Multi-Agent Reinforcement Learning for Dynamic Spectrum Access

Huijuan Jiang*, Tianyu Wang*†, and Shaowei Wang*
*School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
†National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
Email: MG1723063@smail.nju.edu.cn, {tianyu.alex.wang, wangsw}@nju.edu.cn

***Abstract*—Cognitive radio is an efficient spectrum sharing mechanism to solve the contradiction between spectrum shortage and spectrum underutility, where secondary users (SUs) are allowed to access the spectrum licensed to primary users (PUs) in an opportunistic manner. In cognitive radio networks with multiple access points (APs), due to the information exchange cost and system flexibility, APs may not cooperate with each other and there usually does not exist a central controller in practice. We propose a distributed user association scheme based on multi-agent reinforcement learning to achieve load balancing for cognitive radio networks with multiple independent APs. In our proposed scheme, APs execute reinforcement learning process independently to derive optimal policies on user association. In each iteration, APs make decisions on choosing SUs for association and then SUs choose the optimal AP for association based on the offers of all APs, the behaviors of APs and SUs is modeled as a dynamic matching game. Simulation results show that the proposed multi-agent reinforcement learning approach can highly improve the system performance with excellent robustness, compared to the conventional max-SINR method.**

I. Introduction

With the development of information and communication technology, the wireless traffic has experienced an explosive growth in the recent years and the growth rate will accelerate in 5G era. As reported in [1], global traffic will have increased 127-fold from 2005 to 2021 and the traffic is unevenly distributed in time and space. However, due to the traditional spectrum management policy, the spectrum resource is far from full utilization. To solve the contradiction between spectrum shortage and inefficient utilization, cognitive radio (CR) is proposed as an efficient dynamic spectrum sharing mechanism. CR system allows secondary users (SUs) sense and opportunistically access the spectrum licensed to primary users (PUs) as long as the interference to PUs is tolerable [2], [3]. The flexible wireless environment and the requirement of opportunistic access pose a great challenge on resource management in CR networks.

To cope with the immense amount of traffic generated by massive SUs, it is essential for CR systems to deploy multiple access points (APs) in the environment. In a CR system with multiple APs, APs are with less transmission power and hence, small coverage, and the links from APs to SUs becomes shorter, which will improve the transmission quality. However, it raises new challenge on user association, interference management, spectrum sharing and scheduling. User association plays an important role in load balancing, the spectrum efficiency and energy efficiency [4].

Current schemes for user association are classified into central schemes like [5]–[7] and distributed schemes like [8], [9]. Among the centralized mechanisms, the authors in [5] consider a joint optimization problem of use association, subchannel allocation and power allocation in multi-cell multi-association OFDMA heterogeneous networks. In [6], a combinatorial optimization problem for user association and interference coordination in heterogeneous cellular networks is formulated. In [7], user association and user scheduling is jointly optimized for load balancing. Among the centralized mechanisms, the authors in [8] model the competitive behaviors among the user equipments, femtocell APs and macrocell APs as a dynamic matching game and propose distributed algorithms to find the user association and femtocell APs allocation.In [9], online algorithm for the multi-tier multi-cell user association problem is proposed. However, it needs the full network information and a centralized controller. The information exchange is generally expensive.

Consider that APs may not cooperate with each other and there usually does not exist a central controller, central scheme is not practical for user association in CR networks. In most distributed scheme for user association, conventional max-SINR method is not capable for coping with the imbalancing load, some works proposes a matching process between for the users and the serving APs by solving a complex optimization problem [8]. Reinforcement learning algorithm enables an agent to obtain the optimal policy by exploiting the environment, it is essential for APs to learn from the experience and derive an optimal user association policy independently. In multi-agent reinforcement learning algorithm, multiple agents learn optimal policies by exploring the common environment. At each step, each agent chooses an action according the policy that will change the state of environment, and then it will receive a reward indicating the validity of the action and improve the policy. The goal of the agents is to optimize the whole system utility by deriving optimal policies [10], [11]. Multi-agent reinforcement learning is a significant research for distributed problem without central control that the agents

This work was partially supported by the National Natural Science Foundation of China (61801208,61671233), the Jiangsu Science Foundation (BK20170650), the Postdoctoral Science Foundation of China (BX201700118,2017M621712), the Jiangsu Postdoctoral Science Foundation(1701118B), and the open research fund of National Mobile Communications Research Laboratory (2019D02).

