QoE-Fairness Tradeoff Scheme for Dynamic Spectrum Allocation Based on Deep Reinforcement Learning

Le Tong
The Sixty-third Research Institute, National University of Defense Technology, Nanjing, China
tongle63s@163.com

Yangyi Chen*
The Sixty-third Research Institute, National University of Defense Technology, Nanjing, China
chenyangyi09@nudt.edu.cn

Xin Zhou
College of Communications Engineering, Army Engineering University of PLA, Nanjing, China
lgdxzhouxin@126.com

Yifu Sun
The Sixty-third Research Institute, National University of Defense Technology, Nanjing, China
Sunif19@nudt.edu.cn

ABSTRACT

In order to meet the tradeoff of QoE(quality of experience)-Fairness when spectrum resources are insufficient, it is necessary to study the dynamic spectrum allocation problem, especially in the scenario where a base station who acts as a single agent wishes to reliably communicate with the multiple users by centrally managing the spectrum resources. To overcome the fact that user behavior and environment are unknown and dynamic, this paper modeled the dynamic spectrum allocation as an optimization problem, and put forward a dynamic spectrum allocation strategy which based on adaptive deep Q-learning network (ADQN). On this basis, a new reward function is designed to drive the learning process which considering different types of user's communication needs, and a priority experience replay strategy is proposed to accelerate network training speed which based on reducing time error. Moreover, simulation results show that the proposed strategy can accelerate the convergence speed of ADQN and improve the rationality and effectiveness of dynamic spectrum allocation.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Learning paradigms**; • **Reinforcement learning**;

KEYWORDS

dynamic spectrum allocation, quality of experience, deep reinforcement learning, fairness

ACM Reference Format:
Le Tong, Yangyi Chen*, Xin Zhou, and Yifu Sun. 2021. QoE-Fairness Tradeoff Scheme for Dynamic Spectrum Allocation Based on Deep Reinforcement Learning. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3487075.3487137

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487137

1 INTRODUCTION

Dynamic spectrum sharing has attracted great attention in the industry and academia and it is one of the effective ways to solve the spectrum traffic [1]. However, with the explosive growth of the radio equipment types and data traffic, the spectrum management is facing more and more challenge. Specifically, the existing spectrum management approaches which ignore the user's satisfaction and the quality of service (QoS)-based model cannot be applied to all the scenarios. Thus, to improve user's satisfaction, the quality of experience (QoE)-based model is proposed to replace the QoS in [2]. The mean opinion score (MOS) is the most widely adopted QoE models [3], which provides a unified and common metric for different applications, therefore, we can use it to carry out integrated traffic management and resource allocation across traffic of dissimilar features. However, wireless spectrum resources are finite, which cannot withstand the exponential growth in the number of terminals and data traffic. In addition, the fixed frequency allocation strategy wastes spectrum resources and exacerbates the resource gap [4]. To address above problems, the dynamic spectrum allocation has been proposed as a promising approach to improve the spectrum efficiency [1],[5],[6].

In recent years, reinforcement learning (RL) and deep RL (DRL) have been introduced into the field of dynamic spectrum allocation in order to cope with the dynamic characteristics of spectrum environment [7],[8]. The performance of traditional RL methods is limited by the state space and action space of the problem [9]. Deep Q-learning network (DQN) is a classical DRL algorithm, which combines DL and RL, enabling agent to obtain approximate solution of complex system in various complicated states and actions [10]. Experience replay is a typical method of DQN training. In this method, the samples which generated by the interaction between agents and the environment are saved to form an experience pool [11]. During the training, several samples are randomly selected from the experience pool to train the Q-network, so it breaks the correlation between training samples and guarantees the convergence of value function [12],[13].

In the dynamic spectrum sharing system with the limited spectrum resources, it is challenging to satisfy communication requirements of various types of users [14-18]. Considering that there is no cooperation between primary user (PU) and secondary user (SU), a spectrum sharing method based on DQN to dynamically adjust its transmission power is proposed in [14]. Moreover, to address the

multi-channel transmission problem, Li et al. proposed the dynamic spectrum sensing and aggregation method based on the centralized DQN in [15]. In addition, Shi et al. proposed a solution for spectrum resource management in IIoT networks with the aim of facilitating limited spectrum sharing among different types of user equipment in [16]. Since it is very difficult to obtain the environmental information in real-time, the optimization problem was modeled as a Markov decision process (MDP) and a centralized resource management scheme was proposed in [17], whereas a partially observable Markov decision process (POMDP) is modeled in [15]. Furthermore, to obtain the state information and adapt to the dynamic nature of the spectral environment, the Q-learning algorithm and the long short-term memory (LSTM) DQN algorithm were applied to optimize the spectrum resources management policy in [17] and [18], respectively.

