IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 February 2023, accepted 17 April 2023, date of publication 25 April 2023, date of current version 3 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3270316

RESEARCH ARTICLE

Joint DDPG and Unsupervised Learning for Channel Allocation and Power Control in Centralized Wireless Cellular Networks

MING SUN[1], ERZHUANG MEI[1], SHUMEI WANG[2], AND YANHUI JIN[1]
[1]College of Computer and Control Engineering, Qiqihar University, Qiqihar 161006, China
[2]School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China

Corresponding author: Ming Sun (snogisunming@163.com)

This work was supported in part by the Joint Guiding Project of Natural Science Foundation of Heilongjiang Province under Grant LH2019F038, in part by the Scientific Research Foundation of the Education Department of Heilongjiang Province under Grant 145209123, in part by the Humanities and Social Science Fund of Ministry of Education of China under Grant 21YJAZH091, in part by the Philosophy and Social Science Research Project in Jiangsu Universities under Grant 2021SJA1426, and in part by the 14th Five-Year Plan of Education Science in Jiangsu Province under Grant T-c/2021/93.

ABSTRACT In order to solve the resource allocation problem in scenarios of centralized wireless cellular communication with multiple cells, users and channels, a novel resource allocation algorithm based on joint Deep Deterministic Policy Gradient (DDPG) reinforcement learning and unsupervised learning is proposed. Firstly, the proposed algorithm constructs a channel allocation deep neural network based on DDPG to provide an optimized channel allocation scheme. Secondly, the proposed algorithm constructs a power control deep neural network based on unsupervised learning to provide an optimized power control scheme. In order to make the unsupervised learning have perceptions on dynamic wireless environments, the double experience replay is executed to train the channel allocation deep neural network with the DDPG reinforcement learning and the power control deep neural network with the unsupervised learning, respectively. Since the proposed joint algorithm combines the dynamic perception ability of the DDPG reinforcement learning and the continuous optimization ability of unsupervised learning, the energy efficiency can be effectively maximized. Simulation results show that the proposed algorithm outperforms other algorithms in terms of energy efficiency and transmit rate in time-varying dynamic environments. Furthermore, we discuss the implications of our results and possible future research directions. Our work contributes to the advancement of resource allocation techniques in multi-cell cellular networks to meet the increasing demands of modern wireless communication systems.

INDEX TERMS Deep reinforcement learning, DDPG, unsupervised learning, double experience replay, channel allocation, power control, wireless cellular networks.

I. INTRODUCTION

The proliferation of wireless communication devices and data throughput is growing exponentially with the advent of 5G technology. According to the statistics from the China Internet Network Information Center (CNNIC), from December 2018 to December 2022, the number of Chinese netizens increased from 828 million to 1.067 billion. Among them, the number of mobile netizens was 1.065 billion, with an increase of 36.36 million compared with December 2021. The proportion of netizens who use mobile phones to access the Internet is 99.8%. This has brought the issue of limited spectrum resources and increasing communication demands to the forefront [1], [2], [3], [4]. Therefore, improving the utilization of spectrum and power resources has become a key issue to be explored.

Traditional methods for resource allocation in wireless cellular networks mainly include iterative algorithms [5], [6], heuristic algorithms [7], [8], [9], and so on. These methods tend to have relatively high computational burden and

The associate editor coordinating the review of this manuscript and approving it for publication was Tariq Masood.

computation time [10], and are not feasible to handle the time-varying wireless environments in real time. Therefore, these methods are no longer applicable in the new generation of wireless networks.

It is well known that deep reinforcement learning has excellent capabilities in environmental interaction and dynamic perception. In the context of deep reinforcement learning, perception is the ability of an agent to interpret and understand its environment through sensory input, such as images or sensor data. It involves the ability of the agent to recognize patterns, identify objects, and make sense of the sensory information it receives. Perception is a critical aspect of deep reinforcement learning because it enables the agent to accurately perceive and respond to changes in the environment in real time. The agent can also learn by interacting with the dynamic wireless network environments [11], [12]. Currently, the Deep Q Network (DQN), as a research hotspot in deep reinforcement learning, has been widely used for resource allocation in wireless communications. For example, considering the discreteness of the action space in channel allocation, the algorithm [13] used the DQN for the channel allocation in the D2D wireless network. In power allocation, a power control algorithm based on the DQN [11] is used to train micro base stations (MBS) so that the MBS can learn the optimal strategy and help cognitive users (CUs) communicate with appropriate power. Based on globally centralized information processing, the algorithm [14] used the DQN to allocate power to maximize the energy efficiency of the system. In [15], the DQN is used for joint channel allocation and power control in D2D communication to maximize the total transmit rate of the system under the premise of ensuring interference. However, since only one state is mapped from the largest Q value by DQN and the action space for DQN increases exponentially with the number of channels, the above algorithms related to DQN cannot be applied to the resource allocation problem in scenarios of centralized wireless cellular communication with multiple cells, users, and channels.

In addition to deep reinforcement learning, deep unsupervised learning has also attracted a lot of attention from researchers because of the following characteristics. On the one hand, labels are not required in deep unsupervised learning, thus avoiding that the performance of deep learning networks is limited by inappropriate labels [4], [16], [17]. On the other hand, deep unsupervised learning can show better optimization performance than deep reinforcement learning [4], [10], [16], [17], [18]. For example, Sun et al. [10] proposed a centralized cellular network resource allocation method based on two-stage deep unsupervised learning, which improves system energy efficiency by separately optimizing channel allocation and power control; Lee et al. [18] proposed a distributed power control framework based on unsupervised learning that maximizes spectrum efficiency or energy efficiency. However, the above unsupervised learning algorithms cannot promptly detect dynamic changes in the mobile communication environment and often require frequent retraining to cope with dynamic changes in the mobile communication environment, resulting in increased prediction time delay and reduced real-time performance.

According to the above, both DQN deep reinforcement learning and unsupervised learning have their advantages and disadvantages in channel and power allocation in scenarios of centralized wireless cellular communication with multiple cells, users, and channels. Since Deep Deterministic Policy Gradient (DDPG) uses a deterministic policy gradient to learn an optimal policy [3], DDPG is more suitable than DQN to solve the resource allocation problem in scenarios of centralized wireless cellular communication with multiple cells, users, and channels. At the same time, unsupervised learning is capable of mining hidden resource utilization patterns and rules from a large amount of unlabeled data, so as to better meet the actual needs of resource allocation [18]. In addition, unsupervised learning is more efficient in optimizing the continuous power control problem [12]. Therefore, organically combining the environmental interaction and perception capabilities of deep reinforcement learning with the optimization capabilities of unsupervised learning is expected to overcome the shortcomings of previous research in [13], [14], [16], and [19] and improve the performance of deep learning in channel and power allocation. Motivated by this, we propose a resource allocation algorithm for centralized cellular networks by combining power control based on unsupervised learning and channel allocation based on DDPG deep reinforcement learning. To make the unsupervised learning have perceptions on dynamic wireless environments, the double experience replay is executed to train the channel allocation deep neural network with the DDPG reinforcement learning and the power control deep neural network with the unsupervised learning, respectively. Specifically, the proposed algorithm in this paper uses the DDPG reinforcement learning for channel allocation, and uses the experience replay information from the DDPG to train the unsupervised learning, thereby achieving power control.

The rest of the paper is as follows. In the next section, a system model in a downlink centralized multi-cell cellular network is formulated. In Section III, the joint DDPG reinforcement learning and unsupervised learning are proposed for resource allocation in the downlink centralized multi-cell cellular network. Simulations and analysis are performed in Section IV. Conclusions are drawn in the last section.

II. SYSTEM MODEL

A multi-cell cellular network is a type of wireless communication system that consists of multiple cells, each of which is served by a base station. A base station, also known as a cell site or cell tower, is a fixed communication station that serves as the central hub for transmitting and receiving wireless signals within a specific geographic area, known as a cell. Base stations are typically equipped with antennas and other communication equipment and are responsible for providing wireless coverage and facilitating communication between mobile devices, such as smartphones, and the core network infrastructure.

A downlink centralized multi-cell cellular network system considered in this paper is shown in Fig. 1.

