IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 20, 2021, accepted February 2, 2021, date of publication February 8, 2021, date of current version February 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3057911

Power Control for Cognitive Users of Perception Layer in Complex Industrial CPS Based on DQN

XIAOMING ZHANG, (Student Member, IEEE), AND JINGZHAO LI
School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan 232000, China

Corresponding author: Jingzhao Li (ljz_aust@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 51874010, and in part by the Key Technology Research and Innovation Team Project under Grant 201950ZX003.

ABSTRACT Wireless communication is a significant auxiliary technology of data transmission for industrial Cyber-Physical system (CPS). While for the complex industrial scenario of coal mine with long and narrow laneway, lifetime of wireless perception nodes is a potential and nonnegligible problem for safety production. In order to deal with this problem, a power control algorithm based on deep Q network (DQN) is adopted to train micro base station (MBS) by two steps so that the MBS can learn an optimal policy to help the cognitive users (CUs) communicate with a proper transmit power. Firstly, the selection range of transmit power for CUs is calculated by the lower bound of Signal-to-Interference plus-Noise Ratio (SINR) to guarantee the transmission condition of both users. Then, the power control problem is modeled as a Markov Decision Process (MDP) with unknown transition function, where the energy consumption is decreased by giving the upper bound of CUs' SINR or threshold of transmit power in formulation of reward. In modeled MDP, the system state, which collected by primary users (PUs) and fed back to MBS, is reduced dimension by method of principal component analysis and then treated as the input of DQN. After that, DQN is used to train a power control optimal policy by minimizing the loss function. Simulation results demonstrate that the proposed power control algorithm based on DQN has a good performance that the average transition step and energy utility are 3.56 and 1580h, which is better than the existing solutions.

INDEX TERMS Industrial CPS, power control, DQN, energy consumption.

I. INTRODUCTION

The complex industrial CPS is a multi-dimensional complex system that integrates calculation, network, and physical environment in many industrial application fields [1]–[3]. In addition, industrial production has been led into the field of intelligence by CPS thanks to the combination of artificial intelligence technology [4], [5]. As the interface between the physical world and the information world, the perception layer contains a large number of wireless perception nodes, which undertakes the task of data perception and transmission. Generally, energy of the nodes is supplied by external wire or battery. For the transmit power of these battery-supplied nodes, it needs to be restricted to prolong the lifetime [6].

The complex industrial CPS integrates many subsystems and the negative environment is a challenge of data trans-mission. Cognitive radio (CR) technology has been widely used due to its mature technical advantages and unique cognitive ability to the environment [7]–[9]. While the spectrum resource of either daily life or industry production is in shortage. Consequently, we apply the CR technology by method of resource sharing to improve the performance of wireless communication so that the barrier of isolated information can be solved and achieve the data integration of subsystems. Nevertheless, co-channel interference is accompanied by resource sharing [10], [11].

Many existing works have studied the optimization of power control and energy saving for battery-supplied wireless users to obtain a good performance [12]–[15]. The authors in [16] optimized the power control problem of CR networks by a non-cooperative game based on sigmoid function. Ramamonjison *et al.* aimed at energy-efficient problem of power control method for CUs in a two-tier cellular network. Then, PUs are assumed to work in time-slotted manner and retain the same activity during the entire frame duration [17].

The associate editor coordinating the review of this manuscript and approving it for publication was Liang-Bi Chen.

The work in [18] investigated the maximization of energy efficiency for the cognitive femto users by optimizing power control scheme in 5G communications. While they are not suitable for industrial applications because of the non-industrial simulation environment. CUI *et al.* solved the maximization and fairness of energy efficiency by alternating iterative optimization scheme in a practical power consumption model [19]. A distributed power control mechanism was proposed for the energy harvesting CR network in [20], where transmit power of nodes was decided dynamically by themselves based on several parameters. In [21], fixed and dynamic power control schemes are proposed to maximize the rate of a CU and limit the interference to PU according to PU traffic and the temporal correlation of channels. The authors in [22] proposed a distributed power control method for the laneway of coal mine, in which the multi-sink nodes are performed as cluster heads, meanwhile, the optimal transmission range and power are allocated to each sink nodes combined with scoping routing algorithm, while the nodes transmit with multiple hops which does not benefit to the energy efficiency.

Reinforcement learning based method has been applied in the form of Q learning, deep reinforcement learning (DRL), and DQN, etc. in many applications [23], [24]. In [25], a deep reinforcement learning-based power control method is proposed to adjust transmit power of a CU with assistance of several sensor nodes, where the sensor nodes are used to collect the received signal strength information (RSSI) of the CU at different locations. A self-adaptive Q learning based MAC protocol is proposed in [26] to tweaks the MAC parameters by a trial-and-error process method to limit the energy consumption. Chu *et al.* [27] considered both the sum rate and prediction loss with two steps, a long short-term memory (LSTM) based algorithm was used to predict the states of users' battery, then the access and power control problem was simultaneously solved by the proposed DQN based algorithm. In [28], the power control problem was considered as a non-cooperative game process and users are assumed to be selfish, then a stochastic power adaption combined with conjecture-based Q learning algorithm was developed for multiple agents in a sufficient condition for the existence of Nash equilibrium. However, it may be impracticable that the proposed algorithm requires the channel state information of each communication link are known by all users in the network.

In this article, a power control method is adopted to control the transmit power of CUs for energy saving under the condition that the quality-of-service (QoS) of both PUs and CUs are satisfied. Then, we specify the industrial environment as the coal mine, which is one of the most classic, strict, and complex industrial scenarios. Some key experiments have been implemented in Suntuan coal mine, Huaibei City. The limited space in the coal mine results is unfriendly environment to the wireless communication which is indispensable for the construction of coal mine CPS. Currently, a large number of wireless communication terminals has been spread in the coal mine, and the shortage of spectrum resource is an objective existence. We adopt CR technology to improve the efficiency of spectrum resources. Unfortunately, interference caused by spectrum sharing may seriously affects the safety production of coal mine as well as the users' lifetime. In order to prolong the lifetime of battery-supplied CUs and restrain the interference in the perception layer of coal mine CPS, a DQN based power control algorithm is proposed to optimize the energy consumption. We assume that the positions of PUs are fixed and powered with a stable and persistent electricity supply while CUs are movable and powered by limited batteries. PUs and CUs work in a non-cooperative manner which means PUs and CUs have no knowledge about the transmit power of each other. The DQN based power control method is proposed for CUs and experience replay mechanism is adopted to weaken the time sequence of experiences attribute by random sample. The main contributions of this article are summarized as follows:

- The transmit power of CUs are restricted in a selection range which is calculated by SINR threshold of CUs and PUs, so that the QoS of both PUs and CUs are guaranteed to be satisfied.
- The power control is modeled as an MDP problem where state, action, and reward are given in detail.
- The dimension of state is reduced by principal component analysis (PCA), and the state characteristics are retained in a lower-dimension matrix.

The rest parts of the paper are organized as follows. In section II, the node distribution and communication model in the coal mine is introduced as well as the transmit power of CUs are expressed. Preliminaries including discretization of transmit power and energy utility are presented in section III. While in section IV, the implement of DQN based power control algorithm is demonstrated in detail. Simulation results and the associated analysis of the algorithm are investigated in Section V. Finally, the conclusion of the paper and the direction for future research are presented in Section VI.

II. SYSTEM MODEL

In this section, we first present the node distribution in the coal mine. Then, the communication model in the coal mine is introduced. The transmit power of nodes is formulated after that.