Motivated by above, this paper investigates a dynamic spectrum allocation method with the QoE-Fairness tradeoff based on ADQN with prioritized experience replay. The main contributions of this paper are summarized below.

1) A general spectrum sharing model with wide applications based on a novel QoE model is proposed in this paper. Considering the tradeoff of QoE and fairness, the dynamic spectrum allocation problem is modeled as an optimization problem where a certain network utility is maximized while meeting the QoE of different kinds of Users.
2) An adaptive Deep Q-Network is proposed to solve the dynamic spectrum allocation problem, where ADQN can quickly obtaining the optimal strategies via the proposed reward function and prioritized experience replay strategy.
3) Numerical results demonstrate that, the proposed ADQN algorithm can significantly improves the mean of QoE, and has faster convergence speed as compared to existing approaches, which confirms its effectiveness and superiority to solve dynamic spectrum allocation problem.

2 SYSTEM MODEL

We consider an uplink dense wireless network with $N = \{1, 2, \cdot s, N\}$ independent orthogonal subchannels and $K = \{1, 2, \cdot s, K\}$ user requirements. As a typical network scenario, each user can access only one subchannel in one time slot. If multiple users need to transmit signal on the same subchannel at the same time slot, there will be a conflict. Different from the concept of primary users and secondary users in traditional cognitive radio, this paper does not fix the priority attribute of users in dynamic spectrum allocation problem. Users are allowed to transmit data on the channel allocated by the base station only when some constraints are met.

As shown in the Figure 1, all users are divided into two categories and randomly distributed around the base station. Suppose this scenario, there are K_1 first type users, which are data traffic users, and we assume that the spectrum use demand of each data traffic user follows an independent Poisson distribution process. And there are K_2 second category users as video traffic users, and each video traffic user periodically performs data acquisition and upload. In this scenario the transmission on the sub-channel is successfully only when the sub-channel is assigned to one user at the slot $t$and

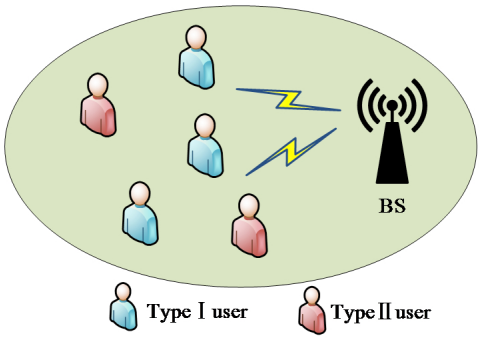

Figure 1: System Model.

the quality of experience of the communication satisfies the user's needs; otherwise, it is transmission failure due to conflict collision, or when the quality of experience is unsatisfactory to the user's needs. At the beginning of each time slot, the users wait for the result of the spectrum resource allocation.

2.1 Transmission Model

Suppose that the total bandwidth is $B$and it is evenly divided into N subchannels, the bandwidth of each subchannel is expressed as B_{w}. We set $D_{kn}(t), (k \in K, n \in N)$ as a binary indicator, where $D_{kn}(t) = 1$ indicates that the subchannel n is allocated to the user k in time slot t, and $D_{kn}(t) = 0$ indicates that the subchannel $n$is not allocated in time slot t. All binary indicators together constitute the spectrum allocation matrix $\mathbf{D}$, which is expressed as follows.

$$\mathbf{D} = \begin{bmatrix} D_{11} & D_{12} & \cdot s & D_{1n} \\ D_{21} & D_{22} & & \vdots \\ \vdots & & \ddots & \\ D_{k1} & \cdot s & & D_{kn} \end{bmatrix} \tag{1}$$

For the users, if it communicates, the SNR can be expressed as follows:

$$SNR_{kn}(t) = D_{kn} * \frac{P^I |h_{kn}(t)|^2}{\sigma^2}, k \in \mathbf{K}, n \in \mathbf{N} \tag{2}$$

where, $\sigma^2$is the received noise power, $P^I$is the transmission power, and $h_{kn}(t)$ is the fading factor allocated to the subchannel n by the user k equipment in slot t.