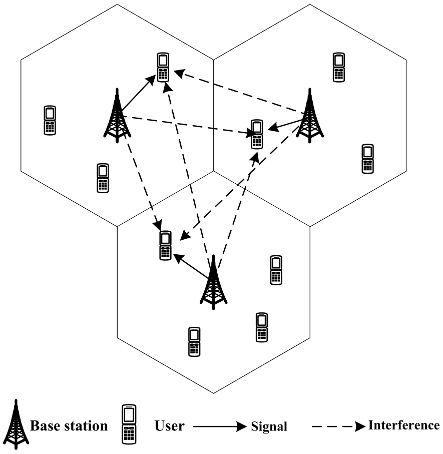


FIGURE 1. **Downlink centralized multi-cell cellular network.**

In the system model shown in Fig. 1, there are K mobile users randomly distributed in M cells. Each cell has one base station. In addition, all base stations and users are single antenna systems. All N orthogonal frequency channels are reused in each cell, and each channel is assigned to a single user. In wireless communications, channel gain plays a critical role in determining the quality of communication links. Channel gain is influenced by various factors such as distance, obstructions, fading, and interference, and is inherently highly dynamic. It is assumed that the channel gain of the system can be expressed by equation (1) [20]:

$$H_{m,k}^{n} = 10^{-(PL_{m,k}+X_{\alpha})/10}\left|h_{m,k}^{n}\right|^{2} \tag{1}$$

where $PL_{m,k}$ represents the log-distance path loss model as base station m communicates with user k; X_α is a random variable subject to a normal distribution with a mean of 0 and a variance of α; $h_{m,k}^n$ represents the Rayleigh fading as base station m communicates with user k over channel n.

When user k communicates over channel n in cell m, the interference from other cells for user k can be expressed as follows [10]:

$$I_{m,k}^{n} = \sum_{\substack{i=1\\ i\neq m}}^{M} \sum_{j=1}^{K} D_{i,j}^{n} p_{i,j}^{n} H_{i,j}^{n} \tag{2}$$

where $D_{m,k}^n$ represents the allocation of channel n in base station m; $D_{m,k}^n = 1$ means that base station m allocates channel n to user k, otherwise, $D_{m,k}^n = 0$; This indicates that channel n is occupied by user k and that communication between base station m and user k is on channel n. The value 1 has no specific meaning here and only indicates whether the channel is occupied by the user. If a user is occupying the channel, the corresponding user index is set to 1, otherwise it is set to 0. $p_{m,k}^n$ denotes the transmit power emitted by base station m when communicating with user k on channel n; $H_{m,k}^n$ represents the channel gain when base station m communicates with user k on channel n.

The total transmit rate R and the total energy efficiency E of the system can be expressed as follows [10]:

$$R = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K} \log_2\left(1 + \frac{D_{m,k}^{n} p_{m,k}^{n} H_{m,k}^{n}}{(N_0 B + I_{m,k}^{n})\Gamma}\right) \tag{3}$$

$$E = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{B\log_2\left(1 + \frac{D_{m,k}^{n} p_{m,k}^{n} H_{m,k}^{n}}{(N_0 B + I_{m,k}^{n})\Gamma}\right)}{10^{6}\cdot p_{m,k}^{n}} \tag{4}$$

where N_0 is the power spectral density of additive white Gaussian noise; B is the bandwidth; $\Gamma = -\ln(5 \cdot BER)/1.6$ where BER is the bit error rate.

In this paper, the optimization problem is formulated as maximizing the expectation of energy efficiency, as shown below:

$$\begin{aligned}
&\max_{\left\{D_{m,k}^{n}, p_{m,k}^{n}\right\}} \quad \mathbb{E}_{H_{m,k}^{n}}[E] \\
&s.t. \quad C_1 : \sum_{k=1}^{K} D_{m,k}^{n} = 1, D_{m,k}^{n} \in \{0,1\} \\
&\qquad C_2 : p_{m,k}^{n} \geq 0 \\
&\qquad C_3 : p_{m}^{n} \geq p_{m,\min}, p_{m}^{n} = \sum\nolimits_{k=1}^{K} D_{m,k}^{n} p_{m,k}^{n} \\
&\qquad C_4 : \sum_{n=1}^{N}\sum_{k=1}^{K} D_{m,k}^{n} p_{m,k}^{n} \leq p_{m,\max}
\end{aligned} \tag{5}$$

where $\mathbb{E}$ is the expectation; E is the energy efficiency; C_1 means that base m allocates channel n to one and only one user in the cell; The C_1 constraint ensures that each channel has only one user, so that the channels are orthogonal to each other and interference can be reduced; C_2 indicates that the transmit power that base station m communicates with user k on channel n is non-negative; The C_2 constraint ensures that the power allocated by the base station to each channel for each user is non-negative because it is consistent with the actual application scenario that the allocated power of the base station cannot be negative. C_3 indicates that the transmit power that base station m communicates on channel n should be greater than or equal to the minimum transmit power;C_4 indicates that the total transmit power that base station m transmits on all channels must not exceed the maximum transmit power $p_{m,\max}$. The C_3 and C_4 constraints are to limit the maximum and minimum transmission power in consideration of user fairness, which can ensure that the power allocated between each channel in a multi-cell system is within a reasonable range, and thus the power allocated to the users can also be guaranteed.

III. PROPOSED JOINT DDPG REINFORCEMENT LEARNING AND UNSUPERVISED LEARNING FOR RESOURCE ALLOCATION

The DDPG reinforcement learning is a type of machine learning where an agent learns to make decisions through trial-and-error interactions with an environment, while unsupervised learning is a type of machine learning where the algorithm learns from unlabeled data. Unsupervised learning tends to have higher optimization efficiency than the DDPG reinforcement learning [12], but has few capabilities of interaction with an environment.

In this section, we present a novel resource allocation algorithm based on joint DDPG reinforcement learning and deep unsupervised learning for the optimization problem (5). Specifically, the proposed algorithm mainly uses DDPG reinforcement learning to obtain the channel allocation scheme and deep unsupervised learning to obtain the channel power control scheme for the optimization problem (5). The proposed algorithm is based on the joint between the DDPG reinforcement learning and the environment and consists of the following parts:

(1) Building a channel allocation deep neural network (DNN) and a power control DNN.

(2) Using the channel allocation DNN with the DDPG reinforcement learning to obtain the channel allocation scheme. During the process, the obtained channel allocation scheme and information of the wireless environment, such as the channel gain information and the interference information, are stored in the experience pool.

(3) The double experience replay is used to obtain different information for the training of the channel allocation DNN and the power control DNN, as shown below:

(3.1) The first experience replay is performed to train the power control DNN with unsupervised learning. Specifically, the channel allocation scheme, channel gain information and interference information from the experience replay are used to train the power control DNN with unsupervised learning, and the optimized channel power control scheme is obtained by the well-trained power control DNN.

(3.2) The second experience replay is performed to train the channel allocation DNN with the DDPG reinforcement learning. Specifically, the channel gain information and interference information from the experience replay are used to train the channel allocation DNN with DDPG reinforcement learning.

In the following, the above parts of the proposed joint DDPG reinforcement learning and unsupervised learning for resource allocation are presented in detail.

A. DDPG REINFORCEMENT LEARNING FOR CHANNEL ALLOCATION

1) BASIC FRAMEWORK OF THE DDPG ALGORITHM

The DDPG reinforcement learning consists of the environment, experience replay, and the DDPG networks. The DDPG networks consist of a main actor network, a target actor network, a main critic network, and a target critic network.