978-1-5386-8088-9/19/$31.00 ©2019 IEEE

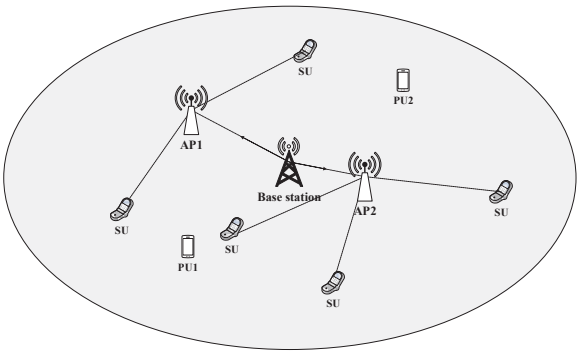

Fig. 1. Illustration of our system model.

learn behavior online. Some researches have studied how reinforcement learning works in CR networks, like [12]–[15]. These works apply reinforcement learning in interference control, power control, spectrum sensing policies and spectrum access. However, in multi-agent reinforcement learning, each agent only has a partial view of the whole system environment and utility, and in addition, all agents make decisions independently and simultaneously, causing that the reward of each agent is influenced by the other agents. It is desirable to design an efficient learning and coordinating mechanism.

In this paper, we propose a multi-agent reinforcement learning for user association in CR networks, where multiple APs independently manage specific spectrum resource in an area with massive SUs. Our system model is shown as Fig. 1. In the system, SUs access spectrum in overlay mechanisms. In each period, APs execute user association and spectrum allocation process. In the user association process, each AP conducts independent reinforcement learning that make a decision on the SUs selection for association based on the spectrum occupation of PUs and the SINR distribution of SUs. Then the SUs choose only one AP with maximal SINR for association. APs and SUs act as a dynamic matching game. In the spectrum allocation process, each AP allocates spectrum to associated SUs with the principle of priority for large SINR users until the spectrum is all exploited.

The rest of the paper is organized as follows. We present system model in Section II. In Section III, we give a brief introduction to multi-agent reinforcement learning and propose the solution for the user association problem. Simulation results are provided in Section IV and we conclude the paper in Section V.

II. System Model

We consider a CR system with M APs with fixed coordinates (x_m, y_m), each AP is responsible for managing a chunk of spectrum and allocating spectrum resource not occupied by PUs to SUs. The licensed spectrum occupation of PUs follows a Possion distribution. We denote L_m as the spectrum managed by $m-th$ AP. The system contains K SUs, they are distributed and move randomly in the area. Different subarea has different SUs distribution density. The packets for transmission arrive in a Possion distribution.

Denote p_m as the power that the $m-th$ AP allocates its associated SUs for transmission and $h_{m,k}$ as the channel gain between the $m-th$ AP and $k-th$ SU. We assume an additive white Gaussian noise (AWGN) at SUs with power σ^2. Then the signal to interference plus noise ratio (SINR) between $m-th$ AP and $k-th$ SU is $\frac{p_m|h_{m,k}|^2}{\sigma^2}$, and the capacity of $m-th$ AP serving k$-th$ SU with unit spectrum is given by:

$$C_{m,k} = \rho_{m,k} log(1 + \frac{p_m|h_{m,k}|^2}{\sigma^2}). \quad (1)$$

Then, the spectrum for serving $k-th$ SU with packet size s_k is $l_{m,k} = s_k/C_{m,k}$. For simplification, we assume that the packets size of all SUs is equal.

In the CR system with multiple APs, the space distribution and spectrum resource of APs would not always match with the space distribution and access requirement of SUs. Consider the communication and collaboration cost, APs can not cooperate with each other, and in addition, there is not a central controller. Each AP need to derive an optimal policy on selecting SUs for association, in order to cope with load imbalance and maximize the total SUs that the CR system can transmit. Each AP should make a decision on user association and resource allocation problem.

