{M} \left( x_m(t) - \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i} \right)^2.
\]

It follows that \( \frac{dV(x(t))}{dt} < 0 \) for \( x(t) \neq x^* \), which completes the proof. \( \square \)

D. Proof of Theorem 5

Since the users are homogeneous and receive the same expected payoff if they choose the same channel. In this case, the optimization problem in (6) is equivalent to the following problem:

\[
\max_{x} \min \left\{ \frac{B_1 \theta_m}{N x_m}, \ldots, \frac{B_M \theta_m}{N x_m} \right\} \quad \text{subject to} \quad \sum_{m=1}^{M} x_m = 1, \quad x_m > 0, \forall m \in \mathcal{M}.
\]
Furthermore, this problem is equivalent to the following problem:

$$\begin{align*}
\max_{x} & \quad C \\
\text{subject to} & \quad \sum_{m=1}^{M} x_m = 1, \quad (23) \\
& \quad x_m > 0, \forall m \in \mathcal{M}, \quad (24) \\
& \quad \frac{B_m \theta_m}{N x_m} \geq C, \forall m \in \mathcal{M}. \quad (25)
\end{align*}$$

First, it is easy to check that $C > 0$. Then from the constant (26), we have

$$x_m \leq \frac{B_m \theta_m}{NC}, \forall m \in \mathcal{M}. \quad (27)$$

Summing over $m = 1, \ldots, M$, we obtain $C \leq \frac{\sum_{m=1}^{M} B_m \theta_m}{N}$. We then check that $x^* = \left(\frac{B_1 \theta_1}{\sum_{i=1}^{M} B_i \theta_i}, \ldots, \frac{B_M \theta_M}{\sum_{i=1}^{M} B_i \theta_i}\right)$ is the optimal point that attains the global maximum value

$$C^* = \frac{\sum_{m=1}^{M} B_m \theta_m}{N}. \quad (28)$$

Next, we prove the uniqueness of the solution. By substituting (28) into (27), we obtain $x_m \leq \frac{B_m \theta_m}{\sum_{i=1}^{M} B_i \theta_i}$. Suppose that there exists a solution $x' = [x'_1, \ldots, x'_M]$ and $x' \neq x^*$, such that it also maximizes the problem (23). In this case, there must exist some $j \in \mathcal{M}$ with $x'_j \neq x^*_j$, which implies

$$x'_j < \frac{B_j \theta_j}{\sum_{i=1}^{M} B_i \theta_i}.$$ 

By summing over the $j = 1, \ldots, M$, we must have

$$\sum_{j=1}^{M} x'_j < 1,$$

which forms a contraction with the constraint (24). Thus, $x^*$ is the unique optimal point for the optimization problem (22).

$\blacksquare$

E. Proof of Lemma 2

First, (14) can be written compactly as

$$\frac{P(T + 1) - P(T)}{\mu_T} = \left[Z(T) - P(T)\right]. \quad (29)$$
By taking $\mu_T \to 0$ and the expectation given the set of learning history $\{P(0), Z(i) - P(i), i < T\}$ of the RHS in (14), we obtain the mean dynamics of (14) as

$$\dot{P}(T) = E[Z(T) - P(T)|P(0), Z(i) - P(i), i < T].$$  \hspace{1cm} (30)

According to (13), $P(T)$ is determined by $Z(T - 1)$ and $P(T - 1)$, and $Z(T)$ is determined by $P(T)$. Thus,

$$E[Z(T) - P(T)|P(0), Z(i) - P(i), i < T] = E[Z(T)|P(T)] - P(T).$$

By conditioning on the event $\{a_n(T) = m\}$ for each $m \in M$ and $n \in N$, we have

$$E[Z^n_m(T)|P(T)] = E_{\{a_n(T) = m\}}\{E[Z^n_m(T)|P(T), a_n(T) = m]\}$$

$$= Pr \{a_n(T) = m\} E[Z^n_m(T)|P(T), a_n(T) = m] + Pr \{a_n(T) \neq m\} E[Z^n_m(T)|P(T), a_n(T) \neq m]$$

$$= \sigma^n_m(T)E[\tilde{U}_n(T)|P(T), a_n(T) = m] + (1 - \sigma^n_m(T))P^n_m(T), \forall n \in N, m \in M. \hspace{1cm} (31)$$