The main actor network is updated by maximizing the function L_A. The function L_A and the policy gradient $\nabla_{\theta^A} L_A$ can be expressed as follows:

$$L_A = \frac{1}{|N_s|}\sum_{s_t\in N_s} Q\left(s_t, \pi\left(s_t|\theta^A\right)|\theta^C\right) \tag{6}$$

$$\nabla_{\theta^A} L_A = \frac{1}{|N_s|}\sum_{s_t\in N_s}\left\{\nabla_{\pi} Q\left(s_t, \pi|\theta^C\right)\nabla_{\theta^A}\pi\left(s_t|\theta^A\right)\right\} \tag{7}$$

where θ^A represents parameters of the main actor network; θ^C stands for parameters of the main critic network; N_s is the set of the mini-batch sampling data in the experience replay, $N_s = \{(s_t, a, r, s_{t+1})\}$, and $|N_s|$ is the size of the set; s_t is the state of the environment at the time t, s_{t+1} is the state of the environment at the time $t+1$ after the state s_t executes the action a; r is the reward received for performing the action a; π is the policy output from the main actor network; Q is the Q value output from the main critic network. θ^A is updated according to equation (8), where η_A is the learning rate.

$$\theta^A = \theta^A + \eta_A \nabla_{\theta^A} L_A \tag{8}$$

The main critic network is updated by minimizing a loss function L_C. The loss function L_C and the gradient $\nabla_{\theta^C} L_C$ can be expressed as follows:

$$\begin{aligned} L_C = &\frac{1}{|N_s|}\sum_{(s_t,a,r,s_{t+1})\in N_s} \\ &\times\left[r+\gamma Q'\left(s_{t+1}, \pi'\left(s_{t+1}|\theta^{A'}\right)|\theta^{C'}\right)\right. \\ &\left.-Q\left(s_t, a|\theta^C\right)\right]^2 \end{aligned} \tag{9}$$

$$\begin{aligned} \nabla_{\theta^C} L_C = &\frac{2}{|N_s|}\sum_{(s_t,a,r,s_{t+1})\in N_s}\left\{r - Q\left(s_t, a|\theta^C\right)\right. \\ &\left.+\gamma Q'\left(s_{t+1}, \pi'\left(s_{t+1}|\theta^{A'}\right)|\theta^{C'}\right)\right\}\nabla_{\theta^C} Q\left(s_t, a|\theta^C\right) \end{aligned} \tag{10}$$

where $\theta^{A'}$ are parameters of the target actor network; $\theta^{C'}$ are parameters of the target critic network; π' is the policy output from the target actor network; Q' is the Q value output from the target critic network; γ is the discount factor, $\gamma \in [0, 1]$. θ^C is updated according to Equation (11), where η_C is the learning rate.

$$\theta^C = \theta^C - \eta_C \nabla_{\theta^C} L_C \tag{11}$$

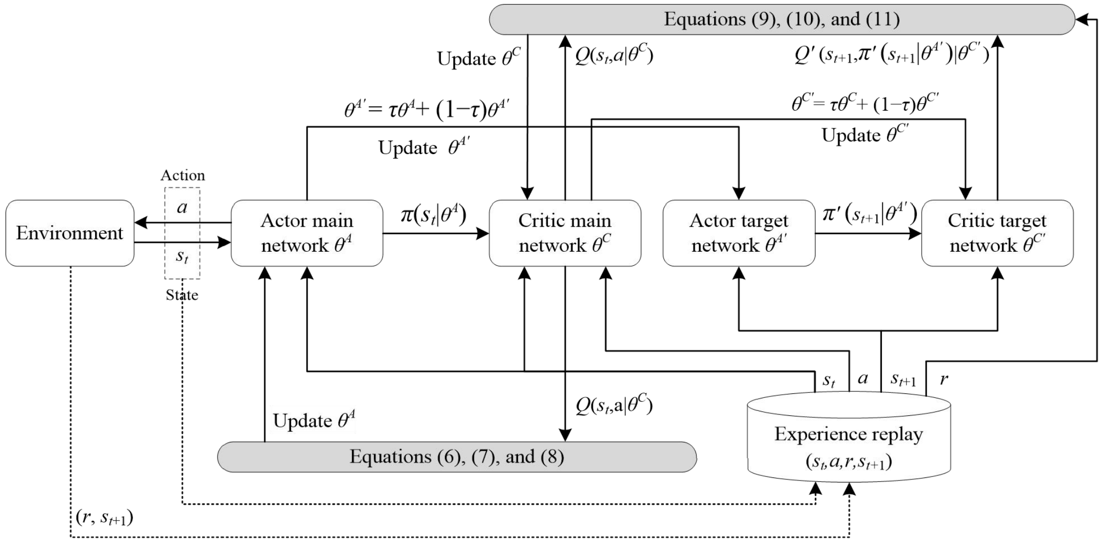

Equations (9), (10), and (11)
Update θ^C
$Q(s_t,a|\theta^C)$
$Q'(s_{t+1},\pi'(s_{t+1}|\theta^{A'})|\theta^{C'})$
$\theta^{A'}=\tau\theta^A+(1-\tau)\theta^{A'}$
Update $\theta^{A'}$
$\theta^{C'}=\tau\theta^C+(1-\tau)\theta^{C'}$
Update $\theta^{C'}$
Action
a
Environment
s_t
State
Actor main network θ^A
$\pi(s_t|\theta^A)$
Critic main network θ^C
Actor target network $\theta^{A'}$
$\pi'(s_{t+1}|\theta^{A'})$
Critic target network $\theta^{C'}$
s_t
a
s_{t+1}
r
Update θ^A
$Q(s_t,a|\theta^C)$
Equations (6), (7), and (8)
Experience replay
(s_t,a,r,s_{t+1})
(r, s_{t+1})

FIGURE 2. The DDPG reinforcement learning framework for channel allocation.

The target actor network and the target critic network are updated by the soft update method shown in equation (12):

$$\begin{cases}\theta^{A'} = \tau\theta^A + (1-\tau)\theta^{A'} \\ \theta^{C'} = \tau\theta^C + (1-\tau)\theta^{C'}\end{cases} \tag{12}$$

where $\tau \ll 1$.

The DDPG reinforcement learning framework for channel allocation is shown in Fig. 2.

2) DEFINITIONS OF STATE, ACTION, AND REWARD

In the DDPG reinforcement learning for channel allocation, the definitions of state s_t, action a, and reward r are described as follows:

(1) State s_t: $s_t = \{H, \hat{H}, \hat{I}\}_t$. Namely, the channel gain H, the normalized interference $\hat{I}$, and the normalized channel gain $\hat{H}$ at time t are taken as the state s_t. $\hat{H}$ and $\hat{I}$ are shown in equations (13) and (14).

$$\hat{H} = \frac{\log_{10}(\tilde{H}) - \min\left[\log_{10}(\tilde{H})\right]}{\max\left[\log_{10}(\tilde{H})\right] - \min\left[\log_{10}(\tilde{H})\right]} \tag{13}$$

$$\hat{I} = \frac{\log_{10}(\tilde{I}) - \min\left[\log_{10}(\tilde{I})\right]}{\max\left[\log_{10}(\tilde{I})\right] - \min\left[\log_{10}(\tilde{I})\right]} \tag{14}$$

where $\tilde{H}$ and $\tilde{I}$ are flattening vectors of the channel gain H, $H = [H_{m,k}^n]$, and the interference signal I, $I = [I_{m,k}^n]$. Logarithmic normalization is used to compress the range of values and reduce the influence of outliers. Logarithmic normalization transforms the data by taking the logarithm of each value, which can make the range of values more manageable and suitable for further analysis. Additionally, logarithmic normalization can also help to stabilize the variance and improve the normality of the data, which can be beneficial in certain types of statistical analyses.

(2) Action a: The channel allocation scheme is taken as an action, i.e., $a = [D_{m,k}^n]$.

(3) Reward r: The optimization objective of this paper is the expectation of energy efficiency, which is the direct response to the action. Therefore, the reward adopted in this paper is shown as follows:

$$r = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{B\log_2\left(1+\frac{D_{m,k}^n p_{m,k}^n H_{m,k}^n}{(N_0B+I_{m,k}^n)\Gamma}\right)}{10^6 \cdot p_{m,k}^n} \tag{15}$$

where $p_{m,k}^n$ is provided by the power control DNN shown in Section III-B.

3) PROPOSED ACTOR NETWORK

In this paper, the channel allocation scheme is taken as the output policy π of the proposed actor network. Note that the channel allocation scheme is generated for the centralized wireless system including multiple cells, channels, and users. Suppose cell m has k_m users that satisfies $\sum_{m=1}^{M} k_m = K$. Then there will be $\prod_{m=1}^{M}(k_m)^N$ kinds of channel allocation schemes. This means that it is difficult to use DQN with the Q-value mechanism as an actor network because the action space for DQN increases exponentially with the number of channels. Hence, we propose a channel allocation DNN with a softmax output layer to act as an actor network. The structure of the proposed channel allocation DNN is shown in Fig. 3. As shown in Fig. 3, the normalized interference $\hat{I}$ and the normalized channel gain $\hat{H}$ are inputs to the channel allocation DNN. They are first passed through a fully connected (FC) layer and a batch normalization (BN) layer, respectively, and then summed and fed into a linear rectifier unit (ReLU) layer. They are then fed into another FC layer, where noise

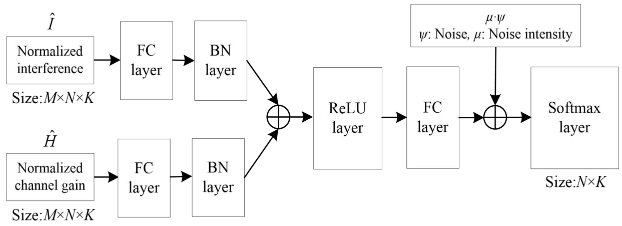


FIGURE 3. Structure of the proposed channel allocation DNN (actor network).