In user association process, each AP makes a decision on how many SUs to notice for association according to spectrum occupancy and SUs load condition, and then notice SUs to associate by the order of SINR. SUs who receive access notification select only one AP with maximal SINR to associate with. SUs who don't receive access notification wait for the next time. The association process of APs and SUs is modeled as a dynamic matching game. After SUs association, each AP need to allocate spectrum resource to associated SUs according to SUs' spectrum demand and SINR. In this process, APs allocate spectrum resource to SUs in the descend order of SINR until all spectrum is allocated.

If an AP allows too many SUs to associate with, some SUs could have been transmitted by other AP if the AP didn't inform it to associate with and SUs associate with other AP with light load. If an AP allows not enough SUs to associate with, the spectrum resource is wasted.

Therefore, APs develop policies on choosing optimal SUs to associate, in order to maximize the associated SUs that have been transmitted while minimizing the associated SUs that have not been transmitted. For SUs associated with $m-th$ AP, let $|TRAN|_{m,k} = 1$ denote that the k$-th$ SU has been transmitted by $m-th$ AP that it associates with, $|TRAN|_{m,k} = -1$ denote that the $k-th$ SU has not been transmitted by $m-th$ AP that it associates with. Then the utility function of $m-th$ AP is given by:

$$U_m = \sum_{k=1}^{K} \rho_{m,k}|TRAN|_{m,k}, \quad (2)$$

where $\rho_{m,k}$ is the association indicator, $\rho_{m,k} = 1$ indicates that k$-th$ SU associates with m$-th$ AP and $\rho_{m,k} = 0$ indicates

that $k-th$ SU doesn't associate with $m-th$ AP.

We try to maximize the whole system utility. Mathematically, the optimization problem can be described as follows:

$$\begin{aligned} & \max_{\rho_{m,k}, l_{m,k}} \sum_{m=1}^{M} \sum_{k=1}^{K} \rho_{m,k} |TRAN|_{m,k} \\ s.t. \quad & C_1 : \sum_{m=1}^{M} \rho_{m,k} = 1, k = 1, \ldots, K, \\ & C_2 : \rho_{m,k} \in \{0,1\}, \forall m, \forall k, \\ & C_3 : \sum_{k=1}^{K} \rho_{m,k} l_{m,k} \leq L_m, m = 1, \ldots, M, \end{aligned} \tag{3}$$

where C_1 and C_2 declares that each SU can only associate with one AP. $C3$ is the spectrum resource constraint.

However, each AP only has a partial view on the overall system, and their actions would affect each other's utility. In addition, consider the uneven distribution and movement of SUs, some SUs with better channel conditions between multiple AP might receive access notifications from multiple AP and choose the best one. That way, for each AP, the number of actual SUs associating with it is not equal to the number of SUs who receive access notification. Therefore, the idle spectrum resource, the distribution of SUs, the policies of other APs influence the performance of each AP concurrently. To solve this distributed optimization problem, we introduce multi-agent reinforcement learning algorithm in user association process, which doesn't need any information exchange among APs. And then APs allocate spectrum resource to SUs in the descend order of SINR until all spectrum is allocated in resource allocation process.

III. Proposed Solution

A. Multi-Agent Reinforcement Learning Algorithm

In multi-agent reinforcement learning process, a set of autonomous agents who share common environment aim to obtain optimal policies, in order to maximize the whole system utility without a prior environment model. Each agent only has a partial view on the overall system environment and utility, and all agents make decision autonomously based on the partial system view. The reward of an agent not only depends on its action and local state but also on other agents' states and policies.

