Deep Reinforcement Learning for Spectrum Sharing in Future Mobile Communication System

Sizhuang Liu[1], Tengjiao Wang[1], Changyong Pan[1,2], Chao Zhang[1,2], Fang Yang[1,2], and Jian Song[1,2]
[1]Department of Electronic Engineering, Tsinghua University
Beijing National Research Center for Information Science and Technology (BNRist), Beijing 100084, P. R. China
[2] Peng Cheng Laboratory, Shenzhen 518055, P. R. China

2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) | 978-1-6654-4908-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/BMSB53066.2021.9547161

***Abstract*—In recent years, the rapid growth of mobile communication services makes spectrum resources become increasingly scarce. This paper considers the multi-dimensional resource allocation problem in unlicensed spectrum communication system. A training method based on deep reinforcement learning is proposed to generate a spectrum sharing and power control strategy for secondary users in the communication system. Deep Q-Network and Deep Recurrent Q-Network are chosen as the structure of neural network. Experiments are conducted to investigate the effectiveness of the algorithm. The results show that collision rate decreases in training while average reward rises.**

***Index Terms*—Resources allocation, machine learning for communications, dynamic spectrum access, deep reinforcement learning, deep Q-network.**

I. Introduction

With the development of mobile communication technologies, the data traffic business is growing rapidly. Scarce licensed spectrum with low utilization can hardly support such large-scale services [1]. Data traffic of licensed spectrum is becoming increasingly saturated. Therefore, researchers propose to adopt the technology of unlicensed spectrum to alleviate the load pressure of licensed spectrum and improve the network capacity. The next generation mobile communication system requires lower latency, higher transmission rate and better spectrum efficiency. Consequently, a more effective spectrum allocation strategy is needed.

Many studies have focused on the dynamic spectrum access technology. Some researches apply game theory to spectrum sharing. In [2] researchers showed a game-theoretic algorithm in cognitive radio (CR) technology. Reference [3] presented a spectrum sharing method using decision tree and game theory. In [4] a pricing strategy based on Stackelberg game in spectrum sharing system was proposed.

Deep learning uses artificial neural network, which can establish complex mapping relationship between input and output data. It has strong feature extraction ability, which also means strong perception ability. Deep reinforcement learning (DRL) combines the perceptual ability of deep learning with the decision-making ability of reinforcement learning. By using their advantages, DRL can deal with the decision making problems of some complex systems. It is a practicable way to apply DRL technology to resource allocation in communication system. In recent years, researchers started using deep reinforcement learning in modern networks[5]. Reference [6] describes a Deep Q-Network (DQN) based model for distributed dynamic spectrum access. In [7] researchers shows a DQN training approach for power control in cognitive radios. A power control strategy based on deep reinforcement learning was introduced in [8] to make secondary users adjust to the appropriate state in a limited time. With an appropriate reward function, secondary users share the spectrum with primary users on the premise of avoiding interference. Evolutionary game theory is combined with DQN to manage the dynamic spectrum in [9]. Double Deep Q-Network (DDQN) [10] is also applicated in spectrum sharing. In [11], a spectrum management framework based on deep Q-learning is studied, and the echo state network is used to realize the process of Q-learning. The users can manage the spectrum intelligently and individually through this method. Since Deep Q-Network is difficult to solve the problem under partial observation conditions, Deep Recurrent Q-Network (DRQN) is adopted in the dynamic access problem of discrete frequency channels [12]-[14]. However, Most of the existing work has limitations. For instance, studies in [12] and [13] mainly focused on dynamic spectrum access, but ignored the influence of transmission power of users. Additionally, they had not fully explored the reward setting problem. In [7] the quality of communication was considered and a power control method was developed, but they built a simple model with only one primary transmitter and one secondary user. In practical application, the system often consists of multiple users. [15] noticed both the allocation of power and spectrum resources in cognitive network, but the traditional reinforcement learning method was used.

The aim of this study is to solve the problem of spectrum sharing and power control of the unlicensed spectrum communication system by a method based on deep reinforcement learning. In this article, we propose an algorithm for spectrum access strategy on the basis of DQN training with multiple agents. Experience replay and freezing target networks are also applied in training. The simulation result presents that SUs can learn the spectrum access strategy through training in the case of only a single SU as well as multiple SUs. Both collision rate and average reward have been improved using the method proposed in this paper. The DRL-based method have a better performance compared with random access and greedy method.