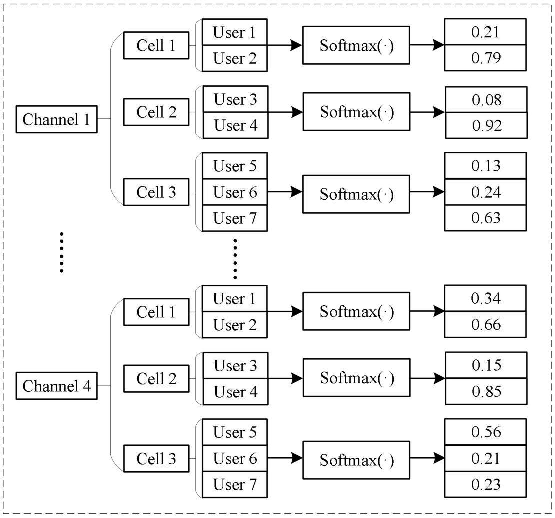


FIGURE 4. Schematic diagram of the softmax layer of the proposed channel allocation DNN (actor network).

is added to them, and then they are fed into a softmax layer. The output is a channel allocation policy of size $N \times K$, which is eventually transformed into a channel allocation policy of size $M \times N \times K$.

Each channel is assigned to the users in the cell according to the probability of each user from the softmax layer. For clarity, a schematic diagram of the softmax layer with 3 cells, 7 users, and 4 channels is shown in Fig. 4. Fig. 4 is explained as follows. In cell 1, users 1 and 2 occupy channel 1 with probabilities of 0.21 and 0.79, respectively, and occupy channel 4 with probabilities of 0.34 and 0.66. In cell 2, users 3 and 4 occupy channel 1 with probabilities of 0.08 and 0.92, respectively, and occupy channel 4 with probabilities of 0.15 and 0.85, respectively. In cell 3, users 5, 6, and 7 occupy channel 1 with probabilities of 0.13, 0.24, and 0.63, respectively, and occupy channel 4 with probabilities of 0.56, 0.21, and 0.23, respectively.

The channel allocation DNN explores and exploits through the product of noise ψ and noise intensity μ. The noise ψ is a random number with a standard normal distribution. And the noise intensity μ is set to be a positive value for exploration, while a zero value for exploitation.

Through the probabilistic occupancy of channels by users, exploration and exploitation based on noise and noise intensity, the proposed channel allocation DNN can prevent the performance degradation caused by the channel space is too large.

B. UNSUPERVISED LEARNING FOR POWER CONTROL

1) PROPOSED POWER CONTROL DNN

This paper proposes a DNN based on unsupervised learning for power control to obtain power $p^n_{m,k}$. The proposed power control DNN is shown in Fig. 5. The proposed power control DNN takes normalized interference, channel allocation scheme, and normalized channel gain as input. After computation through several FC layers, BN layers, and ReLU layers, the power $p^n_{m,k}$ is output through a power constraint processing.

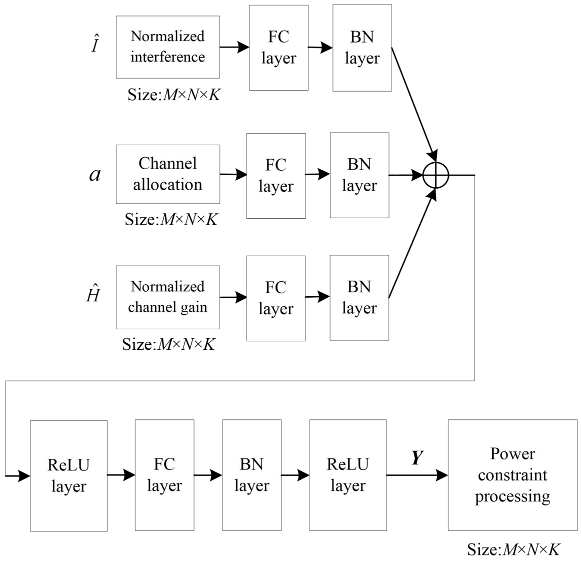


FIGURE 5. Proposed power control DNN.

To enable the power control DNN to meet the requirements of the dynamic wireless environment, the normalized interference $\hat{I}$, the channel allocation a, the normalized channel gain $\hat{H}$, and the normalized channel gain H are extracted from the experience replay of DDPG. Note that the power control DNN shown in Fig. 5 is trained in an unsupervised manner, and no labels are required in the unsupervised learning. Hence, the extracted data is fully used in the following ways. On the one hand, the extracted normalized interference $\hat{I}$, the flattened channel allocation vector a, and the normalized channel gain $\hat{H}$ are used as inputs to the power control DNN. On the other hand, the extracted channel allocation a and the channel gain H are used to construct the loss function of the unsupervised learning to train the power control DNN.

2) POWER CONSTRAINT PROCESSING

The power constraint processing is as follows: Let $\boldsymbol{Y}$ be the output vector of the final ReLU layer in Fig. 5, with size $M \times N$ as shown below:

$$\boldsymbol{Y} = \left[y_1^1, \ldots, y_1^N, \ldots, y_m^n, \ldots, y_M^1, \ldots, y_M^N\right]^{\mathrm{T}} \tag{16}$$

The element y_m^n can be viewed as the transmit power for channel n in cell m. However, it is not guaranteed to meet the transmit power constraints.

To satisfy the transmit power constraints, the power constraint processing first performs min and max operations as shown below:

$$\hat{y}_m^n = \max\left(\min\left(y_m^n, p_{m,\max}\right), p_{m,\min}\right) \tag{17}$$

By means of the operations in (17), $p_{m,\max} \geq \hat{y}_m^n \geq p_{m,\min}$ comes into existence.

The operations in (17) are mainly used to determine whether the power of base stations is fully allocated to the channels or not. For this purpose, we use a sum value of (17), i.e. S_m shown in (18), to control the power allocation of the channels.

$$S_m = \sum\nolimits_{n=1}^{N} \hat{y}_m^n \tag{18}$$

Using the sum value S_m, the power allocation of the channels can be expressed as shown in (19).

$$\hat{y}_m^n = \begin{cases} \dfrac{(p_{\max} - N \cdot p_{\min}) \exp(\hat{y}_m^n)}{\sum\limits_{i=1}^{N} \exp(\hat{y}_m^i)} + p_{\min}, \; if \;\; S_m \geq p_{m,\max} \\ \hat{y}_m^n, \; otherwise \end{cases} \tag{19}$$

As shown in equation (19), if the sum value S_m is greater than or equal to the maximum power of the base stations, i.e., $S_m \geq p_{m,\max}$, it satisfies that $\sum_{n=1}^{N} \hat{\hat{y}}_m^n = p_{m,\max}$, which means that the power of base stations is fully allocated to the channels. If the sum value S_m is smaller than the maximum power of the base stations, i.e., $S_m < p_{m,\max}$, it satisfies that $\sum_{n=1}^{N} \hat{\hat{y}}_m^n < p_{m,\max}$, which means that the power of the base stations is partially allocated to the channels.

Finally, the power $p_{m,k}^n$ can be determined as follows:

$$p_{m,k}^n = \begin{cases} \hat{\hat{y}}_m^n, & if \;\; D_{m,k}^n = 1 \\ 0, & if \;\; D_{m,k}^n = 0 \end{cases} \tag{20}$$

where $D_{m,k}^n$ comes from the action a of the DDPG reinforcement learning.

After the power constraint processing, the output power $p_{m,k}^n$ can satisfy the constraints of C$_2$, C$_3$, and C$_4$ in the constraint optimization problem (5). The detailed flowchart described above is shown in Fig. 6.