The multi-agent reinforcement learning process is denoted as the tuple $\{\mathcal{M}, \mathcal{S}, \mathcal{A}, P(\overrightarrow{s}, \overrightarrow{a}), R(\overrightarrow{s}, \overrightarrow{a})\}$, each of which is described as follows:

- $\mathcal{M} = \{1, \ldots, M\}$: the set of agents, who make decisions based on the environment and try to obtain optimal policies to maximize utility.
- $\mathcal{S}$: the set of possible environment states, yielding joint state vector $\overrightarrow{s} = \{s_1, s_2, \ldots, s_M\}$.
- $\mathcal{A}$: the set of actions available to agents, which yields joint actions set $A_1 \times A_2 \times \ldots \times A_M$ and joint action vector $\overrightarrow{a} = \{a_1, a_2, \ldots, a_M\}$.
- $P(\overrightarrow{s}, \overrightarrow{a})$: the station transition probability function of the system states, given the joint action $\overrightarrow{a}$.

TABLE I
Categorization Criterion of s_{su}

SUs in $[SI_1, +\infty)$	SUs in $[SI_2, SI_1]$	s_{su}
$[0.1, 1]$	–	1
$[0.05, 0.1]$	$[0.25, 1]$	2
$[0.05, 0.1]$	$[0, 0.25]$	3
$[0.01, 0.05]$	$[0.25, 1]$	4
$[0.01, 0.05]$	$[0, 0.25]$	5
$[0, 0.01]$	–	6

- $R(\overrightarrow{s}, \overrightarrow{a})$: the reward function of the agents given the joint action $\overrightarrow{a}$ perform at joint state $\overrightarrow{s}$.

However, each agent only has a partial view on local state and local action, which means that only local reward $R_m(s, a)$ is available to $m-th$ agent.

In the problem, each agent aim to obtain an optimal policy Π_{opt} that maximize the expected cumulative reward during the learning period:

$$\max_{\Pi} E[\sum_{t=1}^{T} \gamma R_m^t(s, a)], \tag{4}$$

where $\gamma, 0 \leq \gamma \leq 1$ is the discount rate, indicting the effect degree of future reward to the present decision.

B. Learning-Based User Association

In the user association problem, the APs can't obtain the information about other APs' space distribution, policy and spectrum resource. Each AP should learn an optimal policy on SUs association in order to maximize the whole system utility, based on its own spectrum resource and the SINR matrix between the AP and SUs.

We define the agent, the state space, the action space and the reward function for the multi-agent reinforcement learning process as follows:

- Agent: the decision maker. In this CR system, the agent represents the APs who are executing the SUs association process.
- State: $s =< s_{pu}, s_{su} >$ consists of the occupation feature of PUs and the SINR feature of SUs who are with transmission requirement. The occupation feature of PUs represents the spectrum resource not occupied by PUs. The SINR feature of SUs represents the overall SINR condition of SUs with transmission requirement. For simplification, we categorize the SINR feature into six levels according to the percentage of SUs with SINR in the range $[SI_1, +\infty)$ and $[SI_2, SI_1]$. The categorization criterion follows as as table I.
- Action: a_m is the action taken by $m-th$ agent, it denotes the number of SUs that the agent allow to associate with.
- Reward: In our problem, the packets of associated SUs may be transmitted or un-transmitted by the AP. We define the reward as $R(s, a) = \sum_{k=1}^{K} \rho_{m,k} |TRAN|_{m,k}$.

For each agent, we construct a Q-table with dimension $R \times C$, where R is possible states and C is possible actions for each agent. The entry $Q_m(s_r, a_c)$ of this Q-table is Q-values

TABLE II
MULTI-AGENT REINFORCEMENT LEARNING PROCESS

1.**Initialization**:
2.$Q_m(s,t)=0$ for $m \in \mathcal{M}$, and $t=1$.
3.***While*** $t \leq T$
4. ***For agent*** $m \in \mathcal{M}$***:***
5. Agent m observe environment state s_m^t at time t.
6. Based on state s_m^t, the agent choose actions a_m^t according to the action choosing policy.
7. ***End for***
8. The joint action generate a transition to new state $\overrightarrow{s}^{t+1}$
9. and an immediate reward $r_m^t(\overrightarrow{s}, \overrightarrow{a})$ for $m \in \mathcal{M}$.
10. update Q-table.
11. $t := t+1$.
12.***Endwhile***

