Journal of Engineering Science and Technology Review 14 (3) (2021) 103 - 110

JOURNAL OF
Engineering Science and Technology Review

Research Article

www.jestr.org

Deep Reinforcement Learning Strategy for Electric Vehicle Charging Considering Wind Power Fluctuation

Anyun Yang[1], Hongbin Sun[1, 2,*] and Xiao Zhang[3]

[1]*School of Electrical Engineering, Changchun Institute of Technology, Changchun 130012, China*
[2]*National and Local Joint Engineering Research Center for Smart Distribution Network Measurement, Control and Safe Operation Technology, Changchun 130012, China*
[3]*Department of Energy Technology, Aalborg University, Aalborg East 9220, Denmark*

Received 13 February 2021; Accepted 1 June 2021

Abstract

Electric vehicles (EVs) can inhibit the wind power fluctuations in the generalized form of energy storage. However, optimizing the charging process of EVs under wind power fluctuations is difficult because of the uncertainties of wind power output and user demands. A charging control strategy based on deep reinforcement learning (DRL) was proposed in this study to address the influence brought by uncertain environmental factors to the control. This strategy mined the deep relation between perceiving the uncertainties of environmental factors and learning charging laws by virtue of the perceptual and learning abilities of DRL. An immediate reward mechanism that acts upon the environment was constructed from the angle of neural network fitting function. The EV charging control model was expressed as a Markov decision process (MDP) that contain the state, action, and transfer functions and reward and discount factors through temporal discretization. Next, the single-step updating and experience replay mode were combined to construct the DRL algorithm, followed by the comparative convergence experiment with the reinforcement learning (RL) algorithm that expressed the reward function in mathematical form. In the end, the agent obtained through training was used for the verification of the calculated example. Results show that the constructed RL algorithm is converged by 8,500 episodes earlier. The charging control strategy based on DRL meets the charging requirements when the proportion of optimization objectives is 0.5 and 0.9, and users are allowed to change the allowed charging time temporarily. This study demonstrates that the charging control strategy based DRL can optimize the EVs charging process under many uncertain factors.

Keywords: Markov decision process, deep reinforcement learning, electric vehicle, immediate reward

1. Introduction

With the development of new energy technology and emergence of the smart grid, the application of new energy technology is important because of the local consumption of renewable energy sources in power distribution networks. Electric vehicle (EV), which is an important carrier of new energy technology, especially the technology of consuming renewable energy sources by controlling the EV charging process, has aroused high attention from numerous researchers. However, this technology can hardly reach a perfect effect, which is ascribed to the uncertainties in the power output of renewable energy sources and users' power utilization behavior. Thus, how to flexibly and effectively control the EV charging process under such uncertain factors has become the key to solving the problems.

The traditional research method solves the objective function by perfecting the environmental model of EVs [1] and simulation [2-4]. However, this method proposes high model requirements and even needs to code the problem, thereby militating against real-time dispatching. The deep reinforcement learning (DRL) that is not based on the aforementioned model has become the breakthrough point with the increasingly complicated EV environment, continuous accumulation of uncertain factors, and requirements for intelligent development. DRL [5] combines the perceptual ability of deep learning and decision-making ability of reinforcement learning and realizes the end-to-end learning through the repeated trials and errors of sequential decision problem. In terms of the present EV charging strategies based on DRL, most charging strategies fail to handle the feedback problems generated by uncertain factors in EV environment [6] to DRL. Literature [7] constructed a multi-agent and multi-objective DRL architecture, but it did not consider the limitation of training capacity; as such, the agent could not contain more environmental states. Literature [8] improved the optimization performance of DRL for real-time EV dispatching by virtue of long short-term memory (LSTM) network but did not consider more characteristic quantities. The hierarchical reward function [9] enhances the perceptual ability of DRL but delays the convergence of DRL. In this study, the multilayer perception (MLP) network fitted reward function was used based on the deep deterministic policy gradient (DDPG), which not only improved the perceptual ability of DRL to some extent but also shortened the convergence time and perfected the EV charging strategy.