3) LOSS FUNCTION IN UNSUPERVISED LEARNING

In this paper, the power control DNN is trained using unsupervised learning. Since no labels are required in unsupervised learning, the loss function can be constructed by the optimization objective of the optimization problem (5) shown

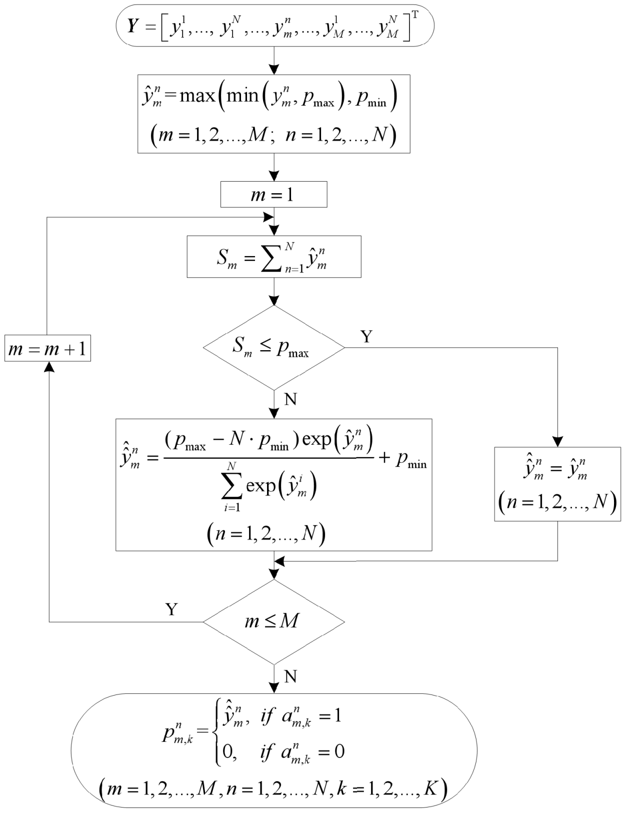

$Y = \left[y_1^1, \ldots, y_1^N, \ldots, y_m^n, \ldots, y_M^1, \ldots, y_M^N\right]^{\mathrm{T}}$
$\hat{y}_m^n = \max\left(\min\left(y_m^n, p_{\max}\right), p_{\min}\right)$
$(m = 1, 2, \ldots, M;\; n = 1, 2, \ldots, N)$
$m = 1$
$S_m = \sum_{n=1}^{N} \hat{y}_m^n$
$S_m \leq p_{\max}$
Y
N
$m = m + 1$
$\hat{\hat{y}}_m^n = \frac{(p_{\max} - N \cdot p_{\min}) \exp(\hat{y}_m^n)}{\sum_{i=1}^{N} \exp(\hat{y}_m^i)} + p_{\min}$
$(n = 1, 2, \ldots, N)$
$\hat{\hat{y}}_m^n = \hat{y}_m^n$
$(n = 1, 2, \ldots, N)$
$m \leq M$
Y
N
$p_{m,k}^n = \begin{cases} \hat{\hat{y}}_m^n, & if\; a_{m,k}^n = 1 \\ 0, & if\; a_{m,k}^n = 0 \end{cases}$
$(m = 1, 2, \ldots, M, n = 1, 2, \ldots, N, k = 1, 2, \ldots, K)$

FIGURE 6. Flowchart of power constraint processing.

in equation (21):

$$L_{PC} = -\mathbb{E}_{\{H, a|H, a \in N_s\}} \times \left[\sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{B \log_2\left(1 + \frac{D_{m,k}^n p_{m,k}^n H_{m,k}^n}{(N_0 B + I_{m,k}^n)\Gamma}\right)}{10^6 \cdot p_{m,k}^n}\right] \tag{21}$$

where $\mathbb{E}$ is the expectation; $p_{m,k}^n$ is obtained from the power constraint processing.

C. TRAINING WITH INFORMATION FROM DOUBLE EXPERIENCE REPLAY

To make the unsupervised learning have perceptions on the dynamic wireless environments, information from the experience pool of reinforcement learning is used not only to train the channel allocation DNN with DDPG reinforcement learning, but also to train the power control DNN with unsupervised learning. Furthermore, to reduce the correlations between the channel allocation DNN and the power control DNN, we use a double experience replay to obtain different information to train the channel allocation DNN and the power control DNN, respectively. The double experience replay involves performing two separate experience replays. Detailed steps for training the channel allocation DNN and the power control DNN with information from the double experience replay are provided below:

Step 1. Initialize the cellular network system environment, the main actor network, the target actor network, the main critic network, the target critic network, the power control DNN, the total episodes (ϕ), the total iterations per episode (κ), the episode recorder ($i = 1$), and the iteration recorder ($j = 1$).

Step 2. Obtain the initial interference I of the cellular network and the state $s_t = \{H, \hat{H}, \hat{I}\}_t$ at time t.

Step 3. Use the main actor network to generate the channel allocation scheme a, and use the power control DNN to generate power $p^n_{m,k}$. Apply channel and power allocation schemes to the cellular network system environment to obtain the interference I, the reward r, and the state $s_{t+1} = \{H, \hat{H}, \hat{I}\}_{t+1}$ at time $t + 1$.

Step 4. Store (s_t, a, r, s_{t+1}) in the experience pool, and set $s_t = s_{t+1}$.

Step 5. If the experience pool is not full, $i = i + 1$ and return to Step 3; otherwise, go to Step 6.

Step 6. Use equation (12) to perform a soft update of the parameters of the target actor network and the target critic network.

Step 7. Replay the experience and use the replay state $s_t = \{H, \hat{H}, \hat{I}\}_t$ and the channel allocation scheme a to train the power control DNN with unsupervised learning.

Step 8. Replay the experience again, and use the replay states $s_t = \{H, \hat{H}, \hat{I}\}_t$ and $s_{t+1} = \{H, \hat{H}, \hat{I}\}_{t+1}$, the channel allocation scheme a, and the reward r to train the main actor network and the main critic network with DDPG reinforcement learning.

Step 9. $i = i + 1$, and perform the corresponding operation according to the following judgment: if $i \leq \kappa$, return to Step 3; if $i > \kappa$, then $j = j + 1$; if $j \leq \phi$, then $i = 1$ and return to Step 2; if $j > \phi$, the training ends.

IV. SIMULATION AND ANALYSES

In this section, we present simulation results to evaluate the performance of our joint DDPG reinforcement learning and unsupervised learning algorithm in resource allocation for a centralized multi-cell cellular network. We compare the proposed algorithm with other algorithms in various performance indicators such as energy efficiency, transmit rate, and computation time, and evaluate the performance of the algorithm in terms of energy efficiency and transmission rate in time-varying dynamic environments. The parameters of the centralized multi-cell cellular network are shown in Table 1.

The proposed algorithm for the centralized multi-cell resource allocation in this paper consists of two parts, i.e., DDPG reinforcement learning for channel allocation and unsupervised learning for power control. The DDPG reinforcement learning for the channel allocation is simulated by the actor neural network and the critic neural network (i.e., the channel allocation DNN), while the unsupervised learning for the power control is simulated by the power control DNN.

In the simulations, the sizes of FC layers, BN layers, and ReLU layers are set to 50 for both the proposed channel allocation DNN and the power control DNN. For fairness, all neural network algorithms were set to have the same order of magnitude of training parameters, and offline training was performed using NVIDIA GeForce RTX 3070 8G GPU, AMD core (TM) R7-5800H 3.80GHz, and 32G memory. In addition, only the CPU was used for online inference of all deep learning algorithms and comparisons.

TABLE 1. Parameters of the centralized multi-cell cellular network.

System parameters (unit)	Value
Number of base stations	3
Cell radius (m)	200
Carrier frequency (GHz)	2.0
Channel bandwidth (kHz)	180
Noise spectral density (dBm/Hz)	-170
Path loss exponent	3.2
Reference distance (m)	100
Standard deviation of shadow fading α	8
Bit error rate (BER)	10^{-3}

In order to effectively train the neural networks, the learning rate for the power control DNN is set to 0.001, and the learning rate for the channel allocation DNN is set to 0.0003. In the simulations, if the energy efficiency obtained by the neural network does not increase, the training is then terminated. Based on the observations in simulations, we set the training episode to 250. After the neural networks are well trained, the proposed algorithm in this paper is compared with other methods, such as the Artificial Bee Colony (ABC) algorithm [9], the joint DQN and DDPG (DQN+DDPG) algorithm [19], the channel allocation and power control algorithm based on unsupervised learning (Unsupervised channel power control) [10], the centralized random/greedy channel allocation and WMMSE channel power control (Greedy/Random+ WMMSE power control) [6]. At the same time, the channel allocation and power control algorithm (unsupervised channel power control) of unsupervised learning proposed in previous research [10] is also simulated and compared with the joint DDPG reinforcement learning and unsupervised learning resource allocation method proposed in this paper. The artificial bee colony algorithm [9] that is used for the comparisons is an improved version of the bee colony algorithm. The traditional artificial bee colony algorithm has the problem of unbalanced local and global search. The authors improved the traditional bee colony algorithm by introducing six different update rules. As for the DQN+DDPG algorithm [19], the DQN network is used for the channel selection, while the DDPG algorithm is used for the power control. In the greedy/random+WMMSE power control algorithm, the channel allocation is performed by the greedy/random algorithm, while the power control is performed by the WMMSE algorithm. The unsupervised channel power control algorithm [10] uses a channel allocation network based on unsupervised learning to output an optimized channel allocation scheme, while using a previously well-trained power control neural network to adjust the optimized channel power to minimize the negative expectations on energy efficiency. To reduce the impact of errors on the performance of the algorithm, the average of 500 Monte Carlo calculations is used as the result of the comparisons.

