Regret of Multi-Channel Bandit Game in Cognitive Radio Networks

Jun Ma and Yonghong Zhang

School of Electronic Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China
Email: majunpaper@foxmail.com; zhangyhh@uestc.edu.cn

Abstract. The problem of how to evaluate the rate of convergence to Nash equilibrium solutions in the process of channel selection under incomplete information is studied. In this paper, the definition of regret is used to reflect the convergence rates of online algorithms. The process of selecting an idle channel for each secondary user is modeled as a multi-channel bandit game. The definition of the maximal averaged regret is given. Two existing online learning algorithms are used to obtain the Nash equilibrium for each SU. The maximal averaged regrets are used to evaluate the performances of online algorithms. When there is a pure strategy Nash equilibrium in the multi-channel bandit game, the maximal averaged regrets are finite. A cooperation mechanism is also needed in the process of calculating the maximal averaged regrets. Simulation results show the maximal averaged regrets are finite and the online algorithm with greater convergence rate has less maximal averaged regrets.

Keywords-cognitive radio networks; adversarial bandit problem; congestion game; online learning; dynamic spectrum access.

1 Introduction

With the emergences of new wireless services and applications, the demand for spectrum increases. However, some studies show that many licensed spectrum bands have not been utilized efficiently [1]. In order to alleviate this contradiction, cognitive radio users called secondary users (SUs) are proposed and allowed to access spectrum bands belong to primary users (PUs) as long as those bands are sensed to be idle. How to select a proper channel to sense and access will affect the idle spectrum usage. Therefore, the channel selection problem is important to each SU.

Firstly, some works about channel selection under incomplete information are introduced. In [2], the problem of how to select a channel set for a SU to access at a time has been investigated under the condition that the statistical information of PUs' traffic is assumed as a stationary and simple distribution which is unknown to SUs in advance. This work is mainly based on classical bandit models [3]. Due to the stochastic nature of cognitive radio networks (CRNs), the real primary traffic distribution is not always stationary. Under this scenario, the channel set selection problem is analyzed in [4]. Furthermore, a similar problem of multiple SUs who need to select and access a channel at a time is studied in [5] considering the effect of the interactions among SUs on individual benefits. Previous works have studied the channel selection problem with incomplete information by bandit models.

From the perspective of game theory, the channel selection problem can be also modeled as a congestion game. Congestion game is a kind of non-cooperative game with the utility of each player by using a certain resource depending on the total number of players who are using the same resource [6], [7]. Some works in the field of game theory have shown that congestion games are a kind of potential game in fact [8]. Potential games are used to study the channel selection problem in [9] and the network selection problem [10] while some online learning algorithms are applied to obtaining the Nash equilibrium (NE) solutions. Potential games always have potential functions which can guarantee at least a pure strategy NE [8]. Although these works claim that NE can be found under incomplete information, they need more information than our model because the potential functions will be given before searching the NE. These works do not consider the convergence rate of the online algorithm. The online algorithm with greater convergence rate will reduce the search time while increase the transmission data time. Therefore, the problem of how to evaluate the convergence rate of online algorithm under incomplete information should be studied.

In this paper, the channel selection problem with incomplete information is modeled as a multi-channel bandit game from the view of bandit models The *MAR* is used to evaluate the convergence rates of online algorithms.

2 System Model and Problem Formulation

Consider a CRN with N SUs as well as S SU base stations and M (M=S<N) primary channels which belong to PUs. All SUs and PUs operate in the slot transmission structure. Each PU has only one channel. If any one of SU nodes wants to communicate with a SU base station, it must use the idle slots of primary channels. A primary channel just serves a SU base station. For example, there are N=3, M=S=2 in a CRN. When SU n=1 communicates with SU base station A, SU node n=1 can only utilize the idle slots of channel m=1 which belongs to PU 1. Similarly, if SU n=2 communicates with SU base station B, SU n=1 can only utilize the idle slots of channel 2 which belongs to PU 2. Here, we assume connecting different SU base station do not affect the further communication with the destination of each SU. The spectrum sensing is assumed to be perfect for each SU and all primary channels are idle during the total simulation time for ease of researching. This assumption is rational. The idle time of some primary channels are much longer than the data transmission durations of secondary users who have little data to transmit (like the temperature sensors). The effect of primary traffics on the performances of convergence rates of online algorithms will be studied in our future works.