Since only one user equipment is allowed to access a single subchannel in one time slot, there is no co-channel interference. Therefore, the instantaneous rate of transmission can be given by the following.

$$R_{kn}(t) = B_{\mathrm{w}} \log(1 + SNR_{kn}(t)), k \in K \tag{3}$$

Assuming that the signal modulation mode is BPSK, the end-to-end packet loss rate P_{loss} can be expressed as follows.

$$P_{\mathrm{loss}} = 1 - (1 - P_{\mathrm{b}})^l \tag{4}$$

$$P_{\mathrm{b}} = \frac{1}{2}(1 - \sqrt{\frac{SNR_{kn}(t)}{1 + SNR_{kn}(t)}}) \tag{5}$$

where P_{b} is the bit error rate of coherent detection, and l is the bit length of each data packet.

2.2 MoS-based QoE Measurement Model

This paper considers the impact of transmission rate R_{kn}, packet loss rate P_{loss}, transmission delay T_{S} and interrupt time T_{I} on the quality of two different QoE. The transmission delay T_{S} is defined as the time from the user waiting for the arrival of the system service, and the interruption time T_{I} is defined as the extra time spent during the service due to the system aborting the resource allocation.

In this paper, the mean opinion score (MOS) is used as the QoE index to measure user's satisfaction. The relationship between transmission characteristics and MOS of different applications is described as follows.

- Data traffic

For data traffic, the MOS is calculated based on the transmit rate, R_{kn}, experienced by the end user, as follows [19].

$$M_{\text{A}}^{i} = \frac{a_1 \log_{10}[a_2 R_{kn}(1 - P_{loss})]}{1 + a_3({T_{\text{S}}}^2 - T_{\text{S}}) + a_4 {T_{\text{I}}}^2} \tag{6}$$

where $M_{\text{A}}^{i} \in [1, 10]$ represents the MOS value of the data traffic user i. $a_1 - a_4$are the control parameters of MOS value of data transfer users, which determined by the highest and lowest QoE value.

- Video traffic

As a real-time transmission service, the QoE model of video traffic is as follows [20]:

$$M_{\text{V}}^{i} = \frac{b_1}{(1 + e^{b_2(\varphi - b_3)})(1 + b_4 {T_{\text{S}}}^2 + b_5 {T_{\text{I}}}^2)} \tag{7}$$

where $M_{\text{V}} \in [1, 10]$ represents the MOS value of the video traffic user i, $b_1 - b_5$ are the control parameters of the MOS value of video traffic user, which determined by the highest and lowest QoE value, and φ is the peak signal-to-noise ratio of the image in the video, which is defined as follows:

$$\varphi = 10 \lg(\frac{a_{\max}^2}{MSE}) \tag{8}$$

$$MSE = \frac{\theta}{R_{kn} - R_0} + G_0 + \delta \cdot P_{\text{loss}} \tag{9}$$

where $a_{\max}$ is the maximum value of the image point color, MSE is the mean square error between the distorted video and the reference video, θ, R_0 and G_0 are the distortion parameters depending on the content and coding structure of the encoded video sequence, and δ is the parameters related to determining the compressed video sequence [21],[22].

2.3 Performance Indicators

For dynamic spectrum sharing system, it is necessary to consider the QoE of all users. And the mean QoE of all users in the system is defined as follows.

$$M_{\text{A+V}} = \frac{1}{10N}(\sum_{i=1}^{K_1} M_{\text{A}}^{i} + \sum_{i=1}^{K_2} M_{\text{V}}^{i}) \tag{10}$$

Due to the different location and requirements of users, the QoE of varies from different users with the same spectrum resources allocated. Therefore, it is necessary to consider the fairness index as a measure of spectrum allocation. In order to measure the change of QoE value after scheduling, especially the change of fairness, the fairness index is introduced. The fairness index F is defined as follow.

$$F = \frac{(\sum_{i=1}^{K_1} M_{\text{A}}^{i} + \sum_{i=1}^{K_2} M_{\text{V}}^{i})^2}{N(\sum_{i=1}^{K_1} {M_{\text{A}}^{i}}^2 + \sum_{i=1}^{K_2} {M_{\text{V}}^{i}}^2)} \tag{11}$$

where F tends to 1 to indicate fairness and tends to 0 to indicate unfairness.