The remainder of the paper is organized as follows. Section

978-1-6654-4908-3/21/$31.00 ©2021 IEEE
Authorized licensed use limited to: Tsinghua University. Downloaded on July 29,2026 at 09:18:14 UTC from IEEE Xplore. Restrictions apply.

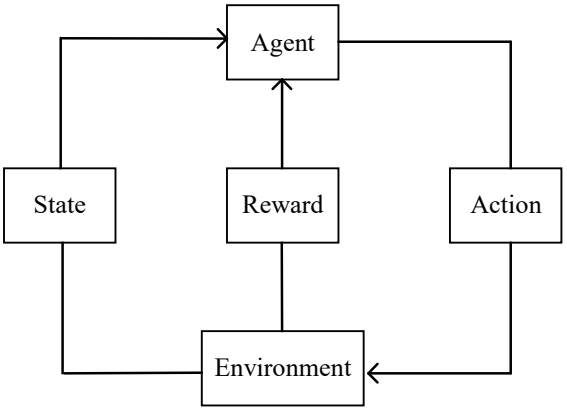


Fig. 1. Principle of the reinforcement learning techniques.

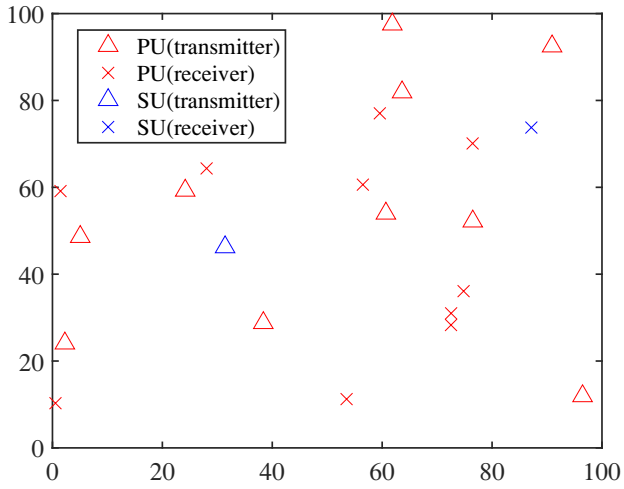


Fig. 2. Positions of SUs and PUs.

II describes the system model we select and the research method we adopt. The result of our simulation is presented in Section III. We conclude our work in Section IV.

II. System Model

We consider an unlicensed spectrum communication system with M primary users (PUs) and N secondary users (SUs). Each user contains a transmitter and a receiver. We assume that for each PU there is one channel. In each time slot, every PU has two possible states: occupying channel or not occupying channel. The state change of PUs is determined by the state transition matrix of Markov process. Every SU has a certain probability to be active. In this case, this SU has the need to transmit data. The power level of every SU is selected from a finite set with cardinality L. The positions of transmitters and receivers are randomly generated in the specified area.

In each time slot, SUs make observations on channels and collect information. According to the learning strategy, every

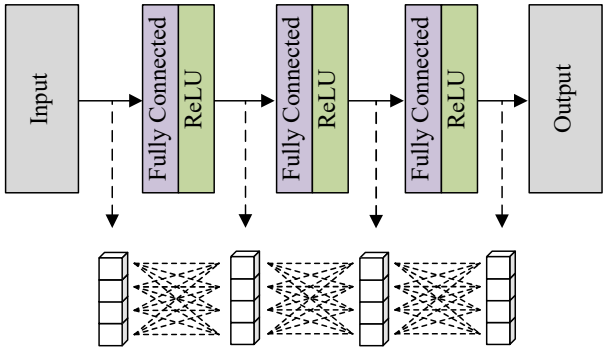


Fig. 3. Illustration of the adopted DQN structure.

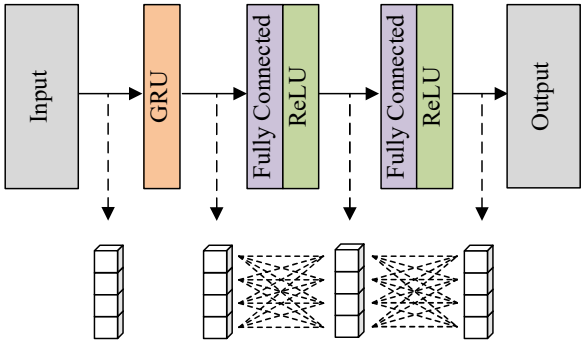


Fig. 4. Illustration of the adopted DRQN structure.