A. NODE DISTRIBUTION IN THE COAL MINE

The laneway in the coal mine is long and narrow, so the architecture of coal mine CPS is mainly constructed by optical fiber and cable. While, wireless transmission performs an indispensable role as an auxiliary communication technology in the coal mine. Recalling that the perception nodes are divided into two types according to the energy supply modes, so we denote the nodes fixed on the laneway wall as PUs and the battery-supplied nodes as CUs. MBS is in charge of data forwarding with both users and connects with core network of coal mine directly. PUs are assume to communicate with fixed transmit power without change. To assist power control

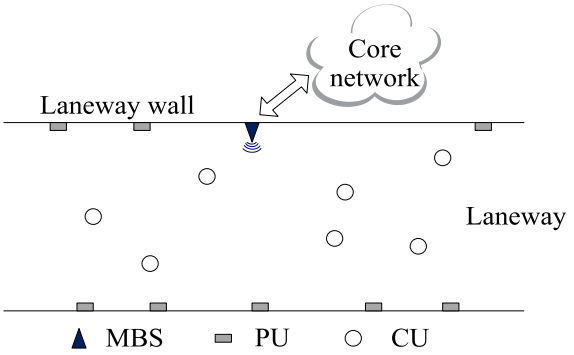

FIGURE 1. Placement sketch of laneway users in the coal mine.

of CUs, all of PUs are employed to measure RSSI, which is the basic function of some nodes such as NB-IoT and LoRa. For instance, the maximum transmission distance of LoRa with transmit power of 20dBm and spreading factor of 12 we have tested is around 200-400m in the coal mine, and the range is different at the different place (300-400m in straight laneway and 200-300 in curved one), which is the universal results of multiple experiments at different places in the coal mine. Thus, the communication scenario is considered to be located in a region of 10*200m which is similar to the laneway in the coal mine. The positions of both users are independent identically distributed (i.i.d.) in the region where CUs and PUs are close enough to satisfy the maximum constraints of communication distance. PUs are on the laneway wall and CUs are located at any position in the region. Ultimately, MBS is located at the middle of laneway wall as illustrated in Figure. 1.

We adopt the point-to-multiple-point (P2MP) as communication manner rather than multi-hop mode for wireless data transmission in underground limited space. The reason is that energy utilization of multi-hop transmission is low and results in the reduction of lifetime especially for the battery-supplied CUs on one hand. On the other hand, the risk of bit error and frame loss are increased because of the multi-hop transmission in a poor communication environment. Thus, each PU and CU communicate with MBS through single hop without relay. Then, the environment parameters of laneway like temperature and pressure are ignored because of little or no influence on the results of our proposed algorithm. For example, the basic version of LoRa E49 can work normally in the temperature from $-40°C$ to $80°C$. PUs transmit signals with sub-channels allocated by MBS and CUs are permitted to reuse these sub-channels. Meanwhile, interference is produced by channel reusing, which means the receiver also receives interference from other transmitter in the same sub-channel when the corresponding transmitter send signals to it. In addition, channel is modeled as the Rayleigh fading channel. Then, each sub-channel allocated to a PU by MBS can be reused by one CU at most and each CU can reuse one sub-channel as well. For the analysis of this article, the other conditions still need to be met as follows

1) Location of PUs and CUs is known, and function of receiving and sending is possessed to both users.

2) The communication link has symmetry, that is, the receiver in a communication link can act as the transmitter, and the transmitter can also act as the receiver.

B. COMMUNICATION MODEL IN THE COAL MINE

In the communication region, the number of CUs is set to be N. The channel is divided into N mutual orthogonal sub-channels and then allocated to PUs, and CUs randomly reuse the sub-channels which occupied by PUs to communicate. Therefore, the number of PUs, which shares sub-channels with CUs, is N as well. There is no interference between PUs because of orthogonal characteristic. In the perception layer of coal mine CPS, j-th PU transmits signals with j-th allocated sub-channel, and i-th CU communicates by reusing j-th sub-channel. Where both $i, j \in X = \{1, 2, \ldots, N\}$. The channel sharing can relieve the scarcity of spectrum resources while results in the interference problem as well. For the interference problem in this paper, parameter of SINR is adopted to measure the quality of communication. The received SINR of i-th CU and j-th PU at MBS are formulated as follows

$$\gamma_{C\_i} = \frac{p_{C\_i} d_{i,B}^{-\alpha} \left|h_{i,B}\right|^2}{\sum_{j=1}^{N} \beta_{i,j} p_{P\_j} d_{j,B}^{-\alpha} \left|h_{j,B}\right|^2 + N_c} \tag{1}$$

$$\gamma_{P\_j} = \frac{p_{P\_j} d_{j,B}^{-\alpha} \left|h_{j,B}\right|^2}{\sum_{i=1}^{N} \beta_{i,j} p_{C\_i} d_{i,B}^{-\alpha} \left|h_{i,B}\right|^2 + N_p} \tag{2}$$

where, p_{C_i} and p_{P_j} denote the transmit power of i-th CU and j-th PU. $d_{i,B}$ and $d_{j,B}$ represent the distance from MBS to users (i-th CU and j-th PU). Then, $h_{i,B}$ and $h_{j,B}$ are the channel response between users and MBS, which follows the i.i.d. complex Gaussian distribution $\mathcal{CN}(0, 1)$ and is known to MBS. The channel is modeled as the Rayleigh fading channel and α denotes the path loss factor where $\alpha = 4$ in view of rough laneway wall and quite serious multi-path fading. N_c and N_p are received additive white Gaussian noise at MBS and follow the distribution $\mathcal{CN}(0, \sigma^2)$. $\beta_{i,j}$ represents the state of channel sharing of PUs and CUs, that is, $\beta_{i,j} = 1$ if i-th CU reuse j-th PU's channel, and $\beta_{i,j} = 0$, otherwise. In order to simplify the interference between CUs and PUs, channel sharing need satisfy the conditions [29] as follows

$$\sum_{i=1}^{N} \beta_{i,j} \leq 1 \quad \text{for } \forall j = 1, 2, \ldots, N \tag{3}$$

$$\sum_{j=1}^{N} \beta_{i,j} \leq 1 \quad \text{for } \forall i = 1, 2, \ldots, N \tag{4}$$

The channels which has been occupied by PUs are randomly allocated to CUs by MBS. Accordingly, each CU cannot reuse multiple channels at the same time and a channel only can be reused by one CU as well. So we can conclude

that there is no interference between CUs because they do not reuse the same channel.

C. TRANSMIT POWER OF CUs

Lifetime is defined as the successive time that CUs can communicate normally in this article. So the transmit power are the significant factor to the lifetime of CUs, which means lower transmit power cannot satisfy the QoS while can contribute to a much longer lifetime on one hand, on the other hand, CUs communicate with larger transmit power will shorten the lifetime. The goal of this article is to prolong the lifetime of CUs by power control on the premise that the communication QoS of both PUs and CUs is satisfied. However, CUs do not know whether the transmit power meets the communication condition by themselves. In this case, the transmit power of CUs are estimated by RSSI. Afterwards, reinforcement learning is used to train a policy for power control and the brief contents of the method is that, in a time slot, CUs broadcast test signals to PUs, then the RSSI of CUs are transmitted to MBS. The neural network is conducted by MBS to obtain the power regulation strategy and feedback to CUs. Finally, CUs adjust the transmit power according to the received strategy. The process of transmitting test signals requires little data overhead of CUs, so energy consumption of this process is ignored.

Before training, the threshold of CUs' transmit power is calculated by the constraint of SINR. When a CU transmits signals to MBS, the PU in the same sub-channel may interfere the signals at MBS. While, communication quality of CUs cannot be satisfied unless the received SINR of CUs' at MBS are not less than the minimum threshold γ_{th}^{C}. So we have the inequality as

$$\gamma_{C\_i} = \frac{p_{C\_i} d_{i,B}^{-\alpha} \left|h_{i,B}\right|^2}{\sum_{j=1}^{N} \beta_{i,j} p_{P\_j} d_{j,B}^{-\alpha} \left|h_{j,B}\right|^2 + N_c} \geq \gamma_{th}^{C} \tag{5}$$

Accordingly, we can obtain the lower bound of CUs' transmit power

$$p_{C\_i} \geq p_{Ci\_\min} = \frac{\gamma_{th}^{C} \left(\sum_{j=1}^{N} \beta_{i,j} p_{P\_j} d_{j,B}^{-\alpha} \left|h_{j,B}\right|^2 + N_c \right)}{d_{i,B}^{-\alpha} \left|h_{i,B}\right|^2} \tag{6}$$

Likewise, PUs will be interfered by the CU in the same sub-channel and the received SINR of PUs at MBS cannot be less than the minimum threshold γ_{th}^{P} as well.

$$\gamma_{P\_j} = \frac{p_{P\_j} d_{j,B}^{-\alpha} \left|h_{j,B}\right|^2}{\sum_{i=1}^{N} \beta_{i,j} p_{C\_i} d_{i,B}^{-\alpha} \left|h_{i,B}\right|^2 + N_p} \geq \gamma_{th}^{P} \tag{7}$$

The upper bound of CUs' transmit power that calculated by SINR threshold of PUs is given as follows.

$$p_{C\_i} \leq p_{Ci\_\max} = \frac{p_{P\_j} d_{j,B}^{-\alpha} \left|h_{j,B}\right|^2 \Big/ \gamma_{th}^{P} - N_p}{\sum_{i=1}^{N} \beta_{i,j} d_{i,B}^{-\alpha} \left|h_{i,B}\right|^2} \tag{8}$$

According to the relevant requirements of safety production in the coal mine, transmit power has an another upper bound $p'_{\max}$, so the maximum transmit power of CUs is shown as follows $p_{\max} = \min(p_{Ci\_\max}, p'_{\max})$.