2.4 Problem Formulation

In a system where multiple types of users coexist, if only maximizing the mean QoE of all users is considered, spectrum resources will be allocated to users with low satisfaction requirements as much as possible, which will cause great unfairness. Therefore, considering the lack of resources, that is, when $N < K$, spectrum resources cannot be allocated to all users at the same time in this paper, propose an optimization problem that considers a tradeoff between the mean of QoE and fairness to address dynamic spectrum allocation problem The optimization problem is expressed as follows.

$$\begin{aligned}
&\max_{\mathbf{D}} U = \lambda M_{\text{A+V}} + (1 - \lambda)F \\
&s.t. D_{kn}(t) \in \{0, 1\}, (k \in \mathbf{K}, n \in \mathbf{N}) \\
&\sum_{n=1}^{N} D_{kn}(t) \le 1, (k \in \mathbf{K}, n \in \mathbf{N}) \\
&\sum_{k=1}^{K} D_{kn}(t) \le 1, (k \in \mathbf{K}, n \in \mathbf{N}) \\
&M_{\text{A}} \ge M_{\text{Amin}} \\
&M_{\text{V}} \ge M_{\text{Vmin}}
\end{aligned} \tag{12}$$

where $\lambda \in (0, 1)$ represents the weighting factor of the mean QoE of all users and fairness, M_{Amin} represents the lowest QoE threshold of data traffic users, and M_{Vmin} represents the lowest QoE threshold of video traffic users. Under the tradeoff between the mean QoE of all users and fairness, it is necessary to find the optimal allocation matrix $\mathbf{D}$ to maximize the objective function.

3 MDP ANALYSIS

3.1 State and Action Space

The dynamic spectrum allocation problem can be modeled as MDP in this paper. The base station (BS) is used as the agent to observe the status information of the whole system and to implement dynamic spectrum allocation. The agent gets observation as follows.

$$\mathbf{O}(t) = \{\mathbf{M}(t), \mathbf{T_S}(t), \mathbf{T_I}(t)\} \tag{13}$$

where $\mathbf{M}(t)$ is the spectrum demand information of all current users composed of binary numbers; $\mathbf{T_S}(t)$ is the service delay information of all current users; and $\mathbf{T_I}(t)$ is the service interruption time information of all current users.

In reality, it is costly to obtain real-time environmental observations. Therefore, unlike existing literatures which take the current observed information as the input state of the agent, this paper defines the real-time input state as:

$$\mathbf{S}(t) = \mathbf{O}(t - 1) \tag{14}$$

We use the observation of the previous time slot as the current input state of the agent, which is fundamentally different from the spectrum allocation strategy where the communication demand

information of all terminals and all channel state information are known in advance.

The input state $\mathbf{S}(t)$, the agent makes a behavior $\mathbf{a}(t)$ on the learned experience and policy. where the element k in $\mathbf{a}(t)$ is the channel index assigned to the user k, and if no assignment is made to it, its corresponding element is set to 0

3.2 The Reward Function

The reward function refers to the immediate return generated due to the allocation of action $\mathbf{a}(t)$ in state $\mathbf{S}(t)$. Our goal is to maximize the objective function under the tradeoff between the mean QoE of all users and fairness. Therefore, the reward function can be expressed as follows:

$$R_{\mathrm{o}}(t) = F \tag{15}$$

In order to make the QoE values in the optimization problem satisfy the minimum QoE threshold, the sub-reward functions $R_{\mathrm{A}}(t)$ and $R_{\mathrm{V}}(t)$ are proposed.

$$R_{\mathrm{A}}(t) = \begin{cases} \frac{1}{K_1}\sum_{i=1}^{K_1} M_{\mathrm{A}}^{i}(t), if M_{\mathrm{A}}^{i}(t) \geq M_{\mathrm{Amin}} \\ \frac{1}{K_1}\sum_{n=1}^{K_1} \mu_1 * M_{\mathrm{A}}^{i}(t), otherwise \end{cases} \tag{16}$$

$$R_{\mathrm{V}}(t) = \begin{cases} \frac{1}{K_2}\sum_{i=1}^{K_2} M_{\mathrm{V}}^{i}(t), if M_{\mathrm{V}}^{i}(t) \geq M_{\mathrm{Vmin}} \\ \frac{1}{K_2}\sum_{n=1}^{K_2} \mu_2 * M_{\mathrm{V}}^{i}(t), otherwise \end{cases} \tag{17}$$

We define sub-reward functions $R_{\mathrm{A}}(t)$ and $R_{\mathrm{V}}(t)$ as the mean QoE value of data traffic users and video traffic users when the threshold constraint is met, otherwise they are set as reward values lower than the threshold, where $\mu_1 \in (0,1)$ and $\mu_2 \in (0,1)$ are attenuation factors.