Each SU selects a channel from M channels at a time to sense and access if the channel is sensed to be idle and the number of using this channel is not too much. Each SU is rational and selfish with the goal of maximizing its own averaged transmission rate during the process of channel selection. At each slot, each SU selects a channel at a time and receives a reward if the channel is accessed, which is analogous to the process of a gambler selecting the arm of a slot machine and receiving a reward. If a SU transmits successfully, the reward is the channel transmission rate which is given by (1) and 0 for the transmission failure. The channel transmission rate of SU n at slot t in channel m can be written in (1), where W is the channel bandwidth, P is the transmit power and σ^2 is the thermal noise level, which are same to all SUs for the simplification of research. Note that $g^t_{n,m}$ denotes the channel gain at slot t and changes over the time but keeps unchanged during each slot.

$$r^t_{n,m} = W \log_2\left(1+\frac{Pg^t_{n,m}}{\sigma^2}\right) \tag{1}$$

When more than one SUs access the same channel simultaneously, the individual transmission success probability decreases because of the mutual interference. Therefore, the rewards obtained by SUs are affected by the number of SUs using the same channel at the same time. We assume $\boldsymbol{s}^t = (s^t_1, \dots, s^t_N)$ is a pure strategy profile at slot t for all SUs, where s^t_n denotes the selection strategy of SU n at slot t. $\mathbf{c}^t(\boldsymbol{s}^t) = (c^t_1, \dots, c^t_M)$ denotes the total number of SUs in each channel corresponding to the strategy profile $\boldsymbol{s}^t$ at slot t. The total number c^t_m affects the successful transmission probability $p(c^t_m)$ of SUs who use the channel m at slot t. When c^t_m increases, $p(c^t_m)$ of SUs decreases. In this paper, we adopt a common MAC protocol (the slotted Aloha) [7] and the specific expression of $p(c^t_m)$ is given in (2). When only one SU is in the channel m, c^t_m=1, it is certain to transmit its data successfully with probability 1. As mentioned before, we model the channel selection problem as a bandit problem. However, the classic bandit model is not fit for our situation because the rewards of each accessing a channel for each SU are not independently drawn from a fixed and unknown distribution.

$$p(\mathrm{c}^t_m)=\begin{cases} 0 & \mathrm{c}^t_m = 0 \\ 1 & \mathrm{c}^t_m = 1 \\ \left(\frac{1}{\mathrm{c}^t_m}\right)\left(1-\frac{1}{\mathrm{c}^t_m}\right)^{\mathrm{c}^t_m-1} & \mathrm{c}^t_m > 1 \end{cases} \tag{2}$$

Here, we utilize a variant of the classic bandit model called adversarial bandit [11] which is a non-stochastic bandit problem to model our scenario. The adversarial bandit is also used in [4] to find the optimal channel. We assume that all SUs use the same online algorithm to find the equilibrium channel for themselves at the same time. In our model, we use T to denote the slot at which each user makes a selection using online learning algorithms. This structure is illustrated in Fig.1. At slot T, each SU will select a channel to access based on a specific distribution over M channels. This specific distribution is decided by the updating rule of an online learning algorithm. After all SUs have completed the process of channel selections at slot T, the pure strategy profile can be represented by $\boldsymbol{s}^T$.

