A Markovian Game-Theoretical Power Control Approach In Cognitive Radio Networks: A Multi-Agent Learning Perspective

Jiandong Li, *Senior Member, IEEE,* Chungang Yang, *Student Member, IEEE*
(State Key Laboratory of ISN, Xidian University, Xi'an ShaanXi, 710071 China)

Abstract: Due to the lack of learning and adaptation abilities in traditional game models, e.g., the strategic Nash game, they can't describe the dynamic behaviors and strategy interactive selections well in the cognitive context. The Markov game theoretical modeling approach is investigated to deal with the power control in the cognitive radio (CR) context, which well captures the learning and adaptation abilities of CRs. With the complex interaction relationship of multiple secondary users (SUs), multiple primary users (PUs) with wireless coexistent environment into consideration, both the secondary overall utility maximization and fairness among the SUs are considered from the mathematical model formulation and the algorithm design perspective. A power control approach to searching for the fair and optimal Nash equilibrium solution (NES) based on the improved multi-agent Q-learning is proposed. Meanwhile, the parameters of the presented algorithm are analyzed through simulations. The numerical results confirm that the proposed algorithm can improve the system utility and also guarantee the fairness among the SUs well.

Index Terms: Markov game; Multi-agent system; resource management; power control; cognitive radio network.

I. INTRODUCTION

The cognitive radio (CR) technology has been developing for a decade since Mitola first coined in 1999 [1]. Many extensions of cognitive concept can be found in the corresponding document of FCC and the newly build protocol of IEEE, e.g. IEEE 802.22 [2]. During the last decade, it has been always a research hotspot of the industry and research institutions over the world [3]. The next generation (xG) networks are expected to be dynamic spectrum access networks with the aid of CR technology [5], which can largely improve radio spectrum efficiency and mitigate physical reason of spectrum scarcity. In the famous cognition cycle concept [1], cognitive intelligence abilities (e.g., learning and adaptation) are both critical for composing a CR network, which are also the most distinctive parts compared to the software defined radio (SDR). Cognitive transceivers should be designed with two crucial features of learning and adapting in mind, which have only received limited attention from the CR community from the resource management perspective.

How to utilize the spectrum opportunity after successfully sensing spectrum holes is actually transmission strategy selection problem. It has always been an important element as a very difficult problem to deal with in the CR networks. Secondary users (SUs) with selfish characteristics always pursue the utility function maximization by adaptively and rationally choosing the best strategies, e.g. the transmission power level. They always have been modeled as a non-cooperative approach [4][6]. Further more, the SUs will be randomly distributed in the wireless environment, these will lead to a distributed algorithm. So the game theory is widely utilized as the powerful toolset to solve many NP-hardness issues in the CR network, especially, the power control problem [7-9, and reference therein]. Neel etc. presented a novel cognition cycle in a utility based method, and the power control approach for the SUs based on the non-cooperative game theory was introduced in [6]. To the best of our knowledge, the most important step of the original cognitive cycle [1], the learning step is ignored, that is to say, the formulated game model for the problem is one-shot game, and the Nash equilibrium solution (NES) of the power control game is just an equilibrium solution, which can not guarantee the optimality [7-9]. There are many ways in which the efficiency of the NE can be improved, e.g. the pricing approach [6, 9]. The most emerging method is the learning technique based on one use. When he gets limit information about the entire network environment, they can obtain better utilities according to the local information, e.g. their own strategies and the achieved utilities. We will focus on multi-agent system to redesign the cognition cycle with the complex interaction relationship of multiple secondary users (SUs), multiple primary users (PUs) and wireless coexistent environment into consideration.

In this paper we observe the intelligence and the interaction of the agents in the multi-agent system (MAS). Then we rebuild the cognition process by employing the learning technique in the MAS to further improve the performance of the system. Spectrum sharing technology is generally divided into two categories which are the overlay and the underlay [6], and the research of the underlay appears more valuable and more flexible. Interference avoidance issue is one of the most important problems in the underlay scenario which should be paid more concerned about and deserved the focus of consideration. Power control technology is considered as an important solution to the problem. On the one hand, it can employ the regulation of the cognitive users' adaptively changing transmitting power level to eliminate the interference