In order to verify the feasibility of the proposed algorithm, we conducted simulation experiments from several aspects, such as the number of channels, the transmit power of the base station, the minimum transmit power of the channels, the number of users, and the number of iterations.

A. EFFECTS OF THE NUMBER OF CHANNELS

As the number of channels, N is taken as 4, 6, 8, 10 and 12. The proposed algorithm is compared with other algorithms in terms of energy efficiency, transmission rate and computation time. The results of energy efficiency, transmission rate, and computation time of different algorithms are shown in Figs. 7, 8, and 9.

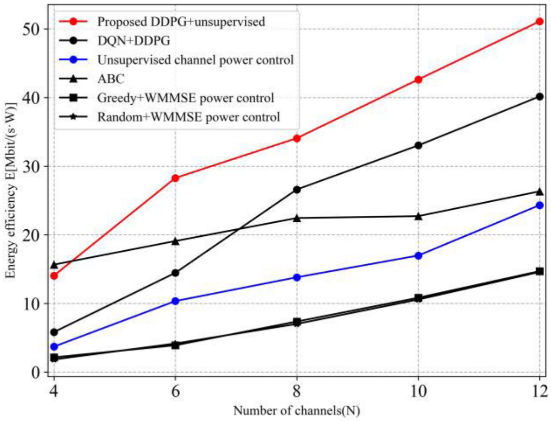


FIGURE 7. Average energy efficiency results obtained by different algorithms.

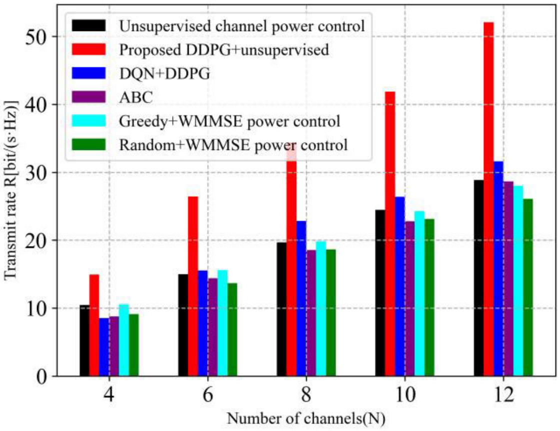


FIGURE 8. Average transmit rate results obtained by different algorithms.

As can be seen from Figs. 7, 8 and 9, the results of the proposed algorithm in terms of both energy efficiency and transmit rate increase as the number of channels in the cell increases. At the same time, as the number of channels varies from 4 to 12, the minimum computation time of the proposed algorithm can reach 10^{-3}(s) order of magnitude, which is not only lower than the computation time of other deep learning algorithms such as DQN+DDPG algorithm [19], Unsupervised channel power control algorithm [10], but also significantly lower than the computation time of other traditional resource allocation algorithms such as ABC [9], Greedy+WMMSE power control algorithm [6], and Random+WMMSE power control algorithm [6]. This indicates that the proposed algorithm not only always achieves higher energy efficiency and transmit rate than other algorithms, but also has lower computation time and delay than other algorithms. It shows that the proposed algorithm has strong optimization ability and effectively improves the utilization rate of wireless resources.

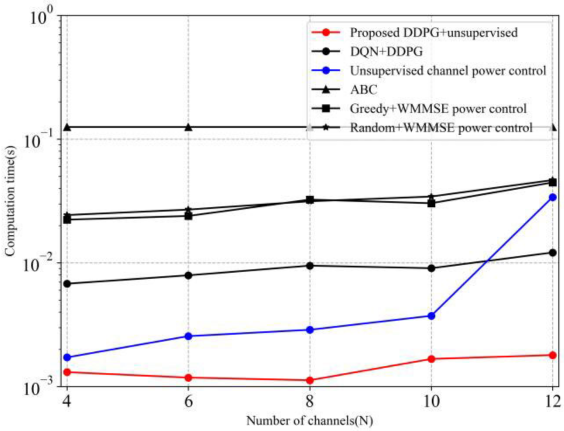


FIGURE 9. Average computation time of different algorithms.

B. EFFECTS OF THE TRANSMIT POWER

Note that both the minimum allocated power of the channels and the maximum transmit power of the base station have impacts on energy efficiency. From the perspective of constraining the minimum allocated power of the channels and the maximum transmit power of the base station, we conducted simulation experiments to comprehensively evaluate the performance of our proposed algorithm. In simulations of constraining the minimum allocated power of channels, the minimum allocated power of channels $p_{m,\min}$ varies from 0.1W to 1.0W, the results of the average energy efficiency obtained by different algorithms are shown in Fig. 10. As can be seen in Fig. 10, the system energy efficiency of all algorithms decreases as the minimum channel power increases. This is because increasing the minimum power of channels easily causes the interference among users to increase and the system transmission rate to decrease, which reduces the overall energy efficiency of the system. However, the proposed algorithm still achieves higher energy efficiency results than other algorithms.

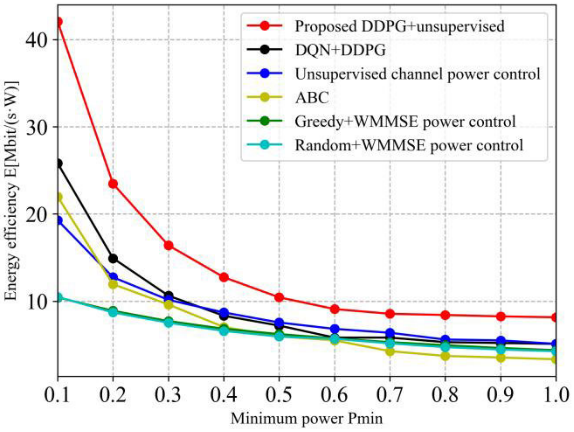


FIGURE 10. Results of energy efficiency obtained by different algorithms as the minimum power varies from 0.1W to 1.0W.

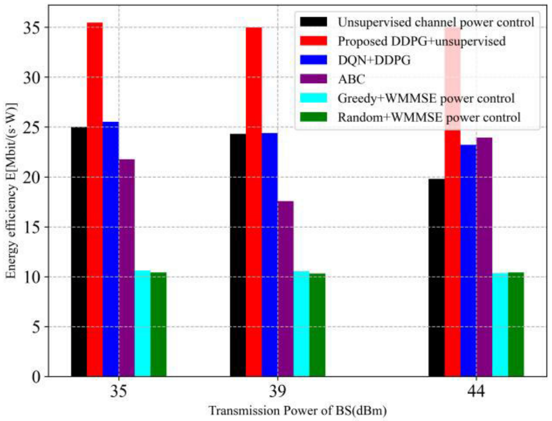


FIGURE 11. Energy efficiency results obtained by different algorithms.

In simulations of limiting the maximum to transmit power of base stations, the maximum transmit power of base stations $p_{m,\max}$ takes 35, 39, and 44 dBm, respectively, results of the average energy efficiency obtained by different algorithms are shown in Fig. 11. As shown in Fig. 11, as the maximum transmit power of the base stations increases, the system energy efficiency obtained by the proposed algorithm remains relatively stable at about 35 [Mbit/(s·W)]. In contrast, the system energy efficiency obtained by other algorithms shows greater variation. This indicates that the proposed algorithm in this paper has a relatively stable energy efficiency performance as the maximum transmit power of the base stations increases. Although the system energy efficiency obtained by the Greedy/Random + WMMSE power control algorithm [6] is also relatively stable, its value always remains at a relatively low level.