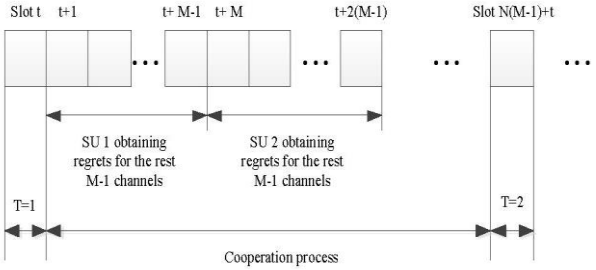


Figure 1. An illustration of the process of calculating regrets under incomplete information in our model

The following M-1slots will be used to calculate the regret for each channel which has not selected at slot T. In order to express our idea clearly, (j, s^T_{-n}) denotes $(s^T_1, \dots, s^T_{n-1}, j, s^T_{n+1}, \dots, s^T_N)$, which is $\boldsymbol{s}^T$. When SU n does not select a channel j at slot T, it may care about how much reward SU n has lost for not playing the pure strategy j at slot T under the condition that all the other SUs keep their selection strategies (s^T_{-n}) at slot T unchanged. If SU n knows its own payoff function and the strategy set $\boldsymbol{s}^T$ at slot T, SU n can calculate the regret of not selecting the channel j and the regret $R^T_n(j,T)$ is given as follows [11],

$$R^T_n(j,T) = U_n\left(j, s^T_{-n}\right) - U_n(\boldsymbol{s}^T) \tag{3}$$

Where U_n denotes the SU n's reward function, $U_n = r^t_{n,m}$ if SU n transmits successfully and otherwise U_n=0. However, in our model, each SU does not know its own payoff function. Hence, each SU needs the cooperation

with other SUs. The cooperation among SUs will help each SU to find the equilibrium channel, which has been showed in Fig.1. Therefore, each SU has an incentive to cooperate with others. We use an example to explain the cooperation process. For example, if SU *n* does not select the channel *j* at slot *T*, it may select channel *j* at slot *t*+1 while other SUs should select the channels which have been chosen at slot *T*, namely, s_{-n}^{T}. In other words, the channel selection process will be repeated at slot *t*+1 and SU *n* will select channel *j* without doubt while other SUs keep s_{-n}^{T} unchanged. Therefore, SU *n* will repeat this process for *M*-1 slots because there are *M*-1 channels not be chosen at slot *T*. Hence, the next *M*-1slots are used to calculate the regrets which are produced by not selecting the channel j ($j \in M-1$) at slot *T*. The selecting order for *M*-1 channels is decided by each SU. All SUs know *M* and When SU *n* completes the regret calculations for *M*-1 channels, SU *n*+1 will be informed and continue to repeat the same process with SU *n* for *M*-1 slots until SU *N* completes this process. The next time for all SUs using the online algorithm to select a channel is at slot *t*+*N*(*M*-1). For example, in the Fig.1, the secondary selection time decided by online algorithms is *T*=2. In our model, if the strategy set (j, s_{-n}^{T}) is the same with $\boldsymbol{s}^{T}$, the strategy *j* is no need to try again. The total number of selecting the channel decided by online algorithms is *Ta*. If the online algorithm selects total *T*=*Ta* times, the averaged regret of SU *n* for channel j ($j \in M$) after *Ta* times is as follows,

$$\begin{aligned}\overline{R_n^j}(Ta) &= 1/Ta\left(\sum_{T=1}^{Ta}[U_n(j,s_{-n}^{T}) - U_n(\boldsymbol{s}^{T})]\right) \\ &= 1/Ta\left(\sum_{T=1}^{Ta}U_n(j,s_{-n}^{T}) - \sum_{T=1}^{Ta}U_n(\boldsymbol{s}^{T})\right)\end{aligned} \quad (4)$$

Here, we define the *MAR* as $R_n^{\max} = \max_j(\overline{R_n^j}(Ta))$ which is the maximal averaged regret and use the *MAR* to evaluate the performances of online learning algorithms.