This research was supported in part by the National Science Fund for Distinguished Young Scholars under Grant 60725105, by the National Basic Research Program of China under Grant 2009CB320404, by the Program for Changjiang Scholars and Innovative Research Team in University, by the 863 projects under Grant 2007AA01Z288, by the National Nature Science Foundation of China under Grant 60572146, 60902032, 60902033, by the Research Fund for the Doctoral Program of Higher Education under Grant 20050701007, by the Teaching and Research Award Program for Outstanding Young Teachers in Higher Education Institutions of Ministry of Education, China, by the Key Project of Chinese Ministry of Education under Grant 107103, by ISN02080001, 2009ZX03007-004, and by the 111 Project under Grant B08038.

to the primary users. On the other hand, it can maximize the cognitive users' QoS requirements or the system throughput depicted as the network utility [7-9]. Most of aforementioned literatures assumed the fixed channel status, namely, the channel state information that the secondary users can obtain remain unchanged during the power adjustment process, which is unreasonable from the practical perspective of view. Further more, for the opportunistic access network established with the emerging cognitive radio technique, the scenario rarely emerges in the future xG network.

Therefore, we attempt to investigate a power control approach which can track the variability of the channel state and the network performance, and this is of great meaning for the practical wireless communication system. We also attempt to find out the robust power control scheme that can adapt to the complex situation, e.g. the spectrum sensing error, channel fluctuation, signal-to-interference-plus-noise ratio (SINR) measurement error and the user dynamics. The contributions of this paper are summarized as follows.

1. A novel cognition circle based on the multi-agent system [11] is proposed, which considers both the interaction among the multi cognitive users and the interaction between the cognitive users and the wireless environment.
2. With the help of the typical Q-learning technique, the power control problem is modeled as the Markov game process [10] and it is investigated in the novel cognition circle framework.
3. The convergence of the proposed power control algorithm is simulated for two states and multi states scenario, and the numerical results tell us that the algorithm is robust.

II. System Model and Agent-based Cognition Cycle

Traditional mathematical tools encounter many unprecedented problems when analyzing modern wireless communications, especially for the cognitive radio network, so a lot of scholars try to adopt utility theory and game theory in economics to study the issues in the modern system [11]. For example, Neel designed a new cognition cycle based on the concept of utility function from the economics field, but he also pointed out that "the effectiveness of cognitive loop" failed to formulate the nature of the learning process [1].

Fig. 1 depicts the proposed cognition cycle based on the multi-agent system, each SU cyclically detect the radio frequency band for the spectrum holes which they can observe to dynamically access in by the Spectrum Sensing①. When they achieve the opportunity, and they also can obtain the channel state information, the minimum interference power requirement of the primary user, and they will extract information, storage information and analyze the information during the Spectrum Analysis② step. Then they will predict the capacity, or the utility function according to the knowledge they earned by the sensing technology①. At this point, they receive the expected utility, as well as a lot of information about the wireless environment. And, they put all information to the Information Fusion Center③, where multi-agent with the reinforcement learning ability share with each other the knowledge and intend to interact with each other by Learning Engine④ so as to formulate the best transmit strategy. Finally the Policy Engine⑤ module sets down the optimal policy with the help of the center manager④, ⑤ and the information fusion center③. The final Spectrum Decision⑥, e.g. the power level, the coding type and the other transmission will act on the radio environment.

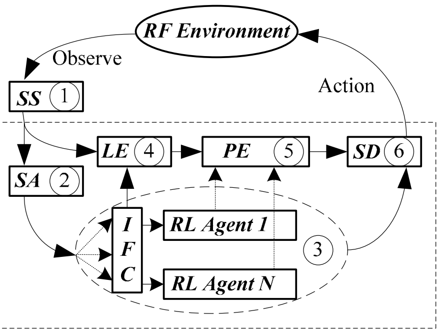

Fig. 1 Multi-agent system (multi-agent learning) based cognition cycle

This process will be ongoing until the best matches between the cognitive users and the wireless environments are achieved for all the users who take possession of spectrum access opportunity. Naturally, during the entire cognitive process, every user constitutes a part of the wireless environment, the optimal strategy selection, e.g. the power level is interactively influence each other. We will utilize the MDP model to describe the interaction process of the cognitive users and the environment, and the Q-learning works well for the multi-user strategy selection process.