TABLE III
ADAPTIVE $\varepsilon-$GREEDY ALGORITHM

1.**Initialization**:
2.$OP_{pre}=0, OP_{cur}=0$, $cnt=0$.
3.***if*** random number $\xi \leq \varepsilon$
4. $OP_{cur}=Q_m^t(a_m^*), cnt=cnt+1$.
5. ***if*** $cnt=T_{ex}$
6. $\triangle=(OP_{cur}-OP_{pre})\times c_a$.
7. ***if***$\triangle>0$
8. $\varepsilon=\varepsilon-sigmoid(\triangle)$.
9. ***end if***
10. $OP_{pre}=OP_{cur}$, $cnt=0$.
11. ***end if***
12. choose action randomly.
13.***else***
14. choose action $a^*=\underset{a\in A}{\operatorname{argmax}} Q_m^t(s_m^t,a)$.
15.***end if***

for agent m, which indicates the expected discounted reward in the future when performing action a_c at state s_r. The learning process for $m-th$ agent is shown as table II, where T is the number of learning periods.

When choosing an action, balancing exploration and exploitation is a great challenge in reinforcement learning, among which exploration means that the agents take as many actions as possible to explore the reward of different actions and exploitation means that the agents choose the best action based on the current Q-table. It's a compromise between short-run and long-run reward. The most widely used method for exploration/exploitation is $\varepsilon-$greedy method. In the $\varepsilon-$greedy method, we denote $\varepsilon, 0 \leq \varepsilon \leq 1$ as the an exploration parameter to avoid suboptimal policy. Agents choose a random action with fixed probability ε. That is:

$$\pi(s)=\begin{cases} \text{random action from } \mathcal{A} & if \quad \xi<\varepsilon, \\ \underset{a\in\mathcal{A}}{\operatorname{argmax}} Q_m(s_m,a) & othersize, \end{cases} \tag{5}$$

where $\xi, 0 \leq \xi \leq 1$ is a uniform random number. We extend the $\varepsilon-$greedy method into an adaptive $\varepsilon-$greedy method, which changes ε according to the received rewards [16], and introduce the softmax method. The adaptive $\varepsilon-$greedy method introduce two parameter T_{ex} and c_a. T_{ex} denotes the maximal time to take action in the exploration mode before adapting ε. c_a is an adaptive parameter that is used to regularize ε. The adaptive $\varepsilon-$greedy algorithm at each step is shown as table III, where OP_{pre} denotes the previous maximal reward before adapting ε and OP_{cur} denotes the maximal reward so far. If the difference of OP_{cur} and OP_{pre} exceeds 0, ε is adapted to a smaller value. In the softmax method, we map a softmax function of $Q(s,a)$ to the action probability. The $m-th$ agent select an action a_m in the state s_m with a probability as follows:

$$P(a_m|s_m)=\frac{exp(Q_m(s_m,a_m)/\tau)}{\Sigma_{i=1}^{N} exp(Q_m(s_m,i)/\tau)}, \tag{6}$$

where τ is a positive parameter and N is the total number of actions for m$-th$ agent.

At each iteration of the learning process, the Q-value is updated according to the following function:

$$Q_m^{t+1}(s_m^t,a_m^t):=(1-\alpha)Q_m^t(s_m^t,a_m^t)+\alpha(R^t(s_m^t,a_m^t)+\gamma\max_{a'\in A}Q_m^t(s_m^{t+1},a')), \tag{7}$$

where α is learning rate.

IV. NUMERICAL RESULTS

In this section, we provide the simulation results of our proposed multi-agent reinforcement learning approach for SUs association in CR networks. Consider an OFDM-based cognitive radio system with multiple acess points, where all users are randomly distributed in a circle area with radius 1100 km, and move randomly within a certain sub-area. The user density is uneven in different sub-aree, where some sub-area with large user density and some sub-area with small user density. 3 APs are fixedly distributed in the coordinate points $(0,500), (-250\sqrt{3}, -250), (250\sqrt{3}, -250)$, and they dominate independent spectrum with bandwidth 70, 80 and 140. A total of 1000 SUs request for transmission in a Possion process with arrival rate 0.7. The spectrum occupation of PUs for these 3 APs follows Possion process of arrival rate 30, 30 and 40. For channel fading, the path loss exponent is 4, the variance of shadowing effect is 10dB and the amplitude of multipath fading is Rayleigh. In the multi-agent reinforcement learning procedure, we set the exploration coefficient ε as 0.1, the discount rate γ as 0.95 and the learning rate α as 0.3.