3 Online Learning Scheme

In this section, we use two existing online learning algorithms to find the NE solution and calculate the *MAR* for all SUs using two existing online learning algorithms. The first one is based on the algorithm Exp3 [11]. The second one is based on the stochastic learning algorithm [12].

3.1 Online Learning Algorithm Based on Exp3

Each SU calculates the selection probability distribution $P_n(T) = \left(p_{n,1}(T),\ldots,p_{n,\mathrm{M}}(T)\right)$ over *M* channels based on (5) according to the corresponding weight value for each channel. The weight value is updated by the normalized reward based on (8). Here, the normalized reward is the ratio between the actual reward and maximal probability reward which is constrained by the hardware of SUs. The maximal probability reward in our simulation part is produced by all random communication conditions.

Algorithm 1: Exp3 based online learning algorithm

1: Initialize total times *Ta*, parameter 0<*b*<1, $R_n(0)$=0, $R_{n,m}(0)$=0, and the weight value $\omega_{n,m}^{T=1}=1$ of each channel $m \in M$ for each SU $n \in N$.

2: Each SU randomly selects a channel according to the probability $p_{n,m}^{T}$ and its expression is as follows,

$$p_{n,m}^{T} = (1-\mathrm{b})\frac{\omega_{n,m}^{T}}{\sum_{m=1}^{M}\omega_{n,m}^{T}} + \frac{b}{M} \quad (5)$$

3: Each SU receives a reward r_n^{T}, if $r_n^{T} \neq 0$ and then

$$R_n(T) = R_n(T\text{-}1) + r_n^{T} \quad (6)$$

4: Each SU updates the weight value $\omega_{n,m}^{T+1}$ of each channel based on the normalized reward $\overline{r}_n^{T}$. The updating rule for each SU is as follows,

$$\hat{\Omega}_{n,m}^{T} = \frac{\overline{r}_n^{T}}{p_{n,m}^{T}} \quad (7)$$

$$\omega_{n,m}^{T+1} = \omega_{n,m}^{T}\exp\left(b\frac{\hat{\Omega}_{n,m}^{T}}{M}\right) \quad (8)$$

5: Cooperation begin

From n=1 to N, each SU *n* selects channel *m* from the set of channels *M* according to the predetermined order and the selection strategy of other SUs s_{-n}^{T} is always unchanged. SU *n* can obtain the rewards for each channel from the set of channels *M*, If the channel *m* has been selected at slot *T*, the reward is $U_n(m,s_{-n}^{T}) = r_n^{T}$, else

$$R_{n,m}(T)=R_{n,m}(T\text{-}1)+U_n(m,s_{-n}^{T}) \quad (9)$$

When SU *n* completes its process of calculating the $R_{n,m}(t)$, the SU *n*+1 is informed and continues until SU *N*.

6: Cooperation end

7: *T*=*T*+1 and back to Step 2.

8: Calculate the MAR based on (4) and find the *MAR*

3.2 Stochastic Learning Algorithm

Each SU selects a channel at random according to a dynamic distribution $P_n(T) = \left(p_{n,1}(T),\ldots,p_{n,\mathrm{M}}(T)\right)$ over *M* channels and uses the normalized reward to update the selection probability distribution for the next round based on (11). At the first slot (*T*=1), the probability distribution $\mathbf{P}_n(T)$ of selection channel is assumed as the uniform distribution over *M* channels.

Algorithm 2: stochastic learning algorithm

1: Initialize total times *Ta*, parameter 0<b<1, $R_n(0)$=0, $R_{n,m}(0)$=0 for each channel $m \in M$ for each SU $n \in N$, and $P_n(T=1)$ is the uniform distribution over *M* channels.

2: Let each SU select a channel randomly according to the probability distribution $\mathbf{P}_n(T)$.