When the length of a decision period is large enough, we have $\frac{1}{k_m(T)} - \frac{1}{k_m(T)} \sim N \left(0, \frac{1}{\sum_{i=1}^{L} \frac{1}{s^n_{m_i}(T)}}\right)$ and $\tilde{\theta}_m - \theta_m \sim N \left(0, \frac{\sigma_m(1 - \theta_m)}{CL}\right)$. Since $\tilde{\theta}_m$ and $\frac{1}{k_m(T)}$ are independent, by conditioning on $\tilde{\theta}$ and $\frac{1}{k_m(T)}$ respectively, we have

$$E[\tilde{U}_n(T)|P(T), a_n(T) = m] = E_{\tilde{\theta}} \{E[\frac{B_m\tilde{\theta}_m}{k_m(T)}P(T), a_n(T) = m, \tilde{\theta}]\} = E[\frac{B_m\tilde{\theta}_m}{k_m(T)}P(T), a_n(T) = m]$$

$$= E_{\tilde{\theta}} \{E[\frac{B_m\tilde{\theta}_m}{k_m(T)}P(T), a_n(T) = m, \frac{1}{k_m(T)}]\} = E[\frac{B_m\tilde{\theta}_m}{k_m(T)}P(T), a_n(T) = m] = (\tilde{\theta} \tilde{\theta})$$

Then substituting (32) into (31), we have

$$E[U_n(T)|P(T), a_n(T) = m] = \sigma^n_m(T)E[U_n(T)|P(T), a_n(T) = m] + (1 - \sigma^n_m(T))P^n_m(T). \hspace{1cm} (33)$$

Combining (30), (31) with (33), we have (17) holds.

\[ \square \]

F. Proof of Lemma 3

Recall that $G^n_m(P(T)) = E[U_n(T)|P(T), a_n(T) = m]$ is the expected payoff of user $n$ choosing channel $m$ given the perceptions of all users $P(T)$. Let $k^{-n}_m(T)$ denote the number of users accessing
channel $m$ except user $n$. Thus,

$$G^m_n(P(T)) = \sum_{i=0}^{N-1} \frac{B_m n_i}{1+i} \Pr\{k^{-n}_m(T) = i | P(T), a_n(T) = m\}, \quad (34)$$

and $\Pr\{k^{-n}_m(T) = i | P(T), a_n(T) = m\}$ follows the Binomial distribution of $N - 1$ (without user $n$) independent non-homogeneous Bernoulli tries, which can be computed as

$$\Pr\{k^{-n}_m(T) = i | P(T), a_n(T) = m\} = \sum_{\sum n'_{i=1,n' \neq n} I_{\{a_{n'}(T) = m\}} = i} (\sigma^m_n(T))^{1-I_{\{a_n(T) = m\}}} (1 - \sigma^m_n(T))^{1-I_{\{a_n(T) = m\}}} . \quad (35)$$

We now consider the difference $G^m_n(P(T)) - G^m_n(\hat{P}(T))$ given two arbitrary perceptions $P(T)$ and $\hat{P}(T)$:

$$|G^m_n(P(T)) - G^m_n(\hat{P}(T))| = B_m n_i \sum_{i=0}^{N-1} \frac{\sum_{\sum n'_{i=1,n' \neq n} I_{\{a_{n'}(T) = m\}} = i} (\sigma^m_n(T))^{1-I_{\{a_n(T) = m\}}} (1 - \sigma^m_n(T))^{1-I_{\{a_n(T) = m\}}} }{1+i}$$

$$- \sum_{j=0}^{N-1} \frac{\sum_{\sum n'_{i=1,n' \neq n} I_{\{a_{n'}(T) = m\}} = j} (\sigma^m_n(T))^{1-I_{\{a_n(T) = m\}}} (1 - \sigma^m_n(T))^{1-I_{\{a_n(T) = m\}}} }{1+j} \leq B_m n_i \sum_{i=0}^{N-1} \frac{N - 1}{\sum_{n'=1,n' \neq n} |\sigma^m_n(T) - \hat{\sigma}^m_n(T)|} \sum_{n'=1,n' \neq n} |\sigma^m_n(T) - \hat{\sigma}^m_n(T)|. \quad (36)$$