Considering the above objectives, a new reward function is designed as follows.

$$R(t) = (1-\lambda)R_0(t) + \lambda(R_{\mathrm{D}}(t) + R_{\mathrm{V}}(t)) \tag{18}$$

where $\lambda \in (0,1)$ is the same as above, our goal is to find a spectrum allocation strategy π that maximizes the expected cumulative discounted rewards [15].

$$V_\pi = E_\pi[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{S}(t+1), \pi(\mathbf{O})] \tag{19}$$

where $\gamma \in (0,1)$ is the discount factor, and $\pi(\mathbf{O})$ is the strategy in $t+1$ slot when the current observation is $\mathbf{O}(t)$. Therefore, the optimal policy π^* can be expressed as:

$$\begin{aligned} \pi^* &= \arg\max_{\pi} V_\pi \\ &= \arg\max_{\pi} E_\pi[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{S}(t+1), \pi(\mathbf{O}))] \end{aligned} \tag{20}$$

4 ADQN FRAMEWORK

In network training, each sample has a different impact on the learning process. Specifically, the significant sample means that the neural network can recover more from it and make the agent more effective to correct its behavior [23]. Therefore, this paper proposes a ADQN algorithm, which introduces a parameter to measure the importance of the training samples of the network. And the agent can learn more effectively from some more important empirical samples, thus improving the performance of the algorithm. In this paper, the importance index of each sample can be measured according to temporal difference error (TD error), which can be defined as follows [16]:

$$I = \left| R + \gamma \max_{\mathbf{a}} Q_{\mathrm{target}}(\mathbf{s}', \mathbf{a}'; \theta_{\mathrm{target}}) - Q(\mathbf{s}, \mathbf{a}; \theta) \right| \tag{21}$$

where $Q_{\mathrm{target}}(\mathbf{s}', \mathbf{a}'; \theta_{\mathrm{target}})$ is the Q value of target network output, $Q(\mathbf{s}, \mathbf{a}; \theta)$ is the Q value of main network output, θ_{target} is the network parameter of target net, θ is the network parameter of main net, and $\mathbf{s}'$ is the next state in which action $\mathbf{a}$ is taken at state $\mathbf{s}(t)$. The larger the I, the greater the upside of the network prediction accuracy and the more worthwhile of the empirical sample. However, certain samples with large I are replayed too frequently, which leads to loss of sample diversity and overfitting. Therefore, the idea of simulated annealing algorithm is used as a reference in this paper. We divide the samples in the memory unit equally according to the mini-batch size Z, and the sampling probability P_{TD} of each sample e is as follows.

$$P_{\mathrm{TD}}(e) = \frac{\exp(I(e)/\zeta)}{\sum_{m=1}^{M/Z} \exp(I(m)/\zeta)} \tag{22}$$

where, ζ is the parameter of the Boltzmann model [24].

$$\begin{aligned} \zeta &= \zeta_0 e^{(-vt)} \zeta \geq \widehat{\zeta} \\ \zeta &= \widehat{\zeta} \zeta \leq \widehat{\zeta}, \end{aligned} \tag{23}$$

whereζ_0 is related to the initial temperature, $\widehat{\zeta}$represents the ending condition in the exploration state. v affects the transition from exploration to exploitation, and M represents the memory size. In the initial stage of training influenced by the annealing temperature, agent extracts samples from the memory unit in an almost random manner, but as the number of iterations increases, the probability of extracting significant samples will gradually increase until they are rejected by the memory unit.

We use the DQN architecture described by Mnih et al. [9] with a few modifications. We do not have the set of convolutional layers since the input to the neural network is not an image. The input to the neural network is state representation and there is a separate output unit for each possible action. We considered a three-layer neural network where the first two layers are fully-connected, each consists of 64 rectifier units. The output layer is a fully-connected linear layer with single output for each possible action. In addition, there are two independent networks with the same structure in the proposed ADQN algorithm. For each input $\mathbf{S}(t)$, the corresponding theoretical Q value can be calculated by Bellman equation as follows [15]:

$$Q_{\mathrm{target}}(\mathbf{s}(t), \mathbf{a}(t)) = R(t) + \gamma \max_{\mathbf{a}} Q_{\mathrm{target}}(\mathbf{s}', \mathbf{a}'; \theta_{\mathrm{target}}) \tag{24}$$

The parameters of θ_{target} are obtained by regularly copying the parameters of the main network. And the loss function is defined as the mean square error of the target value and the Q-value, i.e.,

$$L(\mathbf{s}(t); \theta) = \left| Q_{\mathrm{target}}(\mathbf{s}(t), \mathbf{a}(t) - Q(\mathbf{s}(t), \mathbf{a}(t); \theta) \right|^2 \tag{25}$$