The transmit power of CUs can be written as

$$p_{C\_i} \in \left[p_{Ci\_\min}, p_{\max}\right] \tag{9}$$

In summary, the QoS requirement of both CUs and PUs can be satisfied if the transmit power of CUs are in the range of Eq. (9). Afterwards, how to reduce the energy consumption of CUs by power control method is the goal of this article.

III. PRELIMINARIES

In this section, the transmit power of CUs is discretized and then the energy utility, which is formulated by synthesizing various factors, is given as an indicator to demonstrate the performance of the power control methods.

A. DISCRETIZATION OF TRANSMIT POWER

Power control policy based on reinforcement learning needs CUs carry out an action to select a transmit power. For simplicity, we need transform the continuous value of transmit power to a discrete selection range. So the discrete process for range of each CU's transmit power is given as

$$P_i(u) = p_{Ci\_\min} + (u-1)\mu_i \tag{10}$$

where, $u = 1, 2, \ldots, U$ and U denotes the largest index of selection range. μ_i is common difference of i-th CU. $P_i(u)$ is a discrete value that i-th CU can select. Eq.(10) represents the selection range is an arithmetic progression for different CUs so that each CU has the corresponding range of transmit power. Here, continuous feasible region is transformed into discrete numerical interval. The selection range of i-th CU's transmit power at time slot t is shown as follows

$$p_{C\_i}(t) \in \mathbf{P}_i \triangleq \{P_i(1), P_i(2), \ldots, P_i(u), \ldots, P_i(U)\} \tag{11}$$

$$\text{s.t. } P_i(1) < P_i(2) < \ldots < P_i(U) \tag{12}$$

$$P_i(U) \leq P_{\max} < P_i(U+1) \tag{13}$$

where, $\mathbf{P}_i$ is the definition of discrete range for i-th CU. Eq.(12) denotes the elements of $\mathbf{P}_i$ is arranged from smallest to largest and Eq.(13) indicates that is not larger than $P_{\max}$ while being the largest in the selection range. After discretization, each CU can select the transmit power in its corresponding $\mathbf{P}_i$. Meanwhile, $\mathbf{P}_i$ starts with and satisfies the condition in Eq.(10). Ultimately, Eq.(13) is a discrete process for Eq.(10) so that CUs can select transmit power in a finite range, which simplifies the complexity of action (transmit power) selection in reinforcement learning.

B. ENERGY UTILITY

Transmit power in the Eq.(11) can meet the QoS of both CUs and PUs. But the goal of this article is energy saving on the premise of satisfied QoS. By the way, the signals with huge number of data (e.g. data of video monitoring) need to be reliably transmitted by cable or optical fiber, while wireless communication is usually used to transmit the data of location information or environment parameters (temperature, humidity, pressure, etc.) in the coal mine. The signals with small number of data can be successfully sent to MBS when QoS of CUs is satisfied. Consequently, the communication performance is guaranteed by lower bound of SINR and the goal of energy saving is completed by reinforcement learning with assistance of upper bound of SINR, which will be explained henceforth. In the formulation of energy utility, we take into account several factors to indicate the performance of the proposed power control algorithm. Energy utility is given as follows

$$\mathrm{E} = \frac{E_0 \sum_{i}^{N} \sum_{j}^{N} \beta_{i,j}}{\bar{p}_C N} \tag{14}$$

where E_0 is initial energy of each CU, $\sum_{i=1}^{N} \sum_{j=1}^{N} \beta_{i,j}$ and $\bar{p}_C$ represent the number and the average transmit power of active CUs, respectively. We can observe that Eq.(14) is expected average lifetime of all CUs, actually. In the condition that initial energy and the number of CUs are fixed, the more number of active CUs and the lower transmit power, the higher the energy utility is. But considering the bad communication environment in the coal mine, few CUs communicate unsuccessfully even with maximum transmit power due to poor channel link state and insurmountable interference, that is these CUs are inactive. The ideal situation after training is that all the CUs in the model are active with a proper transmit power. However, the inactive phenomenon existing in the complex communication environment cannot be ignored.

IV. DQN BASED POWER CONTROL ALGORITHM

In this section, a power control problem for multi-user is modeled as an MDP with unknown transition function first. Then DQN based power control algorithm is carried out for optimal policy.

Reinforcement learning (or Q learning) is a promising decision-making method, which has been applied in many fields. It can be seen as an unsupervised deep learning process in which CUs interact with the environment to derive rewards. In this article, multiple CUs are involved in power control by MBS which is the learning agent and performance of both CUs and PUs need to be guaranteed.

The difficulty of reinforcement learning for multi-user is non-static characteristics of environment. For one user, the other users form part of the environment. More precisely in the paper, one CU adjusts the transmit power in the process of training by interacting with the environment which consists of all CUs. Then, the experience tuples of all CUs

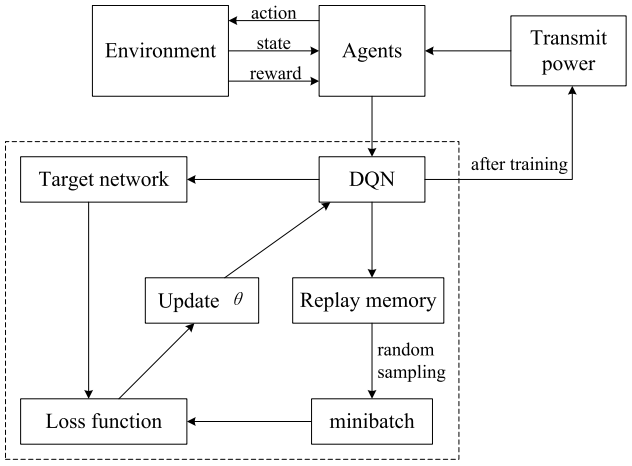

FIGURE 2. Workflow diagram of DQN based method.

are stored in the replay memory which will be introduced henceforth. Replay memory is set for training and its contents are changing with time slot. So the performance of training system is dynamic before DQN is trained well.. In order to overcome such problem, DQN is adopted to deal with the case that the traditional reinforcement learning is difficult to solve the situation of environment change and the large dimension of CUs' state space. The workflow diagram of DQN based method is illustrated in Fig. 2.