For comparison, we introduce max-SINR method, in which SUs choose an AP with maximal SINR to associate with and the load imbalance is unsolved.

Firstly, we compare the performance of the adaptive $\varepsilon-$greedy method and the softmax method on action choosing. Fig. 2 illustrates the whole utility as the iterations process of reinforcement learning using the adaptive $\varepsilon-$greedy method and the softmax method. We can see that the two method performs similarly. The whole system utility convergence to 100 at about the 20 thousand iteration. The curves fluctuate in a narrow range because of the stochastic exploration, the varying PUs occupation and SUs access requirement. For

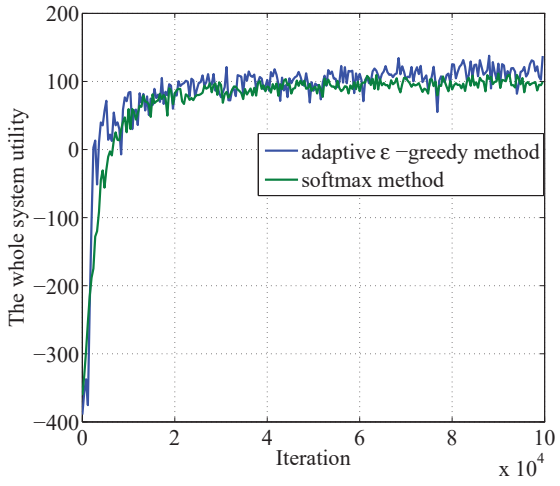

Fig. 2. The whole utility as the iterations process of reinforcement learning of the adaptive $\varepsilon-$greedy method and the softmax method.

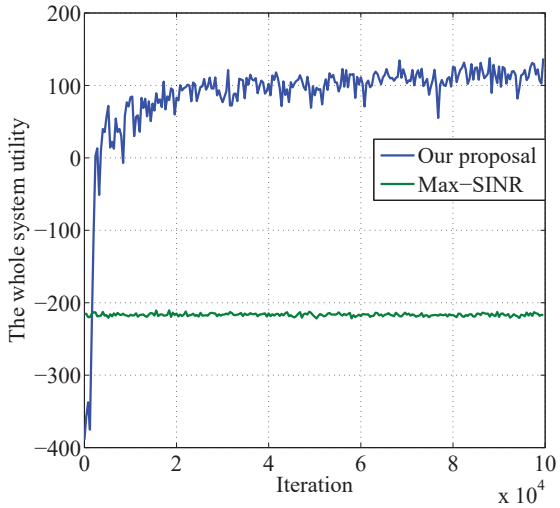

Fig. 3. The whole utility as the iterations process of reinforcement learning under heavy load.

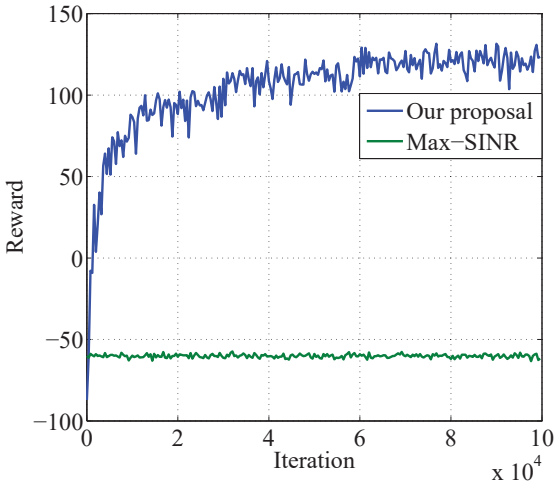

Fig. 4. The whole utility as the iterations process of reinforcement learning under light load.

