Power Control Based on Deep Q Network with Modified Reward Function in Cognitive Networks

Fang Ye, Yinjie Zhang, Yibing Li, Tao Jiang, Yingsong Li
Department of Information and Communication
Harbin Engineering University
Harbin, China
yefang0923@126.com, zhangyinjie1128@126.com, liyibing0920@126.com, jiangtao@hrbeu.edu.cn, liyingsong@hrbeu.edu.cn

***Abstract*—This paper aims to design an appropriate power control policy of the secondary user (SU) to share the spectrum with the primary user without harmful interference. With dynamic spectrum environment, we develop a power control policy based on deep reinforcement learning with Deep Q network (DQN) that the secondary can intelligently adjust his transmit power. And reward function is properly designed to avoid the sparse reward problem which may cause the secondary user cannot adjust to effective power in limited steps and finally fails to transmit. Our experiment result reveals that under the help of the proposed network and reward function, the secondary user can fast and efficiently adjust to effective power from any initial state.**

***Index Terms*—Cognitive radio, spectrum sharing, power control policy, deep reinforcement learning, sparse reward**

I. Introduction

Cognitive radio (CR) put forward by Joseph Mitola is considered as an effective method to improve the spectrum utilization. And the concept of spectrum sharing through CR is highly motivated [1]. During spectrum sharing, the CR user as the secondary user is allowed to use the licensed spectrum without harmful interference.

Most of existing researches consider the power control problem from the perspective of optimization [2] or game theory [3][4]. However, the above methods require some prior information, such as channel state or primary user power. Hence, the method of this paper is based on deep reinforcement learning (DRL) that helps agents to learn from the interaction with environment. Sparse reward problem is often encountered in the reinforcement learning, which reflects in power control that the SU cannot adjust to effective power in limited steps and eventually fails in data transmission. The object of the paper is to make the SU intelligently adjust its transmit power without harmful interference while ensuring quality of service (QoS) of both the primary user (PU) and the SU. Whats more, the reward function is properly designed to avoid sparse reward problem.

The rest of paper is organized as follows. Section Two provides the description of system model. In Section Three, a learning algorithm based on DQN is developed. The experiment results and analysis are conducted in Section Four.

II. System Model

Consider a primary user and a secondary user share the same spectrum in Additive White Gaussian Noise channel. We assume that both the PU and the SU can successfully transmit their data with guaranteeing their QoS, only if the corresponding SINR is over the respective required threshold η_1 and η_2. All possibilities of PU power are regulated to a set P_1 with L_1 elements. Moreover, there are *N* sensor nodes to spatially collect the received signal strength information[5]. $P_n^r(k)$ represents the received power of *n*th node at *k*th time frame,which consists of the transmit power of the PU and SU and a Gaussian random variable, so the state $\mathbf{s(k)}$ is a received power vector of N nodes.

The action of SU is defined as the transmit power it chooses from a finite set P_2 with L_2 elements.In order to avoid sparse reward, we appropriately designed the reward function to make the SU get feedback at every step. The reward function is defined as

$$r(k) = \begin{cases} a \text{ if } \mathrm{SINR}_1 \geq \eta_1 \text{ and } \mathrm{SINR}_2 \geq \eta_2 \\ b \text{ if } \mathrm{SINR}_1 < \eta_1 \text{ and } \mathrm{SINR}_2 \geq \eta_2 \\ c \text{ if } \mathrm{SINR}_1 \geq \eta_1 \text{ and } \mathrm{SINR}_2 < \eta_2 \\ d \text{ if } \mathrm{SINR}_1 < \eta_1 \text{ and } \mathrm{SINR}_2 < \eta_2 \\ b \text{ if steps} > T \end{cases} \tag{1}$$

where the parameter *a* is positive, which means both successfully transmit data; the parameter *b* is negative, which means the SU has affected the PU and should be punished; the parameter *c* is a relatively enough small positive number to protect the PU activity; in the same way, the parameter *d* is a relatively enough small negative number. Lastly, when the SU did not find the proper transmit power within a certain steps *T*, it also should be punished, and the reward is equal to the second condition *b*.

III. DEEP REINFORCEMENT LEARNING BASED ON DEEP Q NETWORK FOR POWER CONTROL

The paper aims to make the SU intelligently adjust its transmitter power within limited steps from any initial state. Abstractly in reinforcement learning, the agent learns the optimal policy π^* for decision-making. The Deep Q Network is proposed by Google Deepmind [6]. The table in conventional Q-learning is replaced by neural network to approximate the

978-1-7281-6670-4/20/$31.00 ©2020 IEEE

state-action function. The training data of DQN is generated from the interaction with environment. In state $\mathbf{s(k)}$, the SU select the action $a(k)$ of the highest Q-value with probability ε_k or a random action in case there is greater state which has not been explored, and get a reward $r(k)$, then reach the next state $\mathbf{s(k+1)}$. The above process consists of a transition $d(k) = \{\mathbf{s(k)}, a(k), r(k), \mathbf{s(k+1)}\}$ stored into Replay Buffer D or memory. And a minibatch N_{batch} is uniformly sampled from Replay Buffer to train the neural network. Motivated by the paper [6], we builds a target network which has the same structure with the former network to get robust model. Therefore, the objective function of DQN is the to minimize the square error of two networks.

$$Loss(\theta) = \min_{\theta} \sum_{k \in N_{batch}} (Q_{target} - Q(\mathbf{s(k)}, a(k); \theta))^2 \quad (2)$$

where θ means the biases and weights in the DQN. When the proposed power control policy reaches the terminal, both the PU and the SU step into a data transmission period, so the state will stay at the terminal until the transmission finishes. And the SU finally learns to adjust its power such that the next state remains the terminal.