A. MDP

The problem of power control is modeled by MDP here. CUs in the current state take selected actions to move into the next state and get an immediate reward. We can observe that the interaction between CUs and environment is a sequence process which consists of state, action, and reward. MDP is an effective method of sequence decision process and thus the problem of this article is modeled by MDP here. MDP is a 4-component tuple $\mathbf{M} = \{\boldsymbol{s}, \boldsymbol{a}, R, T\}$, where $\boldsymbol{s}$ is state space, $\boldsymbol{a}$ is action space, R is reward function and T denotes transition function of state. In this article, the decision maker (MBS) has no way to know the transition relationship between the state of CUs, that is, the transition function T is unknown. Thus, the model-free reinforcement learning is used to train DQN. CUs have to observe the reward in the current state, and then select the corresponding action by a certain method to enter the next state. MDP possesses Markov property. More precisely, i-th CU in the current state $s_i(t) \in s$ selects an action $a_i(t) \in a$ to move into the next state $s_i(t+1)$, while the next state is only related to the current state $s_i(t)$ and the selected action $a_i(t)$, but not to the past states and actions. After moving into a new state, i-th CU will receive an immediate reward $r_i(t)$ which is defined as

$$r_i(t) = \begin{cases} 1, & \text{if} \gamma_{C\_i}(t+1) \geq \gamma_{th}^{C} \text{ and } \gamma_{P\_j}(t+1) \geq \gamma_{th}^{P} \text{ and} \\ & (\gamma_{C\_i}(t+1) \leq c_1 \gamma_{th}^{C} \text{ or } p_i(t+1) \leq c_2 p_{\max}) \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

And immediate reward for all CUs is

$$r(t) = \sum_{i=1}^{N} r_i(t) \tag{16}$$

where c_1 and c_2 are constants. $\gamma_{C\_i}(t+1) \geq \gamma_{th}^{C}$ and $\gamma_{P\_j}(t+1) \geq \gamma_{th}^{P}$ are the premise conditions that PUs and CUs shall communicate normally [16], [24] and [30]. Satisfying the premise conditions merely cannot obtain the immediate reward, which means the purpose of energy saving is not realized at present. By investigating the formulation of SINR, the controllable parameter that affects the energy saving is transmit power after the nodes have been spread in the laneway of coal mine. Therefore, the inequations of $\gamma_{C\_i}(t+1) \leq c_1\gamma_{th}^{C}$ or $p_i(t+1) \leq c_2 p_{\max}$ are the manners to solve the problem of energy saving. The former inequation denotes that the energy consumption is restricted by the upper bound of SINR, and the latter one is for the case that the channel gain of a CU is fine, its SINR may be larger than $c_1\gamma_{th}^{C}$ even with a low transmit power. Whichever of two discussed inequations is true, i-th CU can get an immediate reward based on that the premise conditions are satisfied. The constants and c_2 are set to be and $c_2 = 0.4$ in our experiment. The experiment results suggest that and c_2 can be set to some other similar values as long as doing no harm to the DQN. If and c_2 are too large, it will be meaningless to the power control while too small may result in the case that the algorithm cannot convergent. The constraints are strict enough that the state of i-th CU at time slot t need not only guarantee the QoS of both CUs and PUs, but also reduce the CUs' energy consumption. So CUs cannot obtain the reward unless conforming to the conditions as shown in Eq.(15). The performance which indicated by Eq.(14) is deservedly improved as well.

1) STATE SPACE

The optimization of MDP modeled problem is completed in the process of state transition. The state is input of neural network. In the communication environment of coal mine, we employ PUs to measure RSSI of CUs so that we can obtain a state matrix and each row is a state of one CU, which includes different channel responses between this CU and all PUs. Then RSSI is fed back to MBS as the state input of DQN. The state value is defined as

$$s_{i,j}(t) = p_{i,j}^{(r)}(t) = p_i(t) d_{i,j}^{-\alpha} \left|h_{i,j}\right|^2 \quad \text{for } \forall j \in X \tag{17}$$

where $d_{i,j}$ and $h_{i,j}$ represent the distance and channel response between i-th CU and j-th PU, respectively. The state of i-th CU at time slot t is given as $\boldsymbol{s}_i(t) = \{s_{i,1}(t), s_{i,2}(t), ..., s_{i,N}(t)\}$

Thus, the state space of MDP for the whole CUs is given as follows.

$$\boldsymbol{s}(t) = \{\boldsymbol{s}_i(t) | i \in X\} \tag{18}$$

2) ACTION SPACE

In this article, MBS selects actions for CUs from action space at current state to let CUs move into the next state. Actions are selected through the rule of maximum Q value or randomness with a certain probability and then fed back to CUs by MBS. The action selection of i-th CU at time slot t is $A_i(t) = \mathbf{P}_i$

The action space of MDP model at time slot t is given as follows

$$\boldsymbol{a}(t) = \{A_i(t) | i \in X\} \tag{19}$$

3) REWARD

After learning in each time slot, CUs will move into a new state and get an immediate reward, which has been defined in Eq.(15). Power control of CUs is completed by learning a policy to select an action $\boldsymbol{a}(t)$ based on the current state $\boldsymbol{s}(t)$ in a way of maximizing the discounted cumulative reward which is given as

$$R(t) = \sum_{n=0}^{I'} \tau^n r(t+n) \tag{20}$$

where τ is the discounted factor of immediate reward and $\tau \in [0, 1]$. I' represents the time slot of reaching goal state. The immediate reward obtained in the state which is far away from the current state has little influence on current state, so we use τ to discount the earlier immediate reward.

B. STATE REFORMULATION

In this subsection, we reformulate the state of each time slot into a lower-dimension modality. PCA is a multivariate analysis method for the synthesized simplification of the data with multiple dimensions. It can reduce the dimension of multivariate data on the premise that information loss of data is minimum. Here, we adopt PCA to reduce the dimension of state and the reason is that:

- The state dimension of multiple users is increased with the ascending number of users, which results in the high sparsity of data. Consequently, it is difficult to obtain the character of data.
- Lower dimension of data can reduce the computational complexity of DQN. So the adoption of PCA can promote the performance of our proposed power control method.
- The dimension of state is reduced by PCA while the information of original state can be kept as much as possible by selecting the proper principal component vectors.

In each time slot, the dimension number of each user is N which means the state of each user includes N values. Therefore, all users' states consist of an $N \times N$ matrix $\boldsymbol{\Lambda}$. For simplicity, we omit the parameter t here and $\boldsymbol{\Lambda}$ is written as follows.

$$\boldsymbol{\Lambda} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,N} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,N} \\ \vdots & \vdots & s_{i,j} & \vdots \\ s_{N,1} & s_{N,2} & \cdots & s_{N,N} \end{bmatrix} \tag{21}$$

Then we implement the dimension reduction and the steps are shown as follows.

1) In order to reduce the possibility of overfitting, we centralize the original matrix $\boldsymbol{\Lambda}$ by performing zero-average operation as

$$\hat{s}_{i,j} = s_{i,j} - \frac{1}{N}\sum_{i=1}^{N} s_{i,j} \tag{22}$$

Thus, the centralized state matrix can be rewritten as

$$\hat{\boldsymbol{\Lambda}} = \left[\hat{s}_{i,j}\right]_{N\times N} \tag{23}$$

Then, we can get the states $\hat{\boldsymbol{s}}_j = \left[\hat{s}_{1,j}\ \hat{s}_{2,j}\ \cdots\ \hat{s}_{N,j}\right]^T$ which is measured by PUs and defined as the feature in the PCA.

2) Calculating the covariance matrix of $\hat{\boldsymbol{\Lambda}}$ with feature $\hat{\boldsymbol{s}}_j$ as

$$COV\left(\hat{\boldsymbol{\Lambda}}\right) = \begin{bmatrix} cov\left(\hat{\boldsymbol{s}}_1,\hat{\boldsymbol{s}}_1\right) & cov\left(\hat{\boldsymbol{s}}_1,\hat{\boldsymbol{s}}_2\right) & \cdots & cov\left(\hat{\boldsymbol{s}}_1,\hat{\boldsymbol{s}}_N\right) \\ cov\left(\hat{\boldsymbol{s}}_2,\hat{\boldsymbol{s}}_1\right) & cov\left(\hat{\boldsymbol{s}}_2,\hat{\boldsymbol{s}}_2\right) & \cdots & cov\left(\hat{\boldsymbol{s}}_2,\hat{\boldsymbol{s}}_N\right) \\ \vdots & \vdots & \ddots & \vdots \\ cov\left(\hat{\boldsymbol{s}}_N,\hat{\boldsymbol{s}}_1\right) & cov\left(\hat{\boldsymbol{s}}_N,\hat{\boldsymbol{s}}_2\right) & \cdots & cov\left(\hat{\boldsymbol{s}}_N,\hat{\boldsymbol{s}}_N\right) \end{bmatrix} \tag{24}$$

where $cov(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\left(\boldsymbol{y}_i - \bar{\boldsymbol{y}}\right)$, $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ is the mean value of $\boldsymbol{x}$ and $\boldsymbol{y}$.