C. EFFECTS OF THE NUMBER OF USERS

In the following simulations, we illustrate the impact of the number of users on energy efficiency. When $p_{m,\min} = 0.1$, $p_{m,\max} = 38$, and $N = 10$ remain unchanged, simulation results of energy efficiency are shown in Fig. 12 with the number of users at 10, 15, 20, 25, 30, 35, and 40. From Fig. 12, it can be seen that the system energy efficiency of the proposed algorithm has a slight fluctuation with the increase in the number of users. Except for the cases where the number of users is 15 and 25, the energy efficiency of the proposed algorithm is higher than that of other algorithms. Besides, the simulation results of transmit rate are shown in Fig. 13 with the number of users taking 10, 15, 20, 25, 30, 35, and 40. As shown in Fig. 13, the transmit rate obtained by the proposed joint resource allocation algorithm increases with the number of users and is higher than that of the other algorithms.

Comparisons of the computation time of different algorithms for different numbers of users are shown in Fig. 14. It can be seen that the computation time of the proposed algorithm in this paper is in the order of 10^{-3} seconds, which is about 100 times shorter than the random/greedy channel+WMMSE power control method [6], and about 6 times shorter than the DQN+DDPG algorithm [19] based on deep reinforcement learning. The computation time of the random/greedy channel+WMMSE power control algorithm [6] is caused by multiple iterations and inverse matrix operations in each iteration, which leads to the increase of the computation time cost. On the other hand, due to iterations of heuristic algorithms, the ABC algorithm [9] has a computation time in the order of 10^0 seconds. By combining Figs. 12 and 13, it can be seen that the proposed algorithm can achieve higher energy efficiency and transmit rate by ensuring low computation time. This verifies the effectiveness of the proposed algorithm.

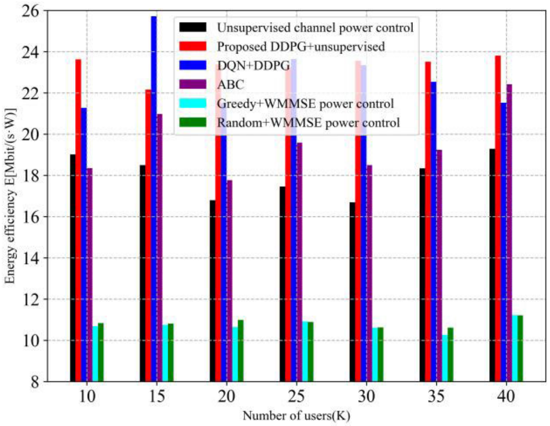


FIGURE 12. Results of energy efficiency obtained by different algorithms with the number of users taking 10, 15, 20, 25, 30, 35, 40.

D. COMPARISONS UNDER TIME-VARYING DYNAMIC ENVIRONMENT

In the following, we consider the performance of the proposed algorithm in a time-varying dynamic environment. The simulation method for the time-varying dynamic environment is described as follows:

TABLE 2. Summary of average energy efficiency and average transmit rate of different algorithms.

Algorithm	Average energy efficiency [Mbit/(s·W)]					Average transmit rate [bit/(s·W)]				
	$N=4$	$N=6$	$N=8$	$N=10$	$N=12$	$N=4$	$N=6$	$N=8$	$N=10$	$N=12$
Proposed DDPG+Unsupervised	14.0420	28.2685	34.0621	42.6126	51.0922	14.9549	26.4229	34.4202	41.8735	52.1083
DQN+DDPG [19]	5.8170	14.4496	26.5932	33.0136	40.1532	8.4944	15.5369	22.8358	26.3651	31.5869
Unsupervised channel power control [10]	3.6972	10.3372	13.8029	16.9711	24.3186	10.4526	14.9611	19.6806	24.4771	28.8458
ABC [9]	15.6485	20.4680	22.2835	20.9209	17.7490	8.7805	14.4824	19.9867	21.4127	30.2052
Greedy +WMMSE power control [6]	1.9830	3.9083	7.1898	10.4506	14.9536	10.6146	15.8013	19.9876	24.0720	27.9614
Random +WMMSE power control [6]	1.8148	4.3837	6.9297	10.0705	14.8314	9.6007	14.7722	17.6334	21.2310	26.5869

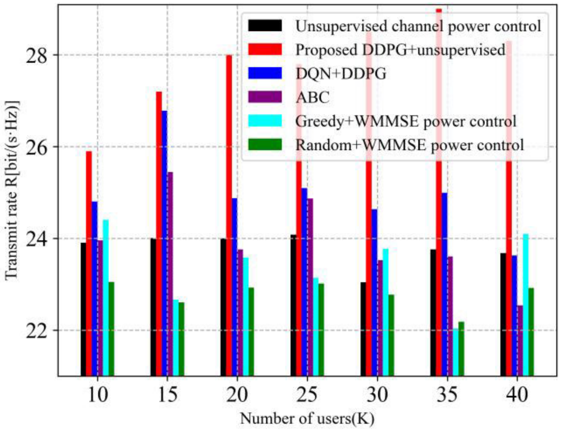


FIGURE 13. Results of transmit rate obtained by different algorithms with the number of users taking 10, 15, 20, 25, 30, 35, 40.

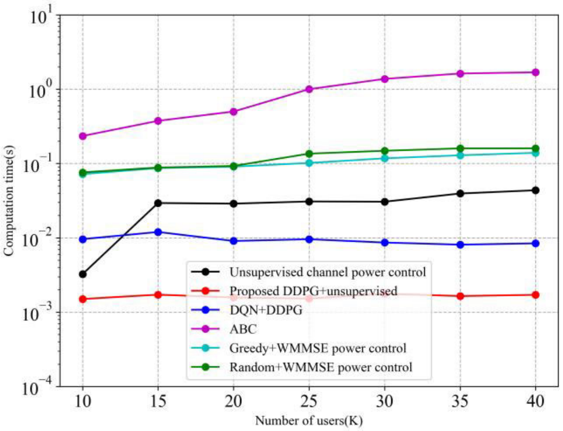


FIGURE 14. Computation time of different algorithms with the number of users taking 10, 15, 20, 25, 30, 35, 40.

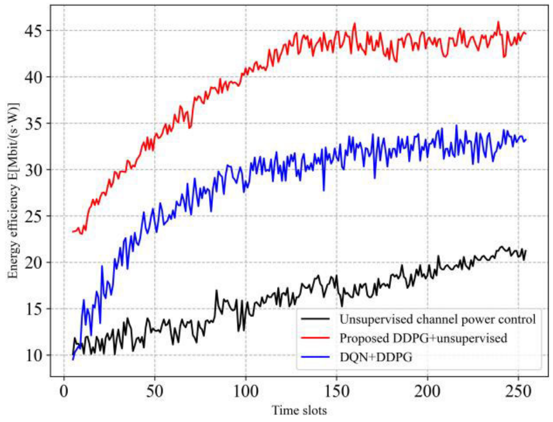


FIGURE 15. Energy efficiency results of different algorithms in the time-varying dynamic wireless environment.

In this dynamic environment, we set up a memory M to store the data in the dynamic environment and use T to represent the time slot index. Firstly, the data collected from the first 50 time slots are filled into the memory M, and T is initialized to zero. Then, the data from 40-time slots are randomly selected from the memory and are used to train the neural network, while the remaining data from 10-time slots are used to test the dynamic performance of the proposed algorithm. After that, the time slot index T is updated by $T = T + 1$. When the next time slot arrives, the memory M is updated on a first-in-first-out basis. In addition, we set the hyperparameters T_1 and T_2 to control the training of the DNN. If $T \leq T_1$, it means that the training of the DNN is not finished; If $T > T_1$, it means that the training of the DNN is finished. At the same time, if $T \leq T_2$, the trained DNN is used for testing; if $T > T_2$, the simulation ends.

With the above simulation method, the results of the energy efficiency and the transmit rate obtained by different algorithms are plotted in Figs. 15 and 16. It can be seen from Figs. 15 and 16 that the proposed algorithm outperforms the DQN+DDPG algorithm and the unsupervised channel power control algorithm in terms of both energy efficiency and transmit rate. The comparison results verify the efficiency of the proposed algorithm in the time-varying dynamic wireless environment.