SU selects a channel and a power level, or does not select any channel. After that, the next time slot comes and the state of channels changes again. If a SU is active in the next time slot, data transmission is performed with the selected channel and power level. Otherwise, the SU does not access any channel in the next time slot. The SU receivers record the SINR. The system gives the reward on the basis of the SINR values and the transmission powers of SUs. Fig. 2 shows a schematic diagram of positions of SUs and PUs ($M = 10$ and $N = 1$).

III. DRL-based spectrum sharing method

Q-learning is a widely-used value-based training algorithm in reinforcement learning. However, Q-learning becomes inefficient when dealing with large-scale problems [16]. In 2013, the DeepMind team proposed a DQN algorithm based on Q-learning and deep learning, and achieved great performance in Atari game [17]. In this study, agent is defined as SUs, and the action denotes the channel and power usage chosen by SUs. Reward function is designed based on the quality of communication, related to whether collision exists and the power used by SU. WINNER II channel model [18] is adopted to calculate the signal to interference plus noise ratio (SINR) at the receivers of SUs. If SU i accesses a channel where the

978-1-6654-4908-3/21/$31.00 ©2021 IEEE
Authorized licensed use limited to: Tsinghua University. Downloaded on July 29,2026 at 09:18:14 UTC from IEEE Xplore. Restrictions apply.

PU is currently inactive, the SINR for the receiver of SU i is calculated as [1][7]

$$\mathrm{SINR}_i = \frac{p_i \left|h_{ii}^2\right|}{\sum_{k \neq i} p_k \left|h_{ki}^2\right| + BN_0}, i = 1, 2, \ldots, N, \quad (1)$$

where h_{ii} and h_{ki} denotes the channel gain from the transmitter i to the receiver i and from the transmitter k to the receiver i. p_i and p_k correspond to transmission power of SU i and k respectively. N_0 and B represent the noise spectral density and channel bandwidth. When a SU accesses a channel where the PU is currently active, collision happens between a SU and a PU. Consequently, the SU receive a negative reward. In another case, if the SU chooses not to access any channel, the reward is set as 0. In other cases, the reward is mainly depended on SINR. When only one transmission power level is considered ($L = 1$), the reward function is designed as [1]

$$r = \begin{cases} 0 & , \text{ no channel access,} \\ -C & , \text{ collision with PU,} \\ \log_2(1 + \mathrm{SINR}) & , \text{ other cases,} \end{cases} \quad (2)$$

where C is a positive constant. Considering multiple transmit powers, the reward function is defined as

$$r = \begin{cases} 0 & , \text{ no channel access or} \\ & \text{SINR not up to standard,} \\ -C_1 & , \text{ collision with PU,} \\ C_2 - C_3 \log_2 P & , \text{ SINR up to standard,} \end{cases} \quad (3)$$

where C_1, C_2, C_3 are positive constants and P represents the transmission power. It should be noted that, compared with the absolute value, the change trend and relative value of reward have greater reference significance. We generate a DQN for each SU. The usage state of the channels is selected as the observation, which is the input of the DQN. Our DQN structure consists of fully connected layer and outputs the Q value of each action. Algorithms based on Gradient Descent are selected as the optimization method, including stochastic gradient descent (SGD) and mini-batch gradient descent (MBGD). After training, the final policy of SU is determined based on the equation

$$a_{t+1} = \mathrm{argmax}_a Q(s, a), \quad (4)$$

where $Q(s, a)$ is the output of DQN. In another simulation, we replace DQN with Deep Recurrent Q-Learning Network (DRQN), which includes a Gated Recurrent Unit (GRU) layer besides fully connected layers. GRU is a simpler neural network structure than Long Short-Term Memory (LSTM), which can process time series data and avoid long-range dependence [19]. The performance of both neural network structures is evaluated.