3) Calculating the eigenvalue and sorting as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, then obtaining the corresponding unit eigenvector of covariance matrix.

$$\boldsymbol{v}_i = \left[v_{i,1} \quad v_{i,2} \quad \cdots \quad v_{i,N}\right] \quad i\imath X \tag{25}$$

4) Calculating the cumulative value of maximum m eigenvalues as

$$\delta_m = \frac{\sum_{i=1}^{m}\boldsymbol{\Lambda}_i}{\sum_{i=1}^{N}\boldsymbol{\Lambda}_i} \quad \mathrm{m} < \mathrm{N} \tag{26}$$

5) While $\delta_m > 0.85$, we select the first m eigenvectors represented as

$$\boldsymbol{v}(m) = \left[\boldsymbol{v}_1^{\mathrm{T}} \quad \boldsymbol{v}_2^{\mathrm{T}} \quad \cdots \quad \boldsymbol{v}_m^{\mathrm{T}}\right] \tag{27}$$

6) we can get the principal component matrix $\boldsymbol{S}$ by

$$\boldsymbol{S} = \hat{\boldsymbol{\Lambda}} \times \boldsymbol{v}(m) \tag{28}$$

The dimension reduction has been completed yet and the multi-dimension state matrix is transformed into a new matrix which contains most information with a few principal components. The reason we take the first m eigenvectors is that the larger eigenvalue can contribute to the larger variance of principal component matrix after dimension reduction, which means the information of original state can be retained as much as possible. The new matrix $\boldsymbol{S}$ with lower dimension is treated as state of DQN henceforth.

C. STATE-ACTION VALUE FUNCTION

State value function is commonly used to indicate the performance of policy learned by decision maker. After training, MBS is expected to select the action according to the optimal policy and let CUs move into the next state. State value function is implemented to calculate the discounted cumulative reward of all the experienced states which includes the current state in the process of training. Noticing that state value function is changeable due to the change of policy, while the policy lies on the selected action in the certain state. The state value function can be written as

$$V^{\pi}(\boldsymbol{S}) = \mathbb{E}[R(\mathrm{t})|\boldsymbol{S}(t) \in \boldsymbol{S}] \tag{29}$$

Recalling that the transition function is unknown and mode-free reinforcement learning is used to train the optimal policy, the state value function cannot be expressed with transition probability and thus we cannot estimate the policy by state value function. That is to say, it is difficult to obtain the optimal policy $\pi^* = \arg\max_{\pi} V^{\pi}(\boldsymbol{S})$ directly by calculating the discounted cumulative reward while interacting with environment. Consequently, we estimate the discounted cumulative reward by selecting action $\boldsymbol{a}$ in current state $\boldsymbol{S}$. Then the optimal policy is obtained by repeated iteration of state-action value function (Q value function) which is established according to Bellman equation as

$$Q(\boldsymbol{S},\boldsymbol{a}) = r(\boldsymbol{S},\boldsymbol{a}) + \tau \max_{\boldsymbol{S}'\in\boldsymbol{S}} \mathrm{Q}\left(\boldsymbol{S}',\boldsymbol{a}'\right) \tag{30}$$

where $\boldsymbol{S}'$ is the state reached after taking the action $\boldsymbol{a}$ in the state $\boldsymbol{S}$, and $\boldsymbol{a}'$ is the action that will be taken in the state $\boldsymbol{S}'$. $Q(\boldsymbol{S},\boldsymbol{a})$ represents the expected discounted cumulative reward after selecting action $\boldsymbol{a}$ in the state $\boldsymbol{S}$. it is obviously that Q value function is a recursive procedure associated with immediate reward. After the policy is trained, MBS selects the optimal actions for CUs which can maximize the Q value. The model-free problem without transition function that cannot evaluate the performance of policy directly can be solved by Q value function [25], [31].

D. DQN BASED POWER CONTROL

The result of Q learning can be demonstrated by a Q table. After training, decision maker selects an action $\boldsymbol{a}$ that the corresponding Q value is maximum from Q table as the next action in the state $\boldsymbol{S}$. while for this article, the dimension of the state space will be extremely large because of the multiple users and non-static characteristic of environment, and even lead to the result of dimensional disaster. Q learning is difficult to cope with the problem. In such a case, we combine Q learning with DNN and expressed as $Q(\boldsymbol{S},\boldsymbol{a};\theta)$, where θ is the weight of DNN. The purpose of DQN training is to approximate target Q value by updating the weight θ. Experience replay mechanism is used to train the neural network. We store the experience tuple $\mathrm{e}(\mathrm{t}) = \{\boldsymbol{S}(t),\boldsymbol{a}(t),r(t),\boldsymbol{S}(t+1)\}$ into replay memory and indicated as $\mathrm{D}(t) = \{\mathrm{e}(1),\mathrm{e}(2),\ldots,\mathrm{e}(t)\}$. When the number of elements in replay memory is enough, we sample randomly to train DQN.

In the process of training, action selection is implemented by employing ε-greedy strategy to balance exploration and exploitation. The experience tuple is stored sequentially with dique manner, so one tuple in replay memory is relational with surrounding ones, which does not benefit to training. Consequently, we select randomly in replay memory to satisfy independent property of sample and increase its utilization. Then, we calculate the target value of Q function by Bellman equation in the current iteration as

$$\phi(\xi) = r(\xi) + \tau \max_{\boldsymbol{a}'} \mathrm{Q}(\boldsymbol{S}(\xi+1), \boldsymbol{a}'; \theta^{-}) \quad \text{for } \forall \xi \in \varphi_t \tag{31}$$

where θ^{-} is weight value of DQN in current iteration. φ_t denotes the index set of experience tuples in the mini batch. Eq.(31) is the target value calculated by Bellman equation which has been proved to convergent to the optimal Q function [32]. While $Q(\boldsymbol{S}, \boldsymbol{a};\theta)$ is the output of DQN i.e. it is the estimated value. Consequently, the DQN is trained to approximate to the target value so that the difference between target value and $Q(\boldsymbol{S}, \boldsymbol{a};\theta)$ becomes small enough. The target value is used to calculate the loss function as follows

$$L(\theta) = \frac{1}{N_{samp}} \sum_{\xi \in \varphi_t} (\phi(\xi) - Q(\boldsymbol{S}(\xi), \boldsymbol{a}(\xi); \theta))^2 \tag{32}$$

where N_{samp} is the size of mini batch. As shown in Eq.(32), loss function $L(\theta)$ is expected to minimize with iterations, where iterations mean the rounds we implement the simulation experiment. If $L(\theta)$ approximate to 0, the value of $Q(\boldsymbol{S}(\xi), \boldsymbol{a}(\xi); \theta)$ will converge to the target value $\phi(\xi)$. In order to speed up the training of DQN and avoid the problems of local optimization, the probability parameter ε is set to be large and decreases linearly with the number of training iteration. Detailedly, when ε is large, CUs can have exploration as much as possible, i.e. randomly select actions with probability ε. While ε becomes small, CUs have much more exploitation with probability 1-ε to select the actions that can maximize the value of $Q(\boldsymbol{S}, \boldsymbol{a}; \theta^*)$.The DQN based power control algorithm is illustrated in Table 1.

In the architecture of DQN, after implementing experiments many times, we adopt a 5-layer neural network with 3 hidden layers which can perform well while with a lower complexity than that of more hidden layers. Then, rectified linear unit (ReLU) function is used as activation for the first two hidden layers because that ReLU can convergent stably and rapidly without gradient vanishing problem. Activation for the third hidden layer is tanh function which can retain the characteristic of data well. The details of hidden layers are introduced in table 2. In order to avoid the system error of the neural network model being independent of the training samples' characteristics and tending to 0, we set the size of mini batch to be $N_{samp} = 128$ to ensure the generalization ability of the DQN. Specifically, we require that the maximum number of experience tuples in replay memory is $N_R = 500$ of recent iterations. The stochastic gradient descent (SGD) is adopted to update the weight θ. The total number of iterations is $\Gamma = 100000$. We validate the performance of DQN based power control method by implementing 200 independent tests after every 500 iterations.

TABLE 1. Implementation of DQN based power control algorithm.