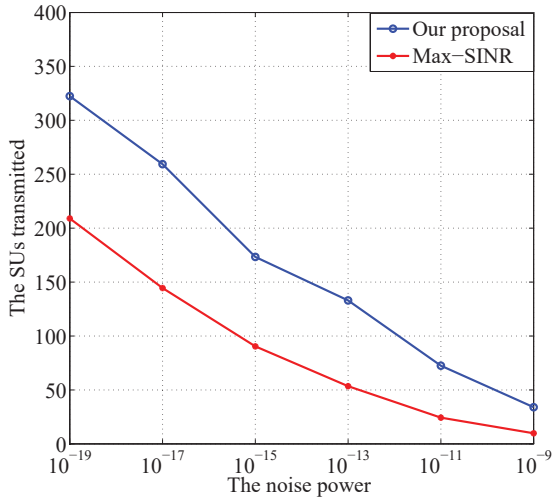

Fig. 5. The total number of SUs that has been transmitted successfully in the area after multi-agent learning process as a function of the noise power.

simplification, we adopt the adaptive $\varepsilon-$greedy method for action choosing in the following simulations.

We demonstrate the whole utility as the iterations process of multi-agent reinforcement learning under heavy load and light load in Fig. 3 and Fig. 4, respectively. The packets arrival rate of SUs is 0.7 under heavy load and 0.35 under light load, respectively. As can be seen in Fig. 3 and Fig. 4, through reinforcement learning, the whole system of our proposal performs much better than the max-SINR method, and the utility converge to an optimal utility. The light fluctuation in a narrow range is caused by stochastic exploration mode and explosive growth of PUs occupancy. In Fig. 3, we can see that the reward of our proposal is about -400 in the beginning and then convergence to about 100 when the iteration number reaches 20 thousand, while the reward of the max-SINR method maintains in -200. In Fig. 3, the reward of our proposal is -100 in the beginning and then convergence to about 130 when the the number of 60 thousand, while the reward of the max-SINR method contains in -50. That's because under heavy load, the limited spectrum resource can't support the massive SUs, larger number of SUs associated with APs could not be transmitted, then the reward is lower in the beginning. With the learning process, the APs derive policies on deciding the number of SUs for association according to the idle spectrum and the SINR distribution of SUs, the SUs that are transmitted successfully reach the capacity of the system.

Then we investigate how the channel condition influences system capacity. Fig. 5 shows the total number of SUs that have been transmitted successfully in the area after learning process as a function of the noise power. It shows that the

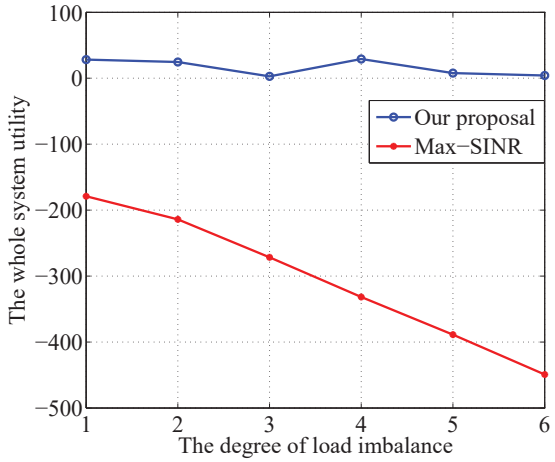

Fig. 6. The whole system utility after the multi-agent reinforcement learning process as a function of the degree of load imbalance.

sum capacity of the whole system decreases as the noise power increases, and our proposal outperforms the max-SINR method obviously. It can been explained intuitively. As the increase of noise power, the required spectrum to transmit the packets with the same size increase, then the whole system capacity decrease.

Finally, we investigate the the performance of our proposed algorithm in different degree of load imbalance. Fig. 6 depicts the whole system utility after the learning process as a function of the degree of load imbalance. We assume that the spectrum is equally managed by the three APs, and then adjust the space distribution of SUs to 6 degrees. The load imbalance is more serious as the degree from 1 to 6. It's obvious that our proposal performs much better than the max-SINR method, especially when the load imbalance is more serious. That's because our proposal intends to derive a policy to optimize the SUs associated to each AP by multi-agent reinforcement learning, the APs with heavy load will select less SUs to association with, then the residual SUs will associate with other APs with small SINR but light load. However, the max-SINR method cannot adjust user association according to the degree of load imbalance. Therefore, our proposed method is robust even when the load imbalance is quiet serious.