We then define a function $f(P_{n'}(T)) \triangleq \sigma^m_n(T) = \frac{e^{\gamma P_{n'}(T)}}{\sum_{i=1}^M e^{\gamma P_{i}(T)}}$. Since $f(P_{n'}(T))$ is continuously differentiable, by the mean value theorem, we know that there exists $\bar{P}_{n'}(T) = \delta(P_{n'}(T) - \hat{P}_{n'}(T))$ with $0 < \delta < 1$ such that

$$\sigma^m_n(T) - \hat{\sigma}^m_n(T) = \frac{e^{\gamma P_{n'}(T)}}{\sum_{i=1}^M e^{\gamma P_{i}(T)}} - \frac{e^{\gamma \bar{P}_{n'}(T)}}{\sum_{i=1}^M e^{\gamma P_{i}(T)}}$$

$$= \gamma \left[ \frac{\sum_{i=1}^M e^{\gamma P_{i}(T)}}{\sum_{i=1}^M e^{\gamma P_{i}(T)}} \cdot \left( \frac{\sum_{i=1}^M e^{\gamma P_{n'}(T)}}{\sum_{i=1}^M e^{\gamma P_{i}(T)}} - \frac{e^{2\gamma P_{n'}(T)}}{\sum_{i=1}^M e^{\gamma P_{i}(T)}} \right)^2 \right] (P_{n'}(T) - \hat{P}_{n'}(T))$$

$$- \sum_{m'=1,m' \neq m} \gamma \frac{e^{\gamma P_{n'}(T)}}{\sum_{i=1}^M e^{\gamma P_{i}(T)}} \cdot \frac{e^{\gamma P_{m'}(T)}}{(\sum_{i=1}^M e^{\gamma P_{i}(T)})^2} (P_{m'}(T) - \hat{P}_{m'}(T)).$$

Let $C_m = \frac{e^{\gamma P_{m'}(T)}}{\sum_{i=1}^M e^{\gamma P_{i}(T)}} \cdot \frac{e^{\gamma P_{m'}(T)}}{\sum_{i=1}^M e^{\gamma P_{i}(T)}}$ and $C_{m'} = \frac{e^{\gamma P_{m'}(T)}}{\sum_{i=1}^M e^{\gamma P_{i}(T)}}, \forall m' \neq m$. It is easy to check that
\[ C_m = \sum_{m' = 1, m' \neq m}^M C_{m'} \quad \text{and} \quad 2C_m \leq 1. \] 

Thus,

\[ |\sigma_m'(T) - \tilde{\sigma}_m'(T)| \leq \gamma C_m |P_m'(T) - \tilde{P}_m'(T)| + \sum_{m' = 1, m' \neq m}^M \gamma C_{m'} |P_{m'}'(T) - \tilde{P}_{m'}'(T)| \]

\[ \leq \gamma (C_m + \sum_{m' = 1, m' \neq m}^M C_{m'}) ||P_n'(T) - \tilde{P}_n'(T)||_\infty \leq \gamma ||P_n'(T) - \tilde{P}_n'(T)||_\infty. \tag{37} \]

Combining (36) and (37), we obtain

\[ |G_m^n(P(T)) - G_m^n(\tilde{P}(T))| \leq \gamma B_m \theta_m (N - 1)(\ln N + 1) \sum_{n' = 1, n' \neq n}^N ||P_{n'}'(T) - \tilde{P}_{n'}'(T)||_\infty \]

\[ \leq \gamma B_m \theta_m (N - 1)^2(\ln N + 1)||P(T) - \tilde{P}(T)||_\infty. \]

It follows that if \( \gamma < \frac{1}{\max_{m \in \mathcal{M}}(B_m \theta_m)(N-1)^2(\ln N + 1)} \), the mapping \( [G_m^n(P(T)), \forall m \in \mathcal{M}, n \in \mathcal{N}] \) forms a maximum-norm contraction. \( \square \)

**G. Proof of Theorem 6**

We complete the proof by checking the assumptions of Theorem 2.1 in [24, pp.127].

(a) Since \( 0 \leq \tilde{\theta}_m \leq 1 \) and \( \tilde{k}_m(T) > 0 \), then \( \tilde{U}_n(T) = \frac{B_m \tilde{\theta}_m}{\tilde{k}_m(T)} \) must be bounded.

From (13) and (15), we have \( P_m^n(T) < \infty \) and \( Z_m^n(T) < \infty \). It follows that \( |Z(T) - P(T)| < \infty \).

Thus, \( \sup_T E[|Z(T) - P(T)|^2] < \infty \).