Algorithm 1. DQN based power control algorithm	
1	Initialize the replay memory $\mathcal{D}$
2	Initialize the DQN $Q(\boldsymbol{S},\boldsymbol{a};\theta)$ with random weight θ^{-}
3	Initialize the selection range of transmit power $\mathbf{P}_i$
4	Initialize the transmit power of CUs randomly
5	Initialize the state $\boldsymbol{s}(1)$ and transform to $\boldsymbol{S}(1)$ by PCA
6	For t=1,2,…, Γ, do
7	Calculate the greedy impact $\varepsilon(t)$
8	$\varepsilon(t) = 0.8(1 - t / (\Gamma + 1))$
9	Choose a random probability p
10	If p$\geq \varepsilon(t)$ then
11	$\boldsymbol{a}(t) = \arg\max_{\boldsymbol{a}} Q(\boldsymbol{S},\boldsymbol{a};\theta)$
12	Else do
13	Select a random action $\boldsymbol{a}(t)$
14	End If
15	With $\boldsymbol{s}(t)$ and $\boldsymbol{a}(t)$ obtain the next state $\boldsymbol{s}(t+1)$ via Eq.(17)
16	Obtain $\boldsymbol{S}$(t+1) by PCA
17	Obtain the immediate reward $r(t)$ via Eq.(15)
18	Store the experience tuple e(t)={$\boldsymbol{S}(t)$, $\boldsymbol{a}(t)$, $r(t)$, $\boldsymbol{S}(t+1)$} in the replay memory $\mathcal{D}$
19	If $t > N_{samp}$, do
20	Get a random mini batch from $\mathcal{D}$ as the training sample
21	Calculate the target $\phi(t)$ via Eq.(31)
22	Train DQN by minimizing the loss function $L(\theta)$
23	End If
24	If $r(t) = N$, do
25	Update the weight with $\theta^* = \arg\min_{\theta} L(\theta)$
26	End If
27	End For

TABLE 2. Details of hidden layers.

Layer	Neurons	Activation
Hidden 1	128	ReLU
Hidden 2	128	ReLU
Hidden 3	256	tanh

After training, MBS can select the optimal transmit power for CUs by the learned policy of power control. Once the transmit power of CUs satisfy the condition as shown in Eq.(15), i.e. the state reaches optimization, CUs will keep the transmit power in the optimal state until the end of data transmission.

V. SIMULATION RESULTS AND ANALYSIS

In this section, a numerical simulation experiment is conducted to demonstrate the performance of the proposed DQN based power control algorithm, where relevant analysis and evaluation are given behind. The number of CUs (or PUs) N is

TABLE 3. Simulation parameters.

Parameter	Value
Maximum/minimum distance between CUs and MBS	100/1 m
Maximum/minimum distance between PUs and MBS	200/1 m
Minimum SINR of CUs / PUs γ_{th}^{C} / γ_{th}^{P}	-10 /0 dB
Transmit power of PUs	100mW
Noise spectral density	-174 dBm/Hz
Initial energy of each node E_0	3 Ah
Discount factor τ	0.99
Size of mini batch N_{samp}	128
Capacity of replay memory N_R	500
Learning rate of DQN	1e-4
Total number of iterations Γ	1e5

set to be $N = 8$, and the rest experiment parameters are listed in Table 3. We use python 3 as the environment to implement experiment. Generally, the path-loss model is applied to both CUs and PUs.

We compared our proposed algorithm with the following solutions.

LQG (Linear Quadratic Gaussian) regulator: the stochastic distribution of the exogenous disturbance is assumed as the Gaussian while the transmit power constraints of CUs are not considered [6].

Q-learning: State and Action are built into a Q-table to store Q values which is hard to handle the case with high-dimension data.

We now investigate the performance of PCA and we randomly select a set of state. The values of state are expanded by a factor of 10^7 to simplify the calculation and illustrated in Table 4. Then, the eigenvalue and the corresponding eigenmatrix are calculated and sorted which is shown in Table 5. Here, we give the maximum 3 eigenvalues and the corresponding eigenmatrices because that the cumulative value of maximum 3 eigenvalues is $3.4316/3.6272 = 0.9461$, which is larger than the threshold 0.85 [33]. Finally, we can obtain 3 principal components as shown in Table 6.

The loss function, SINR, average step, and energy utilization are used to be the performance indicators of the proposed algorithm.

Loss function can indicate the difference between predicted values and target ones as formulated in Eq.(32). The smaller the loss function, the closer the predicted value is to the target value. As shown in Fig. 3, logarithm is implemented to indicate the change of loss. We can conclude that the loss function of our proposed algorithm converges rapidly from value of several hundred to 1 with averaged about 10^4 iterations. Furthermore, the difference becomes smaller as the training goes on, even the difference is small enough, which means the trained DQN is convergent. Though, the curve becomes more unstable as the iteration, the change of peak value is within 1 actually.

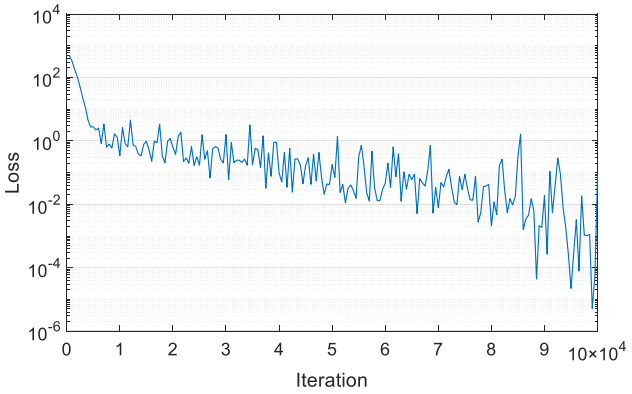

FIGURE 3. Illustration of loss versus iteration.

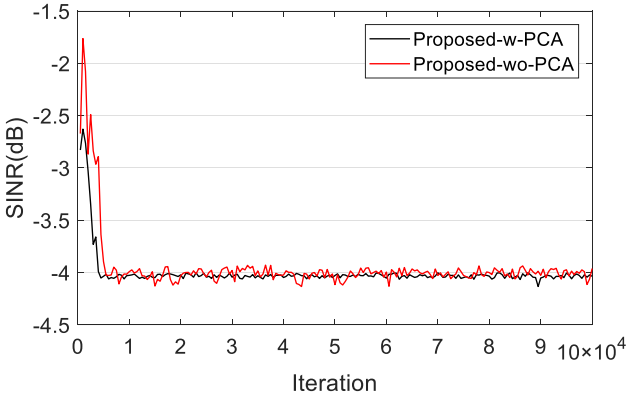

FIGURE 4. SINR of CUs versus iteration.

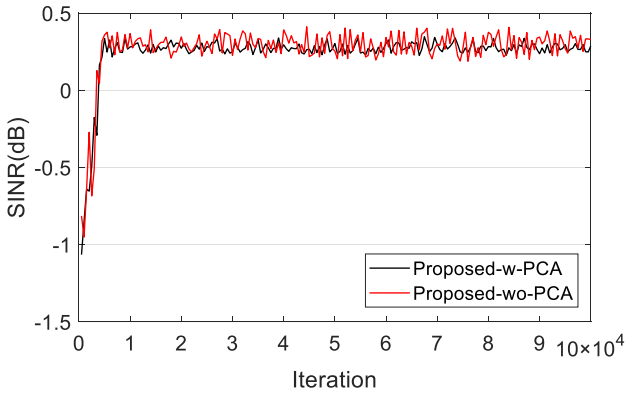

FIGURE 5. SINR of PUs versus iteration.

SINR is the main indicator of communication performance for CUs and PUs. Here, we evaluate the SINR value by illustrating the curves of proposed algorithm with PCA (marked as Proposed-w-PCA) and that without PCA (marked as Proposed-wo-PCA). As shown in Fig. 4 and 5, we can observe that PCA has a certain influence on SINR of both CUs and PUs in the mass. That is, the curves with PCA is some flatter than that without PCA. Then, SINR of both users need to be larger than -10dB, otherwise, the format requirements of modulation and coding scheme (MCS) will not be satisfied for effective communication. So SINR in the

TABLE 4. The state values in a random iteration.