III. Markov game model for power control

Markov game is recognized as the extensive form in the Markov decision process (MDP)-like environment and in the MDP formalization of reinforcement learning, in which a single adaptive agent (in this paper, a cognitive user is termed as a agent) interacts with the wireless environment. In the practical cognitive scenario, the framework of Markov game allows us to widen this view to include multi adaptive agents even with interacting competing goal. Reinforcement learning technique is a promising technique for dealing with the coexistence of multi agents, but the mathematical framework that justifies it is inappropriate for multi-agent environment.

Markov game model with more than one player, which is a specific definition as follows:

Definition 1: Markov Game Model

A Markov game G is defined composed by five components: $G = \langle N, S, P, T, U \rangle$, where $N = \{1, 2, ..., |N|\}$ is the player set, and in this paper we alternatively use the player and agent concept. $S = \{1, 2, ..., |S|\}$ represents a finite set of the network states, and each state $k \in S$ contains lots of information, such as the CSI, the PUs' interference power restriction . In this paper, we quantify the information to a single vector so that easy to describe. The transition probability among discrete network states is depicted as T , and $T = \{\pi_{S_k, S_{k+1}}, k = 1, 2, ..., |S|\}$ where $\pi_{S_k, S_{k+1}}$ subject to Markov process, which is one reason why we call the model as the Markov game G .(Though, the most essential reason is Dynamic Character). The policy for each user is termed as

$P = P_1 \times P_2 \times ... \times P_N$, which is the Cartesian product of each user's policy $P_i = \{p_{i,1}, p_{i,2}, ...\}$. U is the utility function, which is designed in detail in the next section.

From the economic perspective, the utility function of customers must spend to obtain the satisfaction of the description. Utility function is similar to the user's QoS requirements in the field of wireless communication and the objective function approximation in mathematical programming. In recent years, research based on utility theory has attracted wide attention, but most of the utility functions are described in the proceeds of a strategy selection that is a short-sighted behavior. In this paper, we consider that each agent is concerned about not only the effectiveness of the current behavior, but also how the current behavior impact to the utility obtained in the future more, as well as future access opportunity. Therefore, we define the discounted utility function here.

Definition 2: The discounted utility function:

With regard to the power control problem, any of the rational agents presumes the best performance, for example, the capacity maximization or the utility maximization, at the same they all also expect to consume less power. From the point of view of economic, the power-consuming is regarded as the price that each cognitive user should pay for, and the utility is termed as the degree of the satisfaction that the consumer can feel. Therefore, in order to find the tradeoff between power efficiency and long-term gains, the definition of utility function is designated as,

$$C_i(p_i(t), SINR_i(t)) = \log(1 + SINR_i(t)) / p_i(t) \tag{1}$$

The unit of the utility function is bit/s/Hz/Joule, which can be thought as the ratio of the spectrum efficiency to the power-consuming level. The term $SINR_i(t)$ represents the Signal-to-interference-plus-noise ratio (SINR) when the agent chooses the power level $p_i(t)$. In essence, the Markov game is a repeated game and a most simply dynamic game model, and we should take into account of both current rewards and the decision-making in the next step or even more long-term earnings, so there is necessary to define the expected utility function as below:

$$\begin{aligned} U_i(t) &= C_i(t) + \rho C_i(t+1) + \rho^2 C_i(t+2) + ... \\ &= C_i(t) + \rho\{C_i(t+1) + \rho C_i(t+2) + ...\} \\ &= C_i(t) + \rho U_i(t+1) \end{aligned} \tag{2}$$

Here, $U_i(t+1) = C_i(t+1) + \rho C_i(t+2) + ...$ and the term ρ is the discount factor, and the $U_i(t)$ depicts the discounted sum of the immediate reward $C_i(t)$, $C_i(t+1)$,…at time t, $t+1$,….In the Markov game, the discounted factor is of great importance, and it is used to test the stability of the solution, which is called the Nash Equilibrium Solution (NES). The discounted factor provides a random stop-point for the repeated game model, which can be interpreted when the probability of the game process keeps going on.

IV. Q-learning-based Power Control Algorithm

In the proposed cognition cycle framework based on multi-agent system, we study a power control approach based on the Q-learning algorithm to adaptively track the changing status of the wireless environment. We formulate the power control problem as the optimal decision-making process of a Markov Game Model, and with the combination of dynamic programming of the Bellman equation, we have the following analysis.