(b) By (30) and Lemma 2, we have

\[ E[Z_m^n(T) - P_m^n(T)|P(0), Z(i) - P(i), i < T] = \sigma_m^n(T)[G_m^n(P(T)) - P_m^n(T)]. \] \( \tag{38} \)

(c) First, \( \sigma_m^n(T) = \frac{e^{P_m^n(T)}}{\sum_{i = 1}^T e^{P_i^n(T)}} \) is differentiable with respective to \( P_m^n(T) \).

Second, \( G_m^n(P(T)) = E[U_n(T)|P(T), a_n(T) = m] \) is an expectation function, which can be calculated by (34) and (35). Since \( \sigma_m^n(T) \) is differentiable, thus \( P_T\{k_m^n(T) = i|P(T), a_n(T) = m\} \) is also differentiable because summation of differentiable functions is also differentiable. By the same argument, \( G_m^n(P(T)) \) is also differentiable. Thus, \( \sigma_m^n(T)[G_m^n(P(T)) - P_m^n(T)] \) is differentiable and hence continuous.

(d) We have \( \sum_T \mu_T = \infty \) and \( \sum_T \mu_T^2 < \infty \) by assumption.

(e) By (38) and Lemma 2 we know that the expected biased error \( \beta_T = 0 \). It follows that \( \sum_T \mu_T|\beta_T| < \infty \) with probability one.

(f) Since the temperature \( \gamma < \frac{1}{\max_{m \in \mathcal{M}}(B_m \theta_m)(N-1)^2(\ln N + 1)} \), \( G(P(T)) \) is a contraction mapping, with
a unique fixed point $P^*$. Since $\dot{P}_m^n(T) = \sigma_m^n(T)[G_m^n(P(T)) - P_m^n(T)]$ and $\sigma_m^n(T) = \frac{e^{\gamma P_m^n(T)\sum_{i=1}^M e^{\tau_T}(T)}}{\sum_{i=1}^M e^{\tau_T}(T)} > 0$, then the differential equations in (18) have only one limit point $P_m^n$ by setting $\dot{P}_m^n(T) = 0$. We can then define the function $f$ as the integral of $\sigma_m^n(T)[G_m^n(P(T)) - P_m^n(T)]$. It is continuously differentiable since $\sigma_m^n(T)[G_m^n(P(T)) - P_m^n(T)]$ is continuous. Because there exists only one point in the limit set, the function $f$ is a constant on the set. 

H. Proof of Theorem 8

We choose the common Lyapunov function for contraction mapping [23], which is defined as $V(P(T)) = \max_{n,m}\{ |P_m^n(T) - P_m^n| \}$. Since $P^*$ is the unique fixed point of the maximum-norm contraction map $G(P(T))$, it follows that $V(P^*) = 0, V(P(T)) > 0, \forall P(T) \neq P^*$.

Let $(n^*, m^*)$ be solution that maximizes the function $V(P(T))$, i.e., $V(P(T)) = |P_m^n(T) - P_m^n|$. We first consider the case that $P_m^n(T) \geq P_m^n$. Then $V(P(T)) = P_m^n(T) - P_m^n$, and

$$\frac{dV(P(T))}{dT} = \frac{d(P_m^n(T) - P_m^n)}{dT} = \frac{dP_m^n(T)}{dT}.$$ 

From (18), we have $\frac{dV(P(T))}{dT} = \frac{dP_m^n(T)}{dT} = \sigma_m^n(T)[G_m^n(P(T)) - P_m^n(T)].$ Since $G_m^n(P(T))$ is a maximum norm contraction, there exist a Lipschitz constant $0 \leq \xi < 1$ such that $G_m^n(P(T)) - G_m^n(P^*) \leq \xi(P_m^n(T) - P_m^n)$, and $G_m^n(P^*) = P_m^n$. Thus,

$$\frac{dV(P(T))}{dT} = \sigma_m^n(T)[G_m^n(P(T)) - P_m^n(T)] = \sigma_m^n(T)[G_m^n(P(T)) - G_m^n(P^*) + P_m^n - P_m^n(T)] \leq \sigma_m^n(T)\xi(P_m^n(T) - P_m^n) + P_m^n - P_m^n(T) = -(1 - \sigma_m^n(T)\xi)V(P(T)) < 0, \forall P(T) \neq P^*.$$

Similarly, for the case that $P_m^n(T) < P_m^n$, we can also show that $\frac{dV(P(T))}{dT} < 0, \forall P(T) \neq P^*$. Thus, the function $V(P(T))$ is a strict Lyapunov function, and hence the conclusion holds.