The average energy efficiency and average transmit rate of different algorithms at $p_{m,\min} = 0.1$ and $p_{m,\max} = 38$ are summarized in Table 2. As shown in Table 2, it is evident

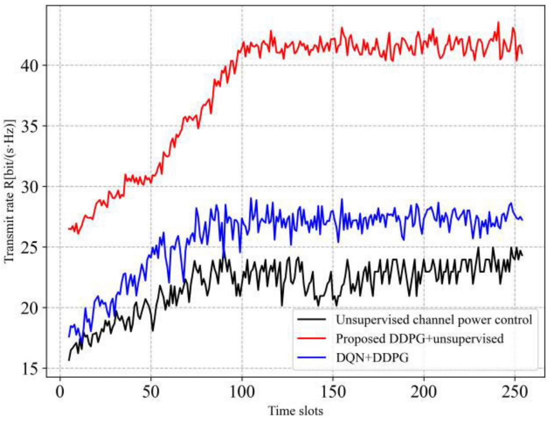


FIGURE 16. Transmit rate results of different algorithms in the time-varying dynamic wireless environment.

that our proposed joint DDPG reinforcement learning and unsupervised learning outperforms other algorithms in terms of energy efficiency and transmit rate, which is consistent with the optimization goal of maximizing energy efficiency.

Through the results of the above simulation experiments, we verified the effectiveness of the proposed algorithm from various aspects, including the number of channels, the number of users, the minimum transmit power of channels, the maximum transmit power of base stations, and the optimization performance in time-varying dynamic environments. This proves the superiority of our proposed joint DDPG reinforcement learning and unsupervised learning for resource allocation problems in scenarios of centralized wireless cellular communication with multiple cells, users, and channels.

V. CONCLUSION

In response to the wireless communication requirements posed by the new era of smart and green networks, this paper proposes a resource allocation algorithm for multi-cell cellular networks based on the joint of deep reinforcement learning and deep unsupervised learning. By making full use of information from the double experience replay to train the channel allocation and the power control DNN, the proposed algorithm fully exploits the dynamic perception of the DDPG reinforcement learning in unknown environments and the performance optimization advantages of unsupervised learning. Our simulation results show that the proposed algorithm is superior to other baseline algorithms. On average, the proposed algorithm achieves up to 432.4% improvement in energy efficiency and up to 75.9% improvement in transmit rate compared to the random/greedy channel allocation +WMMSE power control algorithm. In terms of energy efficiency, it is also better than the ABC, the DQN+DDPG algorithm, and the unsupervised channel power control by up to 74.4%, 64.2%, and 172.2%, respectively. The transmit rate is also up to 78.6%, 64.1%, and 69.2%, respectively. In conclusion, the proposed algorithm demonstrates superior performance in terms of energy efficiency and transmit rate, making it a promising solution for resource allocation in time-varying dynamic environments. Further research can be conducted to explore additional enhancements and optimizations to the proposed algorithm and to investigate its applicability in practical wireless network deployments.

REFERENCES

[1] B. Mao, F. Tang, Y. Kawamoto, and N. Kato, "AI models for green communications towards 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 210–247, 1st Quart., 2022.

[2] M. Zhu, J. Gu, B. Chen, and P. Gu, "Dynamic subcarrier assignment in OFDMA-PONs based on deep reinforcement learning," *IEEE Photon. J.*, vol. 14, no. 2, pp. 1–11, Apr. 2022.

[3] Y. H. Xu, C. C. Yang, M. Hua, and W. Zhou, "Deep deterministic policy gradient (DDPG)-based resource allocation scheme for NOMA vehicular communications," *IEEE Access*, vol. 8, pp. 18797–18807, 2020.

[4] S. Sritharan, H. Weligampola, and H. Gacanin, "A study on deep learning for latency constraint applications in beyond 5G wireless systems," *IEEE Access*, vol. 8, pp. 218037–218061, 2020.

[5] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[6] A. Abrardo, M. Moretti, and F. Saggese, "WMMSE resource allocation for FD-NOMA," *IEEE Commun. Lett.*, vol. 26, no. 11, pp. 2730–2734, Nov. 2022.

[7] M. Sun, Y. Huang, S. Wang, and Y. Xu, "Novel bee colony optimization with update quantities for OFDMA resource allocation," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 889–1020, Jul. 2021.

[8] D. Zhang, D. Zhang, and W. Yan, "A D2D resource allocation mechanism based on pigeon swarm optimization algorithm in heterogeneous networks," *Control Decis.*, vol. 35, no. 12, pp. 2959–2967, 2020.

[9] H. Hakli and M. S. Kiran, "An improved artificial bee colony algorithm for balancing local and global search behaviors in continuous optimization," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 9, pp. 2051–2076, Sep. 2020.

[10] M. Sun, S. Wang, Y. Guo, W. Cao, and Y. Xu, "Deep unsupervised learning based resource allocation method for multi-cell cellular networks," *Control Decis.*, vol. 37, no. 9, pp. 2333–2342, 2022.

[11] X. Zhang and J. Li, "Power control for cognitive users of perception layer in complex industrial CPS based on DQN," *IEEE Access*, vol. 9, pp. 25371–25382, 2021.

[12] M. Sun, Y. Jin, S. Wang, and E. Mei, "Joint deep reinforcement learning and unsupervised learning for channel selection and power control in D2D networks," *Entropy*, vol. 24, no. 12, p. 1722, Nov. 2022.

[13] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, "Deep reinforcement learning for joint channel selection and power control in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1363–1378, Feb. 2021.

[14] F. Zhou, H. Wang, and R. Song, "Downlink power allocation algorithm based on deep reinforcement learning in dense heterogeneous cellular networks," *J. Nanjing Univ. Posts Telecommun.*, vol. 2, pp. 17–24, Jan. 2021.

[15] Y. Du, W. Zhang, S. Wang, J. Xia, and H. A. Mohammad, "Joint resource allocation and mode selection for device-to-device communication underlying cellular networks," *IEEE Access*, vol. 9, pp. 29020–29031, 2021.

[16] S.-M. Tseng, G.-Y. Chen, and H.-C. Chan, "Cross-layer resource management for downlink BF-NOMA-OFDMA video transmission systems and supervised/unsupervised learning based approach," *IEEE Trans. Veh. Technol.*, vol. 71, no. 10, pp. 10744–10753, Oct. 2022.

[17] H. Huang, M. Liu, G. Gui, H. Gacanin, H. Sari, and F. Adachi, "Unsupervised learning-inspired power control methods for energy-efficient wireless networks over fading channels," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 9892–9905, Nov. 2022.

[18] W. Lee and R. Schober, "Deep learning-based resource allocation for device-to-device communication," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5235–5250, Jul. 2022.

[19] Y. S. Nasir and D. Guo, "Deep reinforcement learning for joint spectrum and power allocation in cellular networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Madrid, Spain, Dec. 2021, pp. 1–6.

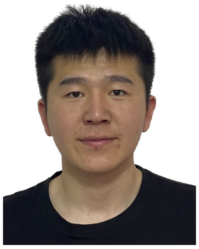
[20] K. I. Ahmed, H. Tabassum, and E. Hossain, "Deep learning for radio resource allocation in multi-cell networks," *IEEE Netw.*, vol. 33, no. 6, pp. 188–195, Nov./Dec. 2019.

MING SUN received the B.S. degree in computer science and technology from Heilongjiang University, Harbin, China, in 2004, the M.S. degree in computer application technology from the Harbin University of Commerce, Harbin, in 2007, and the Ph.D. degree in navigation, guidance and control from Harbin Engineering University, Harbin, in 2010.

He was a Visiting Scholar with the Department of Electrical and Computer Engineering, Baylor University, Waco, TX, USA, in 2015. He is currently a Professor with the College of Computer and Control Engineering, Qiqihar University, Qiqihar, Heilongjiang, China. His current research interests include deep learning, computational intelligence, optimization, and wireless communications.

SHUMEI WANG received the bachelor's degree from the Harbin University of Commerce, in 2004, and the master's degree from Qiqihar University, in 2017. She is currently pursuing the Ph.D. degree with the School of Computer and Information Engineering, Harbin University of Commerce. Her current research interests include deep learning, evolutionary game, and e-commerce.

ERZHUANG MEI received the bachelor's degree from the Anyang Institute of Technology, in 2021. He is currently pursuing the master's degree with the College of Computer and Control Engineering, Qiqihar University, under the tutor Ming Sun. His current research interests include the application of neural networks and the deep learning in the field of communications.



YANHUI JIN received the bachelor's degree from the Anyang Institute of Technology, in 2021. He is currently pursuing the master's degree with the College of Computer and Control Engineering, Qiqihar University, under the tutor Ming Sun. His current research interests include neural networks and deep learning.