In the last section, we have defined the discounted utility function, so the expected utility function of the *ist* SU at the s state is,

$$\begin{aligned} V_i(s) &= E[U_i(t) \mid s_t = s] \\ &= E[C_i(t) + \rho U_i(t+1) \mid s_t = s] \\ &= \sum_{p_i \in A_i} p(s, p_i) \sum_{s'} \pi_{s,s'}^{p_i} (U_{s,s'}^{p_i} + \rho V_i(s')) \\ &= \sum_{p_i \in A_i} p(s, p_i) \sum_{s'} [\pi_{s,s'}^{p_i} U_{s,s'}^{p_i} + \rho \pi_{s,s'}^{p_i} V_i(s')] \end{aligned} \tag{3}$$

Here, the term $\pi_{s,s'}^{p_i}$ is the transition probability of the *ist* SU from the state s to the next state s' with the policy of p_i. $p(s, p_i)$ is the probability that the *ist* SU use the transmit strategy p_i. $V_i(s') = E[U_i(t+1) \mid s_{t+1} = s]$ is the expected utility function that the *ist* SU can obtain at the next state.

In accordance with the Bellman equation, we obtain the definition of Q function,

$$Q(s, p_i) = \sum_{s'} [\pi_{s,s'}^{p_i} U_{s,s'}^{p_i} + \rho \pi_{s,s'}^{p_i} V_i(s')] \tag{4}$$

Then the Q-learning technique which is employed to the power control approach can be mathematically described as,

$$\begin{aligned} Q(s, p_i) &= (1-\alpha) Q(s, p_i) + \alpha [C(s, p_i) + \rho V_i(s')] \\ V_i(s) &= \max_{p_i \in A_i} \{Q(s, p_i)\} \end{aligned} \tag{5}$$

The proposed algorithm for the practical implementation can be summarized as:

1. According to the sensing information ***(SS① and SA②)*** to evaluate the channel gain accordingly, then to initialize a transmission power $p_{i,initial}, i = 1,..,N$ for each SU, referring to the status of the interference power situation of the radio environment ***(SA②)***;
2. Calculate the accumulated rewards $V_i(s), i = 1,..,N$, here taken time-varying channel into account, that is, each user can detect the respective channel status, and predict the one-step rewards with the help of Eq. (8);
3. Execute the Q-learning shown as Eq. (5) based power control algorithm to learn the optimal transmit policy ***(PE⑤)***, and further prediction of Q-values by the ***LE④***;
4. Until the algorithm is convergent, then the optimal transmission policy is achieved and execute on the radio environment ***(SD⑥)***.

V. Simulation and Analysis

For simplicity, we assume that the network state is a function of the channel gain, and this assumption is reasonable because the network status's variation is mainly caused by the CSI in practice. In accordance with the channel gain, the channel model is divided into $|S|$ status, with a fixed interval. Meanwhile, the channel model is subject to exponential distribution as Eq. (6) shown. The state probability π_k is the integral of the distribution function at the fixed interval, and the transition probability $\pi_{S_k, S_{k+1}}$ can be calculated according to the Markov nature of the various status of status.

$$p(g) = 1/g_0 \times \exp(g/g_0) \tag{6}$$

Here g_0 is the average the channel gain. We set the $\Gamma = \{\Gamma_0, \Gamma_1, ..., \Gamma_{|S|}\}$ as the thresholds of the $|S|$ network status, and here $\Gamma_0 = 0$ and $\Gamma_{|S|} = \infty$. If the channel gain stands between the interval $\Gamma_k \le g < \Gamma_{k+1}$, we say that the SU with the channel gain g is at the *kst* networks state S_k. Throughout the simulation section, we assume that the state probability π_k of each S_k is fixed equaling to $\pi_k = 1/|S|$. Because of $\Gamma_0 = 0$, we can conclude that the state probability π_k at $(-\infty, \Gamma_0]$ is the least chance, and we assume that the probability equals to a very small value ξ:

$$p\{g < g_0\} = \int_{-\infty}^{\Gamma_0} p(g)dg = \xi \tag{7}$$

We can get the $\Gamma_0 = f(p(g), \xi)$, and then others threshold $\Gamma_k, k = 1, ..., |S| - 1$. The transition probability $\pi_{S_k, S_{k+1}}$ is well calculated according to,

$$\pi_{S_k, S_{k+1}} = \int_{\Gamma_k}^{\Gamma_{k+1}} p(g)dg \tag{8}$$

A. Analysis of the proposed algorithm (Two SU and Two States)

In this subsection, we first investigate the proposed power control approach based on the Markov game model $G = \langle N = 2, S = 2, P \in R, T, U \rangle$. Both the transition probability and the utility function can be obtained in the method that is introduced in the last sections. One of the most important issues for the power control algorithm in the game-theoretic framework is the convergence and the robustness in line with the variable channel.

Figure 2 depicts how the parameters, the discounted factor ρ and the learning rate α, affect the rate of the convergence and the final performance of the secondary system. The X-coordinate indicates the iteration number, and the Y-coordinate is the expected utility. With the different parameter settings, the performance of the system and the convergence rate vary greatly. We can find that the adaptive transmission power control algorithm employing Q-learning method can well guarantee the convergence.

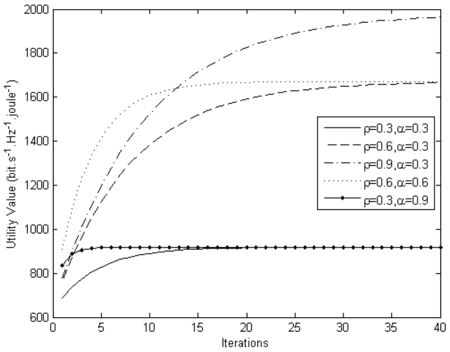

Fig. 2 Performance of the Q-learning algorithm

First, we explore the scenario where the learning-rate is a fixed value, here, $\alpha = 0.3$. Then we observe the three lines with the discounted factor $\rho = 0.3$, $\rho = 0.6$ and $\rho = 0.9$ respectively. We can conclude that: when studying the case with a fixed learning rate， the convergence rate of the proposed algorithm becomes slow as the discount factor increases, that is， the slower the convergence rate is, the greater the ρ becomes.

Remark 1: Combined with the definition of the discounted factor ρ, and its physical meaning is the larger the discount factor is, the longer the Q-learning time. That is to say the Q-learning based power control approach will consume more time to converge to Nash equilibrium solution. That is, the longer the learning time accordingly, the slower the convergence becomes.

However, from Figure 2, we also notice that as the discount factor increases, the expected utility of each player is effectively improved because they can obtain a larger number of interactive and learning strategies to pursue their long-term utility optimal. They use up all of their improvement opportunities until they can not get more utility. Then, we explore another scenario with fixed discounted factor, here, $\rho = 0.6$ .We observe the Fig.2with the learning rate $\alpha = 0.3$ and $\alpha = 0.6$ respectively. We can conclude that: for the fixed discounted factor scenario, the larger the learning rate is, the faster the convergence rate of the proposed algorithm is. Meanwhile, we can find that as long as the discount factor remains the same, the algorithm will converge to the same optimal utility NES though with different convergence rates. The conclusion is verified again when we consider the case of the fixed value $\rho = 0.3$ and the learning rate $\alpha = 0.3$ and $\alpha = 0.9$, respectively.

Remark 2: Therefore, in practical system design process, we take into account the performance of the system requirements, as well as convergence speed, to flexibly select system parameters.

B. Performance Analysis

In Figure 4, for the case with multi-user and multi-status, the algorithm still has a good convergence. No matter how bad a state the SUs start with, they will obtain a total fair approximation to the effectiveness of implementation. This will be further described in Figure 5.

In the last subsection, we consider the scenario that possesses multi SUs and multi states, and the robustness and the convergence property are described as the Fig. 4.