IV. Experiment Results

In our experiment, the simulation parameters are set as following. The sets of PUs power and SUs power (in Watt) are respectively assumed as $P_1 = \{4.0, 4.5, \cdots, 8.0\}$ and $P_2 = \{1.0, 1.5, \cdots, 6.0\}$. The Neural network in DQN has two fully-connected hidden layers and respective activation functions are ReLU and tanh. The episodes of reinforcement learning are 1000 with T=35 steps at most for every episode. It is noted that the above structure and parameters can be modified depending on real environment.

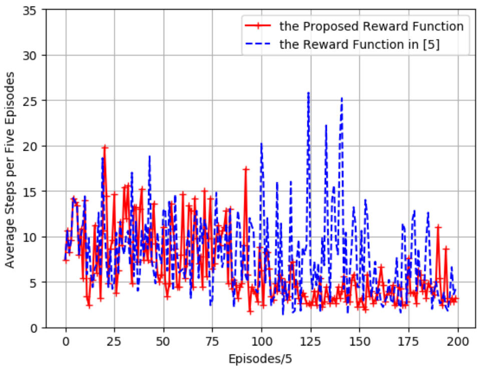

Fig. 1. Average Steps per Five Episodes

We compare the proposed reward function with the reward function in [5] which might cause sparse reward problem. From the Figure 1, we can see that before 500 episodes, the two curves fluctuate with a large variance, since the SU is still at exploration stage and do not learned good policy. After 500 episodes, the SU with the proposed reward function exploits the learned policy, and the average steps obviously decreases, finally converges 3-4 steps with a little fluctuation. The reason why there is still fluctuation is the exploitation probability is 0.9, so the SU explores the environment with 0.1 probability to prevent the better state from not being explored. The exploitation probability can be adjusted to 1.0 according to practical needs. On the other hand, the SU with reward function in [5] still fluctuates a lot at exploitation part, due to sparse reward problem, that is, the SU did not get proper feedback from reward function and did not learn good policy.

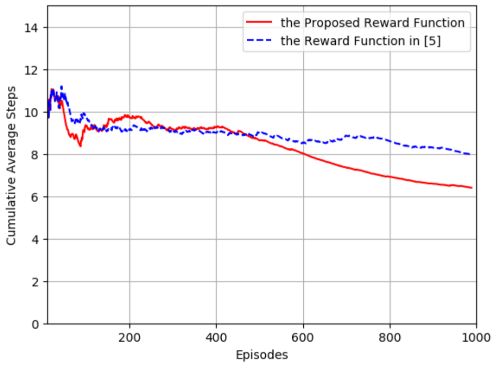

Fig. 2. Cumulative Average Steps

Values of cumulative average steps before 10 episodes are removed from Figure 2 due to random exploration. We can see from Figure 2 that two curves gradually decrease with episodes, which reveals SUs can learn policy from both reward function based on DQN. The curve with the proposed reward function clearly decreases faster and reaches smaller steps, while the curve with reward function in [5] seldom decreases after 200 episodes, because its steps fluctuates with a large variance which can be known from Figure 1.

Acknowledgment

This paper is funded by the National Natural Science Foundation of China (Grant No. 61701134) and the International Exchange Program of Harbin Engineering University for Innovation-oriented Talents Cultivation.

References

[1] S. Haykin, Cognitive radio: Brain-empowered wireless communications, IEEE J. Sel. Areas Commun., vol. 23, no. 2, pp. 201220, Feb. 2005

[2] Das, Gopal Chandra , et al. "Power control in underlay cooperative cognitive radio network under shadowing." International Conference on Microelectronics IEEE, 2016.

[3] G. Yang, B. Li, X. Tan, and X. Wang, Adaptive power control algorithm in cognitive radio based on game theory, IET Commun., vol. 9, no. 15, pp.18071811, Oct. 2015

[4] Junhui, Zhao , et al. "Power Control Algorithm of Cognitive Radio Based on Non-Cooperative Game Theory." China Communications 10.11(2013):143-154.

[5] X. Li, J. Fang, W. Cheng, H. Duan, Z. Chen and H. Li, "Intelligent Power Control for Spectrum Sharing in Cognitive Radios: A Deep Reinforcement Learning Approach," in IEEE Access, vol. 6, pp. 25463-25473, 2018.

[6] V. Mnih et al., Human-level control through deep reinforcement learning, Nature, vol. 518, pp. 529533, 2015.