In the training process of DQN, we generate random weights to initialize the network. Each SU makes an observation and choose the action based on ϵ-greedy method [1]

$$a_{t+1} = \begin{cases} \mathrm{argmax}_a(Q(S, a)) & , \text{ with probability } 1 - \epsilon, \\ \text{random action} & , \text{ with probability } \epsilon, \end{cases} \quad (5)$$

where ϵ is a real number between 0 and 1. The system calculates the reward and the next state, and stores the information. After collecting enough training data, we adopt the optimization method to update the weights of DQN. Freezing target networks [20] is applied in order to reduce oscillation and improve training effect. The loss function is defined as [17]

$$L(\theta) = E[(r + \gamma Q^*(s', a') - Q(s, a))^2], \quad (6)$$

where θ represents weights of the networks, γ is the discounted rate of future rewards. Q^* is the target network, and Q is for estimation. Q is updated in each iteration, but Q^* is not. Every once in a while, we assign the value of Q to Q^*. Before updating, we rearrange the order of training data and apply experience replay technique to break their correlation and improve training efficiency. The experience of the agents is stored in the replay memory [16], waiting to be sampled in subsequent training. The utilization of training data can also be elevated.

In the training process of DRQN, while updating the weights, we keep the sequential information instead of rearranging the order. While updating weights of the network, continuous sequences are sent into the network for calculation. The other procedures are the same as DQN.

IV. Simulation Results

Simulations of changes of collision rate and average reward in DQN and DRQN training are shown in Fig. 5 through Fig. 8. Results of DQN and DRQN are compared with random access and greedy algorithm. In the random access method, SUs randomly select an action in each time slot. In the greedy method, SUs find the channel with the largest expected reward according to the state transition probability of PUs and accesses it.

In the first simulation, we set $M = 10$ and $N = 1$, which represents 10 primary users and a secondary user. The transmission power of SUs is fixed. The reward function is designed as equation 2. In each iteration, new sequences are generated according to equation 5, and are added to the replay memory. Meanwhile, sequences are randomly sampled from the replay memory and are used to update the weight of the network as well as calculating average loss, average reward and collision rate.

Results indicate that the average reward increased significantly while collision rate decreased in training, which verifies the effectiveness of the optimization algorithm. The collision rate has fallen rapidly at the beginning and then gradually becomes stable. The trend of reward is opposite to that of collision rate, which increases in training and approximately converges around 27. There is no significant difference between the two network structures. The convergence speed of DRQN is slower than that of DQN. Both of them are better than random access in collision rate and reward, and are close to greedy method. Since the greedy method is theoretically optimal in the case of single SU, this result is acceptable. The SU can finally get a set of strategies to achieve high reward while avoiding collisions as much as possible.

978-1-6654-4908-3/21/$31.00 ©2021 IEEE
Authorized licensed use limited to: Tsinghua University. Downloaded on July 29,2026 at 09:18:14 UTC from IEEE Xplore. Restrictions apply.

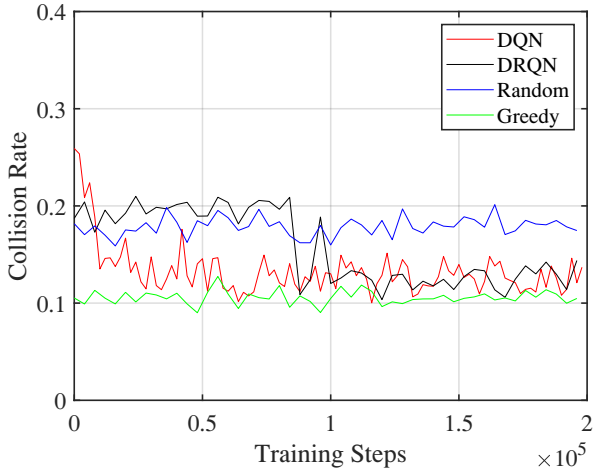


Fig. 5. The collision rate of the proposed method with different training steps (single SU).

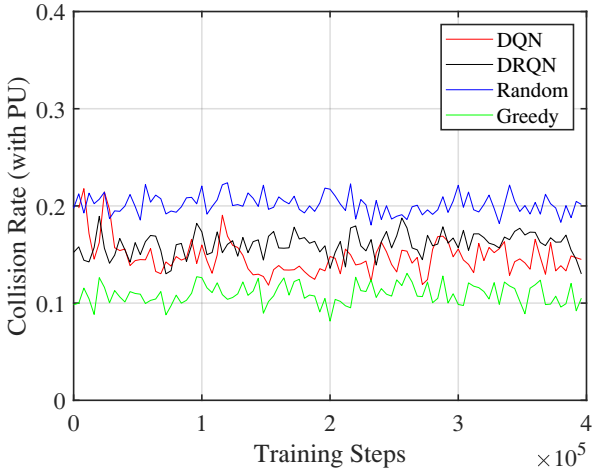


Fig. 7. The collision rate of the proposed method with different training steps (multiple SUs).