I. Proof of Theorem 9

We consider the following optimization problem for each user $n \in N$:

$$\max_{\sigma^n = [\sigma^n_1, ..., \sigma^n_M]} \sum_{m=1}^M \sigma^n_m E[U_n(T)|P^*, a_n(T) = m] - \frac{1}{\gamma} \sum_{m=1}^M \sigma^n_m \ln \sigma^n_m$$

subject to $\sum_{m=1}^M \sigma^n_m = 1, \sigma^n_m \geq 0, \forall m \in M$. 

The objective is to choose a mixed strategy \( \sigma^n \) for user \( n \) to maximize the expected payoff off the term \( \frac{1}{\gamma} \sum_{m=1}^M \sigma^m_n \ln \sigma^m_n \), given the perceptions \( \mathbf{P}^* \) (i.e., other users’ mixed strategies since \( \sigma^m_n = \frac{e^{\gamma P^*_m}}{\sum_{i=1}^M e^{\gamma P^*_i}}, \forall n' \neq n \)). The constraint is to ensure that the mixed strategy is feasible.

By the KKT condition, we obtain the optimal solution as

\[
\hat{\sigma}_m^n = \frac{e^{\gamma E[U_n(T)|\mathbf{P}^*, a_n(T) = m]}}{\sum_{i=1}^M e^{\gamma E[U_n(T)|\mathbf{P}^*, a_n(T) = i]}} \quad \forall m \in \mathcal{M}.
\]

From (20), we have

\[
\hat{\sigma}_m^n = \frac{e^{\gamma P^*_m}}{\sum_{m=1}^M e^{\gamma P^*_m}} = \sigma^m_n. \quad \text{It is known that} \quad \max_{\sigma^n} \left\{ \sum_{m=1}^M \sigma^m_n E[U_n(T)|\mathbf{P}^*, a_n(T) = m] - \frac{1}{\gamma} \sum_{m=1}^M \sigma^m_n \ln \sigma^m_n \right\} \geq \max_{\sigma^n} \sum_{m=1}^M \sigma^m_n E[U_n(T)|\mathbf{P}^*, a_n(T) = m] - \xi, \forall n \in \mathcal{N},
\]

where \( \xi = \max_{n \in \mathcal{N}} \{-\frac{1}{\gamma} \sum_{m=1}^M \sigma^m_n \ln \sigma^m_n \}. \)

\[ \Box \]

J. Spectrum access game is a potential game

**Definition 4 ([28]).** A game \( (\mathcal{N} = \{1, ..., N\}, \mathcal{A} = \mathcal{A}_1 \times ... \times \mathcal{A}_N, U_n : \mathcal{A} \to \mathcal{R}) \) is a potential game if there is an exact potential function \( \phi : \mathcal{A} \to \mathcal{R} \) such that

\[
\phi(a_n', a_{-n}) - \phi(a_n, a_{-n}) = U_n(a_n', a_{-n}) - U_n(a_n, a_{-n}), \forall a_n, a_n' \in \mathcal{A}_n, n \in \mathcal{N},
\]

where \( a_{-n} \) denotes the strategies of other players except player \( n \).

For the spectrum access game, we show that

**Lemma 4.** The spectrum access game is a potential game.

**Proof:** We define the potential function as a function of the population state \( \mathbf{x} \):

\[
\phi(\mathbf{x}) = \sum_{m=1}^M \sum_{i=1}^{N_x_m} B_m \theta_m^i.
\]

Consider a user \( n \) who unilaterally moves from strategy \( a_n \) to strategy \( a_n' \neq a_n \). Then change of potential
function is

\[
\Delta \phi (a_n \rightarrow a'_n) = \sum_{i=1}^{N_x a'_n + 1} \frac{B_m \theta_m}{i} - \sum_{i=1}^{N_x a_n} \frac{B_m \theta_m}{i} + \sum_{i=1}^{N_x a_n - 1} \frac{B_m \theta_m}{i} - \sum_{i=1}^{N_x a_n} \frac{B_m \theta_m}{i} = U_n (a'_n, x) - U_n (a_n, x),
\]

which completes the proof.