3: Each SU receives a reward r_n^T, if $r_n^T \neq 0$ and then

$$R_n(T) = R_n(T-1) + r_n^T \quad (10)$$

4: Each SU updates $P_n(T+1)$ as follows,

$$\mathbf{P}_n(T+1) = \mathbf{P}_n(T) + b\,\overline{r}_n^T(I_m - \mathbf{P}_n(T)) \quad (11)$$

where b is the learning rate, $\overline{r}_n^T$ is the normalized reward, I_m is the indicator function.

5: Cooperation begin

From n=1 to N, each SU n selects channel m from the set of channels M and according to the predetermined order and the selection strategy of other SUs s_{-n}^T is always unchanged. SU n can obtain the rewards for each channel from the set of channels M,

If the channel m has been selected at slot T, the reward is $U_n(m, s_{-n}^T) = r_n^T$, else

$$R_{n,m}(T) = R_{n,m}(T\text{-}1) + U_n(m, s_{-n}^T) \quad (12)$$

When SU n completes its process of calculating the $R_{n,m}(t)$, the SU n+1 continues until SU N

6: Cooperation end

7: T=T+1 and back to Step 2.

8: Calculate the MAR based on (4) and find the *MAR*

During the cooperation, if SU n has selected channel m at slot T, SU n has no need to try again. Therefore, we think $U_n(m, s_{-n}^T) = r_n^T$.

3.3 The Relationship between NE and MAR

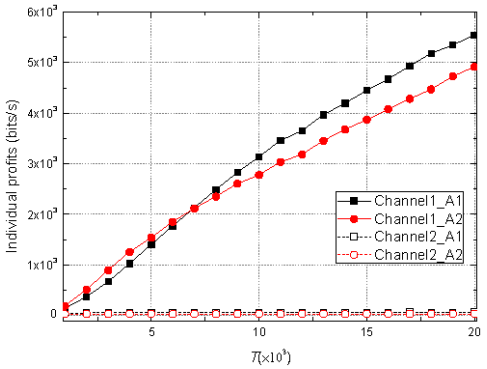
In this subsection, we use the definition of regret to evaluate the convergence rates of online algorithms. We can find based on the theorem 1 that the *MAR* will be reduced if the selection strategy converges to the equilibrium channel quickly.

Theorem 1: When the multi-channel bandit game has a pure strategy NE, the *MAR* of online algorithms for each SU is finite.

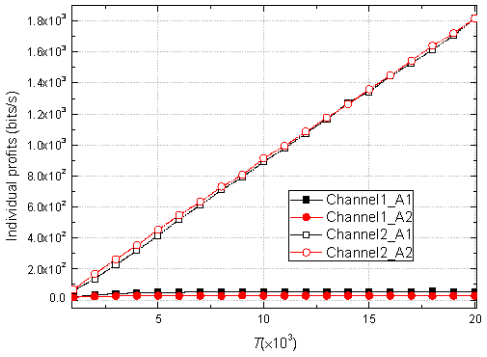
Proof: The total selection times which is decided by online algorithms is *Ta*. We assume channel J is the equilibrium strategy for SU n. At a special selection time T', each SU will select optimal channel J and does not change its selection after slot T'. In other words, when the total *Ta* is large enough, $\sum_{T=T'}^{Ta} R_n(T) = 1/(Ta - T')(U_n(J, s_{-n}^T) - U_n(\boldsymbol{s}^T)) = 0$ after each SU passes slot T'. This is because the (J, s_{-n}^T) is the same with $\boldsymbol{s}^T$ at slot T= T' and the reward of each accessing the optimal channel J is random for SU n. When all SUs do not change their selections, the result of searching optimal channel will converge to an NE SU n. Therefore, $\sum_{T=1}^{Ta}[U_n(J, s_{-n}^T) - U_n(\boldsymbol{s}^T)] < \infty$ and $R_n^{\max} = \max_j(\overline{R_n^j}(Ta)) < \infty$.