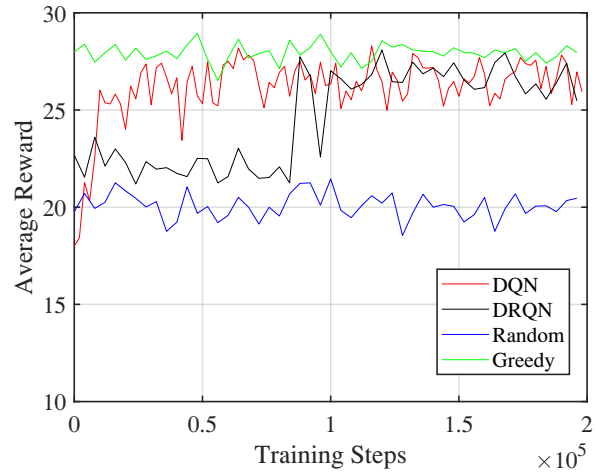


Fig. 6. The average reward of the proposed method with different training steps (single SU).

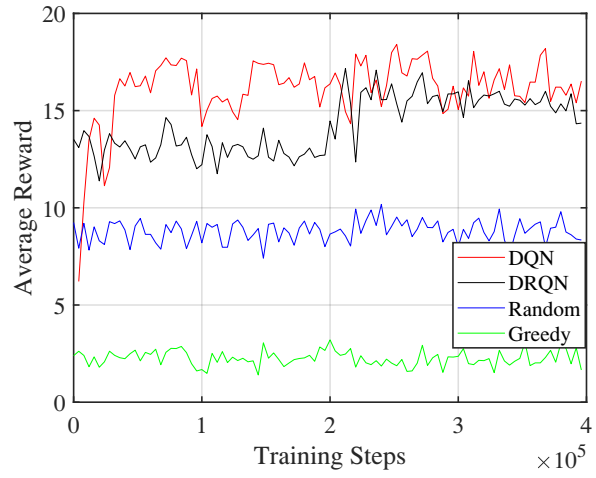


Fig. 8. The average reward of the proposed method with different training steps (multiple SUs).

In the other simulation, we set $M = 8$, $N = 3$ and $L = 3$. Multiple SUs exist and different power levels are available. Unlike the case of single SU, the reward function is calculated by equation 3. The simulation results show that DQN and DRQN method can converge to a high reward compared with random access and greedy method in this case. While adopting the greedy algorithm, all of SUs tend to access the best channel, so the collision rate between SUs rises and therefore the average reward is low. The DQN structure has advantages over the DRQN structure in convergence speed again. In both circumstances, the SUs can finally get a set of strategies to access channels with proper power while avoiding collisions with PUs and SUs.

V. Conclusion

This article proposed a spectrum sharing method for unlicensed users in the communication system. Secondary users are modeled as agents in reinforcement learning and the target of communication is modeled as the reward function based on the signal to interference plus noise ratio and whether collision happens. DQN and DRQN are selected as the neural network structure. In the process of deep reinforcement learning, experience replay, freezing target networks and mini-batch gradient descent are utilized for improving training effect. The algorithm converges at high reward and low collision rate, making good use of spectrum resources. The correctness and effectiveness of the algorithm are confirmed in the simulation.

978-1-6654-4908-3/21/$31.00 ©2021 IEEE
Authorized licensed use limited to: Tsinghua University. Downloaded on July 29,2026 at 09:18:14 UTC from IEEE Xplore. Restrictions apply.

ACKNOWLEDGMENT

This work was supported by the China National Key R&D Program of China under Grant 2020YFB1807600, Peng Cheng Laboratory under Grant PCL2021A10, and Guoqiang Institute of Tsinghua University.

REFERENCES

[1] H. Chang, H. Song, Y. Yi, J. Zhang, H. He and L. Liu, "Distributive Dynamic Spectrum Access Through Deep Reinforcement Learning: A Reservoir Computing-Based Approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1938-1948, Apr. 2019.

[2] W. Zhou, Q. Zhu and Y. Ling, "An Improved Game-Theoretic Algorithm for Competitive Spectrum Sharing," in *2010 International Conference on Communications and Mobile Computing*, Shenzhen, China, Apr. 2010.