State	PU_1	PU_2	PU_3	PU_4	PU_5	PU_6	PU_7	PU_8
CU_1	0.8315	1.8075	1.1150	0.2192	0.8470	1.3117	1.2887	0.4563
CU_2	0.3589	0.3781	0.8192	0.4776	1.3950	0.7551	1.0676	0.3177
CU_3	0.8053	0.7308	1.1206	1.0794	1.3978	1.1853	1.1915	3.0916
CU_4	0.6499	1.0926	0.1326	0.8122	0.4399	1.1139	0.3393	0.9452
CU_5	1.0678	0.9464	0.1284	1.1757	0.6357	0.4770	0.8608	1.3984
CU_6	0.6700	0.2490	0.4048	1.1459	0.5121	0.7626	0.684	0.5348
CU_7	1.1617	2.9601	1.0484	1.0493	1.2342	0.6299	0.9639	0.3151
CU_8	1.1982	1.5343	0.7873	1.3500	1.1789	0.7661	1.0898	4.1712

TABLE 5. The eigenvalue and the corresponding eigenmatrix.

Eigenvalues	Eigenvectors
$\lambda_1 = 2.2082$	$\boldsymbol{v}_3 = [0.0911 \quad -0.0354 \quad 0.0424 \quad 0.1572 \quad 0.0929 \quad 0.0086 \quad 0.0592 \quad 0.9756]^{\mathrm{T}}$
$\lambda_2 = 0.8805$	$\boldsymbol{v}_2 = [-0.2049 \quad -0.9253 \quad -0.2533 \quad 0.0182 \quad -0.1480 \quad 0.0001 \quad -0.1230 \quad 0.0152]^{\mathrm{T}}$
$\lambda_3 = 0.3429$	$\boldsymbol{v}_3 = [0.2270 \quad 0.2354 \quad -0.5534 \quad 0.3944 \quad -0.4687 \quad -0.2503 \quad -0.3856 \quad 0.0181]^{\mathrm{T}}$

TABLE 6. The principal component of original state.

-0.5809	-0.7939	-0.6455	0.6023	0.6832	0.3030	0.2415	0.1902
-0.7093	0.7338	0.2772	0.4335	0.4045	1.0877	-1.8312	-0.3962
-1.0231	-1.0898	1.7632	-0.5816	0.0043	-0.8623	-1.0330	2.8223

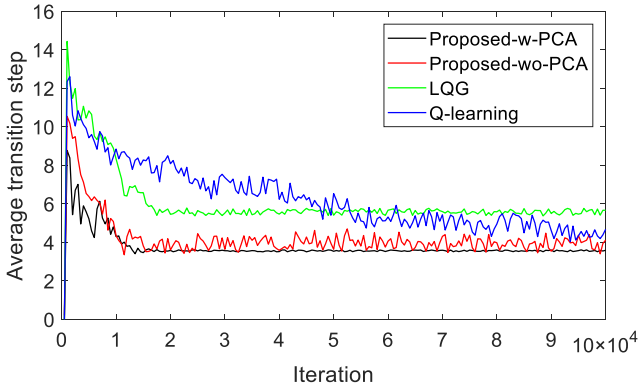

FIGURE 6. Average transition versus iteration.

both figures satisfy the QoS of communication even at the beginning of iteration. Ultimately, SINR value of CUs and PUs reach rapidly to the level around −4dB and 0.2dB in the process of training, which satisfies the condition of energy saving.

As illustrated in Fig. 6, the experiment of average transition step versus iteration for four kind of methods is implemented based on the number of CUs is 8, which means after average transition step of training, CUs and PUs can communicate with satisfied transmit power. Transition steps are the number of steps that DQN needs implement to obtain the stable state in each iteration. Firstly, proposed algorithm has a good performance regardless of dimension reduction by PCA or not. And both of curves are convergent with over 10^4 iterations. While for the proposed algorithm with PCA, it has a fractional transition superiority of 0.34 compared with proposed algorithm without dimension reduction which the former can learn the optimal policy by average 3.56 transition steps and the latter is about 3.90, respectively. Then, curve of the former is much flatter than that of the latter because the PCA performs well on retaining principle components. LQG method has a good performance as well after 2×10^4 iterations with a stable curve. While average transition step of LQG needs at last 5.72 which is some inferior to our proposed method. After that, performance of Q learning is barely satisfactory because it never converges in the whole iterations, though the average transition is less than that of LQG after 5×10^4 iterations. Such an evidence proofs that Q learning is not a reliable method to handle the case with multi-dimension states. In summary, the proposed algorithm which reduces dimensions by PCA performs much better.

In this case, we evaluate the energy utility of aforementioned four methods. Fig. 7 illustrates the energy utility of CUs versus transition step with $N = 8$ in each iteration after training. As mentioned previously, energy utility is

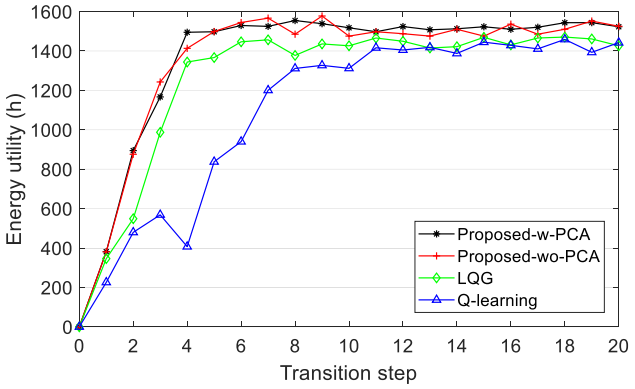

FIGURE 7. Energy utility of CUs versus transition with $N = 8$.

expected average lifetime of all active CUs. The ordinate value in Fig. 7 indicates the continuous time that CUs can communicate normally. We can observe from Fig. 7 that all the curves increase with the transition while the former three curves (i.e. proposed-w-PCA, proposed-wo-PCA, and LQG) increase more fleetly compared with Q learning and then become gentle after several transitions. Our DQN based method can converge to the value around 1520h and the proposed-w-PCA has a good performance with a more stable curve than that of proposed-wo-PCA, though the latter performs well likewise in the mass. The energy utility of LQG regulator is inferior to that of the proposed DQN based method. Q-learning is a method with low computational complexity, while the performance of energy utility is unstable. Though the training of DQN needs a high computational complexity, while after the DQN is trained, the optimal transmit power of CUs can be obtained within several transition steps which involves a much lower complexity. That is, giving a random state, the CU can move into the optimal state with several transition steps. Conclusively, the proposed algorithm can realize a high utility to prolong the lifetime of CUs in coal mine CPS.

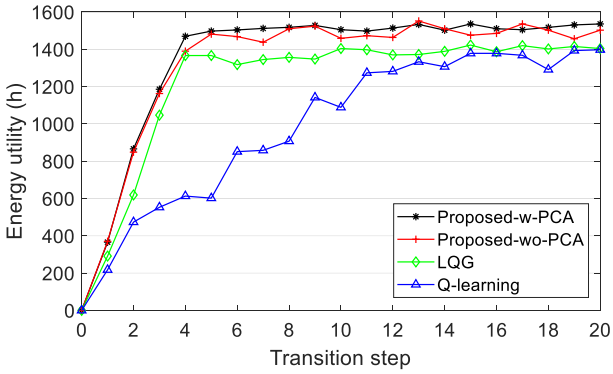

FIGURE 8. Energy utility of CUs versus transition with $N = 16$.

In order to evaluate the stability of the different method, we give an illustration of energy utility with $N = 16$. As shown in Fig. 8, we can observe that the proposed method with or without PCA can converge to a stable level as well as LQG regulator in the case that the number of CUs is increased. Though the utility values of the proposed method and LQG regulator have a little reduction compared with Fig. 7, the stability is much better than Q-learning. The curve of Q-learning illustrates that the high energy utility needs more transition steps than that with $N = 8$. In summary, the proposed method with PCA can perform well with a satisfied stability.

VI. CONCLUSION

In this article, we propose a power control algorithm to improve the energy utility for CUs based on DQN in the perception layer of coal mine CPS, which is a classic case of industrial CPS. We construct the communication scenario by referring to the environment of laneway in the coal mine. The transmit power of CUs is initialized based on the calculated selection range in which the transmit power can satisfy the QoS of PUs and CUs. The power control policy is trained by DQN in a way of reward manner to assist CUs in adjusting the transmit power. Numerical simulation result demonstrates that the proposed algorithm is convergent and has a good performance of energy saving.