4 Simulation Results

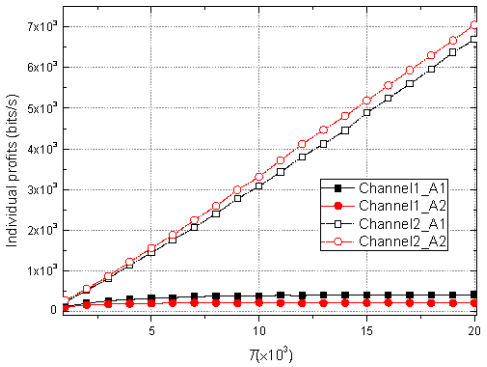
The performances of online learning algorithms are evaluated in this section. There are N=4 SUs and M=2 primary channels in a CRN covering a 500m×500m area. S=2 SU base stations locate in (0,250), (500,250) which are accessible for all SUs. When the 4 SUs are selected randomly, the locations of SUs are fixed during the simulations. All SUs adopt the channel model of Flat/Light tree density proposed in [13]. The transmission power is 10^{-3}W and the noise power level for all SUs is assumed to be 10^{-12}W. The bandwidth for all SUs is assumed as 1Hz and the learning rate b=0.01 for the two algorithms. In figures, *Ta* is showed in abscissa axis, which is from 1000 to 20000. Each *Ta* runs 100 simulations.



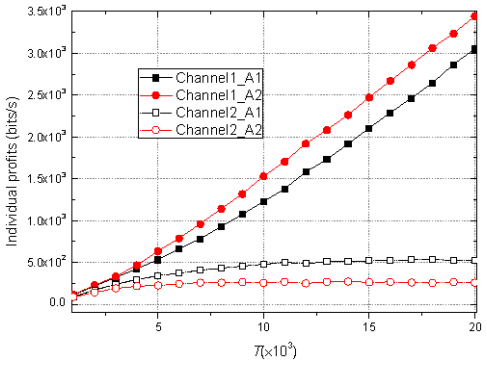
(a) The equilibrium channel for SU 1 is channel 1



(b) The equilibrium channel for SU 2 is channel 2



(c) The equilibrium channel for SU 3 is channel 2



(d) The equilibrium channel for SU4 is channel 1

Figure 2. Individual profits obtained from different channels using two online learning algorithms.

We can obtain NE based on (2) which is that SU1 and SU4 selects channel 1 as well as channel 2 for SU2 and SU3. The NE strategy can be written as (1, 2, 2, 1). First, we are interested in the behaviors of SUs in different channels. In Fig. 2, we show the individual accumulated transmission rates of each SU obtaining from channel 1 and channel 2 using algorithm 1 (A1) and algorithm 2 (A2). From the Fig. 2, we can see that each SU has a dominant strategy which is consistent with the pure strategy NE (1,2,2,1). From Fig. 2(a), we can find SU 1 has obtained more individual profits from the channel 1 regardless of A1 and A2. The same result is also showed in Fig. 2(d). Therefore, the channel 1 is the equilibrium solutions for both SU 1 and SU 4. In Fig. 2(b), SU 2 has received more individual profits from selecting the channel 2 and so dose SU 3 which is showed in Fig.2(c).

In Fig.3, we can note that all *MARs* are finite. In order to evaluate the convergence rates of two online algorithms, we compare the *MARs* of two online algorithms for all SUs in Fig. 3. From Fig.3, we can see the *MAR* of the equilibrium strategy for each SU by A1 is more than A2. Due to the lower convergence rate, each SU using A1 will select non-equilibrium channel more frequently, which will increase the *MAR*. The differences of *MAR*s between SUs using the same online algorithm are dependent on the different locations of SUs.