REFERENCES


TABLE I

EVOLUTIONARY SPECTRUM ACCESS MECHANISM

1: **initialization:**
2: set the global strategy adaptation factor $0 < \alpha \leq 1$.
3: select a random channel for each user.
4: **end initialization**

5: **loop** for each time slot $t$ and each user $n \in \mathcal{N}$ in parallel:
6: sense the selected channel.
7: if the channel is idle then
8: compete for the chosen channel and transmit data packets if successfully grabbing the channel.
9: end if
10: broadcast the chosen channel ID to other users through the common control channel.
11: receive the information of other users’ channel selection and calculate the population state $x(t)$.
12: compute the expected payoff $U_n(a_n, x(t)) = \frac{B_{an} \theta_{an}}{N x_{an}(t)}$.
13: if $U_n(a_n, x(t)) < \frac{1}{\sum_{i=1}^M B_i \theta_i}$ then
14: generate a random value $\delta$ according to a uniform distribution on $(0, 1)$.
15: if $\delta < \alpha(1 - \frac{U_n(a_n, x(t))}{\frac{1}{\sum_{i=1}^M B_i \theta_i}})$ then
16: select a new channel $m$ with probability

$$p_m = \frac{\left[ \frac{B_{an} \theta_{an}}{\sum_{i=1}^M B_i \theta_i} - x_{m}(t) \right]^+}{\sum_{m'=1}^M \left[ \frac{B_{an} \theta_{an}}{\sum_{i=1}^M B_i \theta_i} - x_{m'}(t) \right]^+},$$

where $[c]^+ = \max\{c, 0\}$.
17: else select the original channel.
18: end if
19: end if
20: **end loop**
TABLE II
DISTRIBUTED REINFORCEMENT LEARNING MECHANISM FOR DYNAMIC SPECTRUM ACCESS

1: initialization:
2: set the temperature $\gamma$.
3: set the initial perception values $P_n(0) = 0$ for each user $n \in \mathcal{N}$.
4: set the period index $T = 0$.
5: end initialization

6: loop for each decision period $T$ and each user $n \in \mathcal{N}$ in parallel:
7: select a channel $m \in \mathcal{M}$ according to (12).
8: for each time slot $t$ in the period $T$ do
9: sense and compete to access the channel $m$.
10: record the observations $S_n^m(T, t)$ and $I_n^m(T, t)$.
11: end for
12: estimate $k_m(T)$ and $\theta_m$ by the maximum likelihood estimation according to (7) and (10).
13: compute the estimated expected payoff $\tilde{U}_n(T)$.
14: update the perceptions value $P_n(T)$ according to (13).
15: set the period index $T = T + 1$.
16: end loop
Fig. 1. Structure of each time slot.

Fig. 2. Illustration of dynamic spectrum access. User B accesses channel 2, which is occupied by primary users. Users A and D access the same idle channel 3, and thus lead to collision and a reduced expected throughput for each other. User C accesses an idle channel 5 without any collision, and achieves the maximum throughput on that channel.

Fig. 3. The number of iterations for the evolutionary spectrum access mechanism with different strategy adaptation factor $\alpha$.

Fig. 4. The fraction of users on different channels with the number of users $N = 20$.

Fig. 5. The fraction of users on different channels with the number of users $N = 40$. 
Fig. 6. The expected payoff of choosing different channels with the number of users $N = 20$.

Fig. 7. The expected payoff of choosing different channels with the number of users $N = 40$.

Fig. 8. Stability of the evolutionary spectrum access mechanism. Among a total of $N = 40$ users, 50% choose mutant channels randomly at time slot 30.

Fig. 9. Stability of the evolutionary spectrum access mechanism. Among a total of $N = 40$ users, 90% users choose mutant channels randomly at time slot 30.

Fig. 10. Maximum likelihood estimation of the number of secondary users $k_m(T)$ on a channel. The vertical bar represents the range of the estimated values.

Fig. 11. Maximum likelihood estimation of the channel idle probability $\theta_m$. The vertical bar represents the range of the estimated values.
Fig. 12. The system performance of the distributed reinforcement learning algorithm with different temperature $\gamma$.

Fig. 13. Users’ time average payoff in the distributed reinforcement learning algorithm with the number of users $N = 3$. 
Fig. 14. Users’ time average payoff in the distributed reinforcement learning algorithm with the number of users $N = 20$.

Fig. 15. Stability of the distributed reinforcement learning algorithm. 2 out of 3 users reset their perception values at the period $T = 1000$.

Fig. 16. Stability of the distributed reinforcement learning. 10 out of 20 users reset their perception values at the period $T = 1000$.

Fig. 17. Average system payoff by the evolutionary spectrum access mechanism, distributed reinforcement learning algorithm and random access algorithm with different number of users.