The structure of ADQN is shown in Figure 2

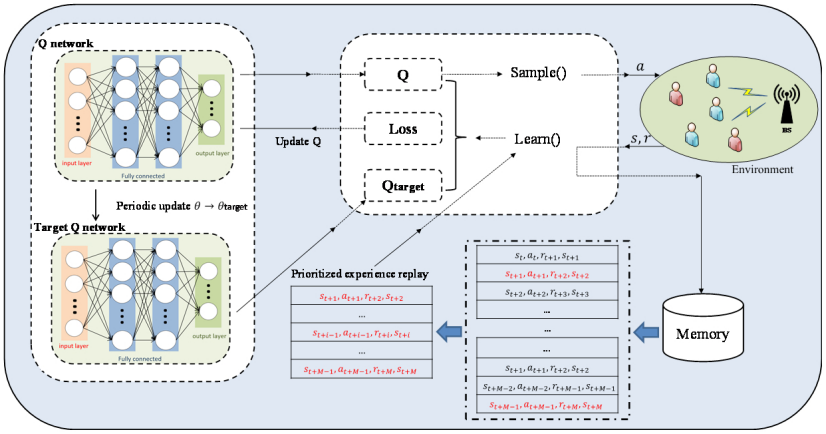

Figure 2: ADQN-based Dynamic Spectrum Allocation System.

ADQN Algorithm for Dynamic Spectrum Allocation

Input: memory size M, mini-batch size Z, discount rate γ, learning rate α, ε in ε-greedy policy, target network update frequency F, and budget T.

Initialize main Q with random weights θ.

Initialize target weights $\theta_{\text{target}} \leftarrow \theta$

For $t = 1$ **to** T

Observe $\mathbf{s}(t), \mathbf{a}(t), R(t)$

If $t \le M$ **Then**

Store transition $\mathbf{s}(t-1), \mathbf{a}(t-1), R(t), \mathbf{s}(t)$ in memory unit.

Else

Remove the oldest experience tuple in memory unit.

End If

Compute TD-error I.

Update sampling probability P_{TD}.

Sample tuples are extracted from the memory unit with P_{TD}.

Choose action $\mathbf{a}(t)$ with ε-greedy policy

Update $\theta_{\text{target}} \leftarrow L(\theta)$.

End For

Algorithm End

5 SIMULATION RESULTS

In this section, numerical experiments were conducted to verify the performance of the proposed dynamic spectrum allocation algorithm. In all simulations, considering the dynamic characteristics of wireless channel, the signal-to-noise ratio at the receiver is assumed to be a random variable with uniform distribution of 15~20dB, $B_{\text{W}} = 0.2$MHz. Using (K, N, K_2) represents the spectrum sharing system with N subchannels and K users (K_2 type II users). The combinatorial space of hyper-parameters is too large for an exhaustive search. We did not perform a systematic grid search owing to the high computational cost. Instead, we have only performed an informal search. The values of all hyper-parameters are provided in Table 1. And all results were obtained based on the deep learning framework in TensorFlow 1.13.1.

Table 1: Hyper-Parameters of ADQN

Hyper-Parameters	Value
Memory size M	1000
Mini-batch size Z	128
Discount rate γ	0.9
Learning rate α	0.005
Target network update frequency	300
Activation function	ReLU
Optimizer	Adam

We consider the scenarios when the number of users is 2, 3, 4 and 5 times the number of subchannels. Figure 3 shows the relationship between the proposed ADQN algorithm and the iteration

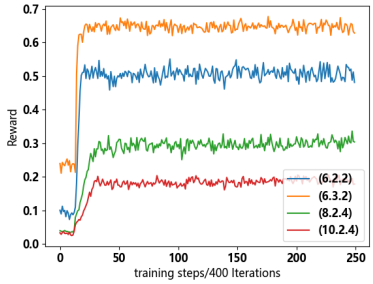

Figure 3: Training Effect Demonstration of ADQN under Different Scenarios.

step under the scenario of different number of users when $\lambda = 0.7$. It can be seen that even if the number of users is more than 5 times the number of subchannels, all reward values converge to stability as the training step increases, this indicates that the ADQN algorithm has excellent convergence. Furthermore, it can be noted that increasing the number of users or decreasing the number of available subchannels will result in a smaller final reward, which is consistent with the definition of the reward in (18).

To better illustrate the performance of ADQN algorithm, we compare it with DQN algorithm and random policy in the same networks and hyperparameters.

We must point out that there is not any coordination and exchange between users in our proposed algorithm, which starts from random exploration. Therefore, as shown in Figure 4 (a), after training, the return value of our algorithm is more than 3 times that of the random policy in (6,2,2). And as shown in Figure 4 (b), when the random allocation algorithm almost fails (the reward is close to 0), the reward of the algorithm can still reach about 0.3. As shown in Figures 4(c) and 4(d), the ADQN algorithm significantly improves the QoE average and fairness performance compared to the random policy. In addition, compared with DQN, the algorithm has better performance (about 4% mean of QoE and 7% fairness index).