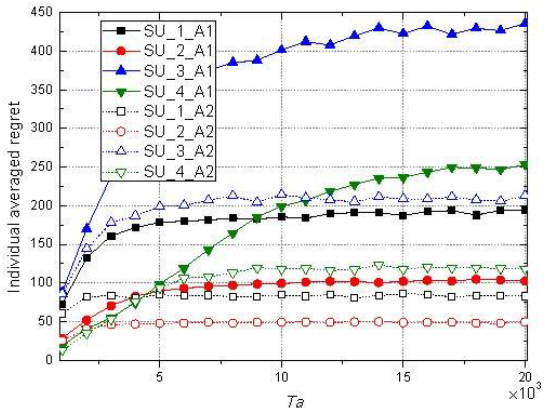


Figure 3. Comparison of 4 SUs' MAR of two selection strategies

5 Conclusion

In this paper, we have modeled the channel selection problem under incomplete information from the perspective of connection between then bandit model and the game model. When the channel selection game has a pure strategy NE, the *MAR* of selection strategies is finite. Two existing online learning algorithms are used to obtain the equilibrium channel for all SUs. The stochastic learning algorithm outperforms the Exp3 based online learning algorithm in terms of convergence rate to the Nash equilibrium solution, which is because the *MAR* of stochastic learning algorithm is less than Exp3. This work provides the connection between non-cooperative games and bandit problems, which will help us to illustrate the convergence rate of different online algorithm without knowing the payoff functions and other SUs' selection strategies.

References

1. B. Wang and K.J.R. Liu, "Advances in cognitive radio networks: A survey," IEEE J. Sel. Topics Signal Process., vol. 5, no. 1, pp. 5–23, Feb. 2011.
2. Z. Zhou, J. Hai, T. Peng, and J. Slevinsky, "Channel exploration and exploitation with imperfect spectrum sensing in cognitive radio networks," IEEE J. Sel. Areas Commun., vol. 31, no. 3, pp. 429–441, Mar. 2013.
3. Auer P, Cesa-Bianchi N, Fischer P. "Finite-time analysis of the multiarmed bandit problem," Mach. Learn., vol.47, no.2-3, pp. 235-256, 2002.
4. X. Fang, D. Yang, and G. Xue, "Taming wheel of fortune in the air: an algorithmic framework for channel selection strategy in cognitive radio networks," IEEE Trans. Veh. Technol., vol. 62, no. 2, pp. 783–796, Feb. 2013.
5. K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," IEEE Trans. Signal Process., vol. 58, no. 11, pp. 5667–5681, Nov. 2010.
6. L. Blumrosen and S. Dobzinski, "Welfare maximization in congestion games," IEEE J. Sel. Areas Commun., vol. 25, no. 6, pp. 1224–1236, Aug. 2007.
7. L. M. Law, J. Huang, and M. Liu, "Price of anarchy for congestion games in cognitive radio networks," IEEE Trans. Wireless Commun., vol. 11, no. 10, pp. 3778–3787, Oct. 2012.
8. D.Monderer and L.S.Shapley, "Potential games," Games Econ. Behav., vol. 14, pp. 124–143, 1996.
9. Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in unknown dynamic environment: a game-theoretic stochastic learning solution," IEEE Trans. Wireless Commun., vol. 11, no. 4, pp. 1380–1391, Apr. 2012.
10. T. Li-Chuan, C. Feng-Tsun, Z. Daqiang, et al., "Network selection in cognitive heterogeneous networks using stochastic learning," IEEE Commun. Lett., vol. 17, no. 12, pp. 2304–2307, Dec. 2013.
11. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," SIAM Journal on Computing,vol. 32, no. 1, pp. 48–77, Nov. 2002.
12. P. Sastry, V. Phansalkar, and M. Thathachar, "Decentralized learning of nash equilibria in multi-person stochastic games with incomplete information," IEEE Trans. Syst., Man, Cybernm., vol. 24, no. 5, pp. 769–777, May 1994.
13. V. Erceg, L. J. Greenstein, S. Y. Tjandra, et al., "An empirically based path loss model for wireless channels in suburban environments," IEEE J. Sel. Areas Commun., vol. 17, no. 7, pp. 1205–1211, Jul. 1999.