As for the future work, joint power control and channel allocation for CUs and PUs are remained to be the further investigation. Meanwhile, the application of the proposed algorithm is planned to adopt in Suntuan coal mine, Huaibei province and many other industrial productions like transport service and construction industry.

REFERENCES

[1] B. Lei, J. Wang, Y. Wu, and X. Li, "From bounded reachability analysis of linear hybrid automata to verification of industrial CPS and IoT," in *Engineering Trustworthy Software Systems—5th International School*. Cham, Switzerland: Springer, 2020, pp. 10–43.

[2] M. T. Higuera-Toledano, J. L. Risco-Martin, P. Arroba, and J. L. Ayala, "Green adaptation of real-time Web services for industrial CPS within a cloud environment," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1249–1256, Jun. 2017.

[3] P. Ren, J. Li, and D. Yang, "The research on model construction and application of coal mine CPS perception and control layer," *Int. J. Embedded Syst.*, vol. 11, no. 4, pp. 483–492, 2019.

[4] S. Luo, Y. Wen, W. Xu, and D. Puthal, "Adaptive task offloading auction for industrial CPS in mobile edge computing," *IEEE Access*, vol. 7, pp. 169055–169065, 2019.

[5] J. Li, D. Yang, and X. Zhang, "Research on modelling and scheduling strategy for mine transportation control system based on CPS," *Int. J. Embedded Syst.*, vol. 11, no. 5, pp. 678–686, 2019.

[6] S. Zhang and X. Zhao, "Distributed power control based on linear quadratic optimal controller for cognitive radio network," *China Commun.*, vol. 15, no. 8, pp. 77–91, Aug. 2018.

[7] G. Kakkavas, K. Tsitseklis, V. Karyotis, and S. Papavassiliou, "A software defined radio cross-layer resource allocation approach for cognitive radio networks: From theory to practice," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 2, pp. 740–755, Jun. 2020.

[8] S. Demirci and D. Gozupek, "Switching cost-aware joint frequency assignment and scheduling for industrial cognitive radio networks," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4365–4377, Jul. 2020.

[9] J. Ren, Y. Zhang, R. Deng, N. Zhang, D. Zhang, and X. Shen, "Joint channel access and sampling rate control in energy harvesting cognitive radio sensor networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 7, no. 1, pp. 149–161, Jan. 2019.

[10] S. Chaudhari and D. Cabric, "QoS aware power allocation and user selection in massive MIMO underlay cognitive radio networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 4, no. 2, pp. 220–231, Jun. 2018.
[11] A. Tsakmalis, S. Chatzinotas, and B. Ottersten, "Interference constraint active learning with uncertain feedback for cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4654–4668, Jul. 2017.
[12] F. Zhou, N. C. Beaulieu, Z. Li, J. Si, and P. Qi, "Energy-efficient optimal power allocation for fading cognitive radio channels: Ergodic capacity, outage capacity, and minimum-rate capacity," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2741–2755, Apr. 2016.
[13] J. Ren, Y. Zhang, N. Zhang, D. Zhang, and X. Shen, "Dynamic channel access to improve energy efficiency in cognitive radio sensor networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3143–3156, May 2016.
[14] C. Yang, W. Lou, Y. Fu, S. Xie, and R. Yu, "On throughput maximization in multichannel cognitive radio networks via generalized access strategy," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1384–1398, Apr. 2016.
[15] X. Yang, M. Sheng, H. Sun, X. Wang, and J. Li, "Spatial throughput analysis and transmission strategy design in energy harvesting cognitive radio networks," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 5938–5951, Dec. 2018.
[16] Y. A. Al-Gumaei, K. A. Noordin, A. M. Mansoor, and K. Dimyati, "Acceleration improvement of a sigmoid power control game algorithm in cognitive radio networks," in *Proc. Int. Conf. Smart Comput. Electron. Enterprise (ICSCEE)*, Jul. 2018, pp. 1–5.
[17] R. Ramamonjison and V. K. Bhargava, "Energy efficiency maximization framework in cognitive downlink two-tier networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1468–1479, Mar. 2015.
[18] H. Park and T. Hwang, "Energy-efficient power control of cognitive femto users for 5G communications," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 772–785, Apr. 2016.
[19] M. Cui, B.-J. Hu, X. Li, H. Chen, S. Hu, and Y. Wang, "Energy-efficient power control algorithms in massive MIMO cognitive radio networks," *IEEE Access*, vol. 5, pp. 1164–1177, 2017.
[20] M. Zareei, C. Vargas-Rosales, R. V. Hernndez, and E. Azpilicueta, "Efficient transmission power control for energy-harvesting cognitive radio sensor network," in *Proc. IEEE 30th Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC Workshops)*, Sep. 2019, pp. 1–5.
[21] S. Chaudhari and D. Cabric, "Power control and frequency band selection policies for underlay MIMO cognitive radio," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 2, pp. 304–317, Jun. 2019.
[22] W. W, X. Xia, M. Wozniak, X. Fan, R. Damaševičius, and Y. Li, "Multi-sink distributed power control algorithm for cyber-physical-systems in coal mine tunnels," *Comput. Netw.*, vol. 161, pp. 210–219, Oct. 2019.
[23] Y. Wu, H. Tan, J. Peng, H. Zhang, and H. He, "Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus," *Appl. Energy*, vol. 247, pp. 454–466, Aug. 2019.
[24] S. Liu, X. Hu, and W. Wang, "Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems," *IEEE Access*, vol. 6, pp. 15733–15742, 2018.
[25] X. Li, J. Fang, W. Cheng, H. Duan, Z. Chen, and H. Li, "Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach," *IEEE Access*, vol. 6, pp. 25463–25473, 2018.
[26] C. Savaglio, P. Pace, G. Aloi, A. Liotta, and G. Fortino, "Lightweight reinforcement learning for energy efficient communications in wireless sensor networks," *IEEE Access*, vol. 7, pp. 29355–29364, 2019.
[27] M. Chu, X. Liao, H. Li, and S. Cui, "Power control in energy harvesting multiple access system with reinforcement learning," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 9175–9186, Oct. 2019.
[28] X. Chen, Z. Zhao, and H. Zhang, "Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 11, pp. 2155–2166, Nov. 2013.
[29] J. Li, X. Zhang, Y. Feng, and K.-C. Li, "A resource allocation mechanism based on weighted efficiency interference-aware for D2D underlaid communication," *Sensors*, vol. 19, no. 14, p. 3194, Jul. 2019.
[30] P. Zhou, Y. Chang, and J. A. Copeland, "Reinforcement learning for repeated power control game in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 54–69, Jan. 2012.
[31] B.-N. Trinh, L. Murphy, and G.-M. Muntean, "A reinforcement learning-based duty cycle adjustment technique in wireless multimedia sensor networks," *IEEE Access*, vol. 8, pp. 58774–58787, 2020.
[32] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 2003.
[33] J. Xi, H. Jiang, B. Chen, and L. Fu, "Infrared multispectral radiation temperature measurement based on PCA-ELM," *J. Shanghai Jiaotong Univ.*, vol. 42, no. 10, pp. 963–968, 2020, doi: 10.16183/j.cnki.jsjtu.2020.027.

XIAOMING ZHANG (Student Member, IEEE) received the bachelor's degree in electrical engineering from the College of Electrical and Information Engineering, Anhui University of Science and Technology (AUST), Huainan, China, in 2015. He is currently pursuing the Ph.D. degree in mining and electromechanical engineering with AUST. His current research interests include heterogeneous cellular networks, the mine Internet of Things, and cyber physical systems.

JINGZHAO LI received the M.A. degree from the China University of Mine and Technology, Xuzhou, China, in 1992, and the Ph.D. degree from the Key Laboratory of Power Electronics and Power Drives, Hefei University of Science and Technology, Hefei, China, in 2003. He is currently a Professor with the College of Electrical and Information Engineering, Anhui University of Science and Technology (AUST), Huainan, China. His research interests include computer control technology, the Internet of Things (IoT), cyber physical systems (CPSs), and embedded systems. His current research interests include design and analyze the mechanism of mine information and physical interface, power control and resource management for the IoT networks, and method of secure interaction in CPS.