As shown in Figure 5, with the increase of the number of users, the time required for calculation increases exponentially with the number of state spaces, and the convergence speed of ADQN algorithm is significantly faster than that of DQN algorithm. In addition, compared with DQN algorithm, the convergence time of the proposed algorithm is reduced by 18% in (8,2,4) and 22% in (10,2,2).

Figure 6 shows the performance comparison of different algorithms under different compromise coefficients λ in (6,2,2). As we can see, the reward value of the learning algorithm decreases as λ increases because the growth rate of the QoE mean is smaller than the reduction rate of fairness. It is important to note that the trend of the reward function varies in other scenarios where the growth rate of the average QoE is faster than the reduction rate of fairness and the reward value of the learning algorithm increases as λincreases, which is determined by the concavity and convexity of the optimization problem. Furthermore, we can see that the ADQN algorithm can obtain approximate solutions under different λ and it still has a slight performance advantage over the DQN algorithm.

6 CONCLUSION

In this paper, we study a dynamic spectrum allocation scheme for multiple heterogeneous users in the presence of resource shortage. By introducing a tradeoff factor λ, we modeled the dynamic spectrum allocation problem as an optimization problem which balances user QoE mean and fairness. Based on this, we propose a ADQN

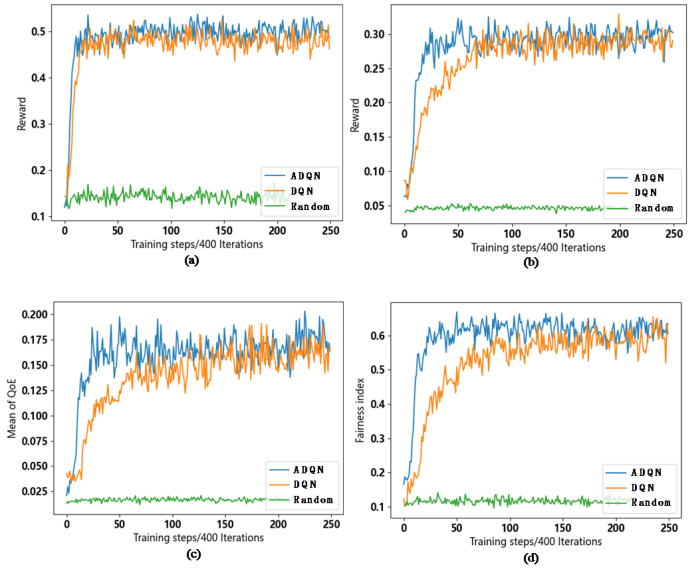

Figure 4: Training Effect Comparison of Different Algorithms. (a) Comparison of Reward Function in (6,2,2). (b) Comparison of Reward Function in (8,2,4). (c) Comparison of QoE Mean in (8,2,4). (d) Comparison of Fairness Index in (8,2,4).

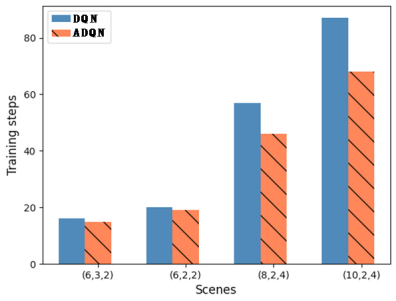

Figure 5: Convergence Speed Comparison of Different Algorithms.

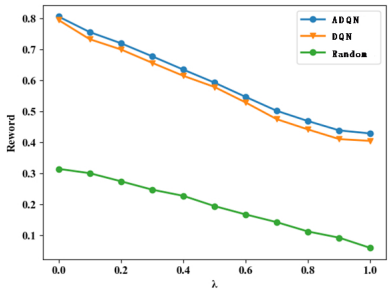

Figure 6: Reward Function Comparison of Different Algorithms Versus λ.

algorithm with prioritized experience playback. Numerical results show that the algorithm is able to converge to different number of users and λ with excellent robustness. In addition, the algorithm has a faster convergence rate and slightly better performance compared to other algorithms.