V. Conclusions

In this paper, we have proposed a multi-agent reinforcement learning approach to user association in CR networks. The CR system consists of multiple APs and massive SUs that distributed unevenly in an area. The APs are responsible for managing a certain spectrum that is licensed to PUs, and in addition, APs can not cooperate with each other and there is not a central controller. At each step, the APs use reinforcement learning to make optimal decisions on the SUs for association and notice selected SUs, then SUs choose one AP with maximal SINR. To illustrate the performance of our proposed multi-agent reinforcement learning for user association, we have conducted several simulations. The simulation results show that the multi-agent reinforcement learning process can derive optimal policies. Our proposed method outperforms the max-SINR method and has an excellent robustness performance when the degree of load imbalance is serious.

References

[1] Cisco whitepaper, "Cisco Visual Network Index: Forecast and Methodology, 2016-2021," June 5, 2017.

[2] F. C. Commission, "Facilitating opportunities for flexible, efficient,and reliable spectrum use employing cognitive radio technologies," *FCC Report.*, ET Docket 03-322, Dec. 2003.

[3] J. Mitola and G. Q. Maguire, "Cognitive radio: making software radios more personal," *IEEE Person. Commun.*, vol. 6, no. 4, pp. 13-18, Aug. 1999.

[4] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. Wong, R. Schober, L. Hanzo, "User Association in 5G Networks: A Survey and an Outlook," *Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018-1044, Jan. 2016.

[5] F. Wang, W. Chen, H. Tang, Q. Wu, "Joint Optimization of User Association, Subchannel Allocation, and Power Allocation in Multi-Cell Multi-Association OFDMA Heterogeneous Networks," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2671-2684, Mar. 2017.

[6] Y. Jin, L. Qiu, "Joint User Association and Interference Coordination in Heterogeneous Cellular Networks," *IEEE Commun. Lett.*, vol. 17, no. 12, pp. 2296-2299, Oct. 2013.

[7] X. Ge, X. Li, H. Jin, J. Cheng, V. C. M. Leung, "Joint User Association and User Scheduling for Load Balancing in Heterogeneous Networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3211-3225, Feb. 2018.

[8] S. Bayat, R. H. Y. Louie, Z. Han, Branka Vucetic, Yonghui Li, "Distributed User Association and Femtocell Allocation in Heterogeneous Wireless Networks," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 3027-3043, Jul. 2014.

[9] W. C. Ao, K. Psounis, "Approximation Algorithms for Online User Association in Multi-Tier Multi-Cell Mobile Networks," *IEEE/ACM Trans. Netw*, vol. 25, no. 4, pp. 2361-2374, Apr. 2017.

[10] R. S. Sutton, A. G. Barto,"Reinforcement Learning: An Introduction," *The MIT Press*, 2014.

[11] L. B. R. Babuska,B. D. Schutter, "Multi-agent reinforcement learning: An overview," *Studies in Computational Intelligence*, vol. 310, Berlin, Germany, pp. 183C221, 2010.

[12] A. G. Serrano, L. Giupponi, "Distributed Q-Learning for Aggregated Interference Control in Cognitive Radio Networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 1823-1834, Feb. 2010.

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

[14] J. Lundn, S. R. Kulkarni, V. Koivunen, H. Vincent. Poor, "Multiagent Reinforcement Learning Based Spectrum Sensing Policies for Cognitive Radio Networks," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 5, pp. 858-868, Apr. 2013.

[15] V. Raj, I. Dias, T. Tholeti, S. Kalyani, "Spectrum Access In Cognitive Radio Using a Two-Stage Reinforcement Learning Approach," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 20-34, Jan. 2018.

[16] A. S. Mignona, R. L. A. Rocha, "An Adaptive Implementation of -Greedy in Reinforcement Learning," *Procedia Comput. Sci.*, vol. 109, pp. 1146-1151, 2017.