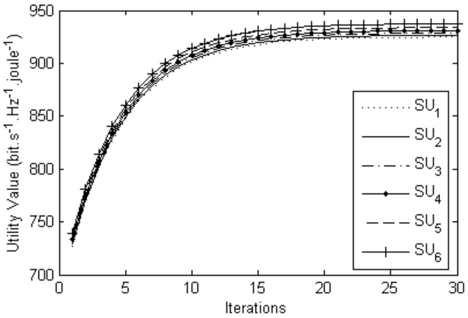

Fig. 4 the robustness and the convergence property

Figure 5 is the multiple SUs case, and it shows us that the optimal utility obtained by each SU after the power NES is achieved. The proposed Q-learning algorithm can effectively improve the performance and meanwhile we can find that comparing to the no Q-learning one, it is essentially a one-shot algorithm. The algorithm also guarantees well the fairness among the multiple SUs. For SU5 and SU6, they achieve more utility but destroy the fairness as a result of selfish behavior. Then they get certain punishment, and all cognitive users expect to achieve a fair approximation of the proceeds eventually. Therefore, this proposed algorithm can guarantee degree fairness among SUs without damaging the more utility maximization. In game theory, the learning algorithm is an implementation of the non-cooperative game approach to cooperation, so that the comparison between the two algorithms is concerned in Figure 5.

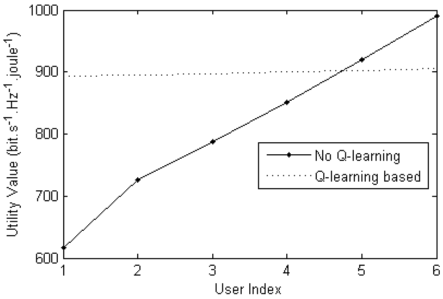

Fig. 5 Comparison of the effectiveness of algorithm

VI. Conclusion

A novel cognition cycle based on the multi-agent system is proposed, which takes into account of the interaction among the rational and individual agents as well as the interaction between the various agents with wireless environment. From the Markovian game-theoretic perspective, the spectrum sharing issue among multiple secondary users is formulated as a expected utility maximization problem. To observe the historic information, each SU can well obtain the fair and optimal power control strategy. Employing the multi-agent Q-learning technique, an adaptive transmission power control approach is proposed, which is distributed and with low computation complexity. Simulation results show that the presented algorithm can effectively guarantee the fairness among users. Compared to a static game, Nash equilibrium solution of the power control is also proved more effective. The proposed algorithm can always guarantee good convergence and the system robustness.

References

[1] J. Mitola, Cognitive radio: An integrated agent architecture for software defined radio. Doctor of Technology [c]. Royal Institute of Technology. Stockholm. Sweden. 2000.

[2] S. Haykin, Cognitive Radio: Brain-Empowered Wireless Communications [J]. IEEE JSAC, 2005, 23(2):201-220.

[3] A. N. Mody, S. R. Blatt, Recent Advances in Cognitive Communications [J]. IEEE Communications Magazine, October, 2007, 45(10):54-61

[4] J. Zhu, K J Ray Liu, Cognitive radios for dynamic spectrum access-Dynamic spectrum sharing: A game theoretical overview [J]. IEEE Communications Magazine, 2007 (5), 45(5):88-94.

[5] I. F. Akyildiz, W. Y. Lee, M.C. Vuran, S. Mohanty, NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey [J]. Computer Networks, 2006(9),50 (13):2127-2159.

[6] J. Neel, Game theory can be used to analyze cognitive radio Source: Electronic Engineering Times, Aug 29, 2005, 1386:69 -72.

[7] C. G. Yang, J. D. Li, A Game-Theoretic Approach to Adaptive Utility-Based Power Control in Cognitive Radio Networks, VTC2009-Fall:1–6.

[8] C. G. Yang, J. D. Li, and Z. Tian, Optimal Power Control for Cognitive Radio Networks with Coupled Interference Constraints: A Cooperative Game-Theoretic Perspective, IEEE Transactions on Vehicular Technology, May 2010, 59(4): 1696-1706.

[9] C. G. Yang, J. D. Li, et al, Game Theory based Power Allocation in Cognitive Radio, Journal of Xidian University, Feb. 2009, 36(1): 1-4.

[10] J. Huang, V. Krishnamurthy, Transmission control in cognitive radio as a Markovian dynamic game: structural result on randomized threshold policies, IEEE Transactions on Communications, 2010, 58(1): 301 – 310.

[11] Kaelbling, Leslie Pack. Reinforcement learning: a survey. Source: Journal of Artificial Intelligence Research, 1996, 4: 237-285.