REFERENCES

[1] Gupta A, Jha R K (2015). A survey of 5G network: Architecture and emerging technologies[J]. IEEE access, 3: 1206-1232.
[2] Singh K D, Hadjadj-Aoul Y, Rubino G (2012). Quality of experience estimation for adaptive HTTP/TCP video streaming using H. 264/AVC[C]//2012 IEEE Consumer Communications and Networking Conference (CCNC). Las Vegas, United States. IEEE, 127-131.
[3] ITU-T SG12 (2007). Definition of Quality of Experience[S], COM12 - LS 62 - E, TD 109rev2 (PLEN/12), Geneva, Jan. 2007.
[4] Hu F, Chen B, Zhu K (2018). Full spectrum sharing in cognitive radio networks toward 5G: A survey[J]. IEEE Access, 6: 15754-15776.
[5] Tsiropoulos G I, Dobre O A, Ahmed M H, *et al.* (2014). Radio resource allocation techniques for efficient spectrum access in cognitive radio networks[J]. IEEE Communications Surveys & Tutorials, 18(1): 824-847.
[6] Ahmad W S H M W, Radzi N A M, Samidi F S, *et al.* (2020). 5G technology: Towards dynamic spectrum sharing using cognitive radio networks[J]. IEEE Access, 8: 14460-14488.
[7] Liu S, Wu J, He J (2021). Dynamic Multichannel Sensing in Cognitive Radio: Hierarchical Reinforcement Learning[J]. IEEE Access, 9: 25473-25481.
[8] Liu X, Sun C, Zhou M, *et al.* (2021). Reinforcement learning-based multislot double-threshold spectrum sensing with Bayesian fusion for industrial big spectrum data[J]. IEEE Transactions on Industrial Informatics, 17(5): 3391-3400.
[9] Mnih V, Kavukcuoglu K, Silver D, *et al.* (2015). Human-level control through deep reinforcement learning[J]. nature, 518(7540): 529-533.
[10] Sutton R S, Barto A G (1998). Introduction to reinforcement learning[M]. Cambridge: MIT press.
[11] Li S, Li O, Liu G, *et al.* (2021). Trajectory Based Prioritized Double Experience Buffer for Sample-Efficient Policy Optimization[J]. IEEE Access, 9: 101424-101432.
[12] Kang C, Rong C, Ren W, *et al.* (2021). Deep Deterministic Policy Gradient Based on Double Network Prioritized Experience Replay[J]. IEEE Access, 9: 60296-60308.
[13] Glatt R, Costa A H R (2017). Policy reuse in deep reinforcement learning[C]// Thirty-First AAAI Conference on Artificial Intelligence. Menlo Park, United States: AAAI, 4929-4930.
[14] Li X, Fang J, Cheng W, *et al.* (2018). Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach[J]. IEEE access, 6: 25463-25473.
[15] Li Y, Zhang W, Wang C X, *et al.* (2020). Deep reinforcement learning for dynamic spectrum sensing and aggregation in multi-channel wireless networks[J]. IEEE Transactions on Cognitive Communications and Networking, 6(2): 464-475.
[16] Shi Z, Xie X, Lu H, *et al.* (2021). Deep-Reinforcement-Learning-Based Spectrum Resource Management for Industrial Internet of Things[J]. IEEE Internet of Things Journal, 8(5): 3476-3489.
[17] Chu M, Li H, Liao X, *et al.* (2019). Reinforcement learning-based multiaccess control and battery prediction with energy harvesting in IoT systems[J]. IEEE Internet of Things Journal, 6(2): 2009-2020.
[18] Zhu J, Song Y, Jiang D, *et al.* (2018). A new deep-Q-learning-based transmission scheduling mechanism for the cognitive Internet of Things[J]. IEEE Internet of Things Journal, 5(4): 2375-2385.
[19] Dobrijevic O, Kassler A J, Skorin-Kapov L, *et al.* (2014). Q-point: Qoe-driven path optimization model for multimedia services[C]//International Conference on Wired/Wireless Internet Communications. Springer, Cham, 134-147.
[20] Zhou L, Wang X, Tu W, *et al.* (2010). Distributed scheduling scheme for video streaming over multi-channel multi-radio multi-hop wireless networks[J]. IEEE Journal on Selected Areas in Communications, 28(3): 409-419.
[21] Hanhart P, Ebrahimi T (2014). Calculation of average coding efficiency based on subjective quality scores[J]. Journal of Visual communication and image representation, 25(3): 555-564.
[22] Khan S, Duhovnikov S, Steinbach E, *et al.* (2007). MOS-based multiuser multi-application cross-layer optimization for mobile multimedia communication[J]. Advances in Multimedia.
[23] Schaul T, Quan J, Antonoglou I, *et al.* (2015). Prioritized experience replay[J]. arXiv preprint arXiv:1511.05952.
[24] Han C, Huo L, Tong X, *et al.* (2020). Spatial anti-jamming scheme for internet of satellites based on the deep reinforcement learning and stackelberg game[J]. IEEE Transactions on Vehicular Technology, 69(5): 5331-5342.