[3] F. Cai, Y. Gao, L. Cheng, L. Sang and D. Yang, "Spectrum sharing for LTE and WiFi coexistence using decision tree and game theory," in *2016 IEEE Wireless Communications and Networking Conference*, Doha, Qatar, Apr. 2016.

[4] I. Ahmad, Z. Wei, Z. Feng, Y. Bai, Q. Zhang and P. Zhang, "Joint price and power allocation under interference constraint for dynamic spectrum access networks," in *2014 IEEE International Symposium on Dynamic Spectrum Access Networks*, McLean, VA, USA, Apr. 2014.

[5] N. C. Luong et al., "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133-3174, Fourthquarter 2019.

[6] X. Gao, Z. Dou and L. Qi, "A New Distributed Dynamic Spectrum Access Model Based on DQN," in *2020 15th IEEE International Conference on Signal Processing*, Beijing, China, Dec. 2020.

[7] 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," *IEEE Access*, vol. 6, pp. 25463-25473, Apr. 2018.

[8] F. Ye, Y. Zhang, Y. Li, T. Jiang and Y. Li, "Power Control Based on Deep Q Network with Modified Reward Function in Cognitive Networks," in *2020 IEEE USNC-CNC-URSI North American Radio Science Meeting (Joint with AP-S Symposium)*, Montreal, QC, Canada, Jul. 2020.

[9] P. Yang et al., "Dynamic Spectrum Access in Cognitive Radio Networks Using Deep Reinforcement Learning and Evolutionary Game," in *2018 IEEE/CIC International Conference on Communications in China*, Beijing, China, Aug. 2018.

[10] Y. Xu, J. Yu, W. C. Headley and R. M. Buehrer, "Deep Reinforcement Learning for Dynamic Spectrum Access in Wireless Networks," in *MILCOM 2018 - 2018 IEEE Military Communications Conference*, Los Angeles, CA, USA, Oct. 2018.

[11] H. Song, L. Liu, J. Ashdown and Y. Yi, "A Deep Reinforcement Learning Framework for Spectrum Management in Dynamic Spectrum Access," *IEEE Internet Things J.*, Jan. 2021.

[12] Y. Xu, J. Yu and R. M. Buehrer, "Cache-Enabled Dynamic Spectrum Access via Deep Recurrent Q-Networks with Partial Observation," in *2019 IEEE International Symposium on Dynamic Spectrum Access Networks*, Newark, NJ, USA, Nov. 2019.

[13] Y. Xu, J. Yu and R. M. Buehrer, "The Application of Deep Reinforcement Learning to Distributed Spectrum Access in Dynamic Heterogeneous Environments With Partial Observations," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4494-4506, Jul. 2020.

[14] Y. Xu, J. Yu and R. M. Buehrer, "Dealing with Partial Observations in Dynamic Spectrum Access: Deep Recurrent Q-Networks," in *MILCOM 2018 - 2018 IEEE Military Communications Conference*, Los Angeles, CA, USA, Oct. 2018.

[15] S. Maulik, R. Roy, A. De and A. Bhatttacharya, "Online dynamic resource allocation in interference temperature constrained cognitive radio network using reinforcement learning," in *2012 International Conference on Signal Processing and Communications*, Bangalore, India, Jul. 2012.

[16] O. Naparstek and K. Cohen, "Deep Multi-User Reinforcement Learning for Distributed Dynamic Spectrum Access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310-323, Jan. 2019.

[17] V. Mnih et al., "Playing Atari with deep reinforcement learning," *Deepmind Technol.*, London, U.K., Rep., Dec. 2013.

[18] J. Meinilä, P. Kyösti, T. Jämsä, and L. Hentilä, "Winner II channel models," in *Radio Technologies and Concepts for IMT-Advanced*. Hoboken, NJ, USA: Wiley, 2009, pp. 39–92

[19] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, "Empirical evaluation of gated recurrent neural networkson sequence modeling," *arXiv preprint arXiv:1412.3555*, Dec. 2014.

[20] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-533, Feb. 2015.

978-1-6654-4908-3/21/$31.00 ©2021 IEEE
Authorized licensed use limited to: Tsinghua University. Downloaded on July 29,2026 at 09:18:14 UTC from IEEE Xplore. Restrictions apply.