2. State of the art

To cope with the influence of environmental uncertainties on the EV charging process, the present dispatching methods

*E-mail address: win_shb@163.com
ISSN: 1791-2377
doi:10.25103/jestr.143.12

include day-ahead dispatching [10,11] and real-time dispatching [12,13], where the former is mainly used to study historical data and control EV according to priori knowledge. Nevertheless, its flexibility is restricted. The real-time dispatching method mainly aims to study the present data, where the optimal dispatching is implemented by the control strategy, and the optimization speed of the algorithm is the key. During the optimization of charging process through dynamic programming [14,15], the state space that contains all features in the environment is obtained by the state prediction model through reverse calculation, and the state space connected by the transfer function defines the optimal strategy through the real-time search; thus, the abundance of sample data influences the discreteness of state space and optimization scope of strategy. In Literature [16], the historical and present data, as well as prediction data, were considered to improve the abundance of sample data, and a RL algorithm of batch learning was proposed to learn the optimization strategy from the samples. Next, the optimal charging decision dataset was created using a method based on linear programming. However, the charging strategy based on DRL mainly concerns two aspects: the perceptual ability of DRL for charging environment and the learning performance of DRL. Literature [17] not only formulated the EV charging control model as a Markov decision process (MDP) and proposed a charging control strategy based on DRL but also used LSTM network to extract the day-ahead energy price information to improve the dynamic perceptual ability of the strategy. Then, the DRL training efficiency is enhanced by using two experience replay buffers and adding Gaussian noise into the network. To consider the vehicle-to-grid (V2G) ability of EVs and the discreteness of their charging/discharging level, Literature [18] used the two-layer optimization formula to simulate the pricing of EVs and established the problem in multidimensional continuous state and action space by combining DDPG and experience replay buffers with priority levels. This method could effectively improve the effective utilization efficiency and optimization ability of experience replay buffers, but the priority setting, which depended on priori knowledge, was the key. Literature [19] constructed DRL using a competitive mechanism that improved the DRL learning performance; however, but the strategy of updating the neural network parameters of this method was crucial. In time and space, the laws included in the samples are mainly fed back by the reward function to DRL. Similarly, the uncertain factors in EV environment are mainly presented by the reward function. Literature [20,21] used the mathematical method to express immediate reward, the formulized immediate reward contained the setting relation between state quantities, the calculation results directly acted upon the environmental quality, and the obtained numerical values were transmitted to the neural network, thereby directly impacting the updating of DRL parameters. This method either could not guarantee enough state quantities or could not feed effective information back.

When the EV environment feeds back to DRL, although the perceptual ability of DRL can be improved through the clustering analysis and feature extraction, feeding back enough effective information using the reward function in mathematical form is difficult in consideration of the uncertain factors contained in EV environment. Therefore, starting from solving the form of reward function, a reward function in the form of neural network was designed, a Markov decision process was established for EV charging, and then the DDPG algorithm was combined to complete the EV charging control strategy considering the wind power fluctuation.

The remainder of this work is organized as follows: Section 3 first expounded the principle of EV collaborative wind power digestion, introduced the MDP of EV charging control strategy, and designed a neural network to fit the reward function and improve DDPG. Section 4 analyzed the calculated example and mainly introduced the setting of RL environment and analysis of training results. Section 5 summarized the whole study.

3. Methodology

3.1 Principle of EV collaborative wind power digestion

Under the influence of natural environmental factors and technical limitations, wind power output is prone to reverse peak load regulation and evident fluctuation. The wind curtailment phenomenon can easily occur when the power supply structure is single, namely, when the adjustable power supplies are limited and there is limiting value on the power transmission section. When used as adjustable loads, EVs can effectively inhibit the wind power fluctuation and further improve the quality of wind power.

As shown in Figure 1, the equivalent load of wind power and EV can be adjusted by changing the charging power so as to reduce the fluctuations of wind power at the grid-connected side during the EV charging process, especially when EVs are connected to continuous adjustable charging piles.

In order to simplify the limitation of charging time permitted by EV users and consider the optimization operation under multiple objects (charging cost and wind power fluctuation), the weight method was used to transform the multi-objective problem into a single-objective problem, so the optimization objectives of EV charging behavior are as follows:

$$\min\left(F,C\right)=\alpha\cdot\min\left(F\right)+\left(1-\alpha\right)\cdot\min\left(C\right) \tag{1}$$

$$\min\left(F\right)=\sum_{t_{on}}^{t_{off}}\left(F_{t+1}-F_t\right) \tag{2}$$

$$\min\left(C\right)=\sum_{t_{on}}^{t_{off}}P_{c,t}\cdot E_t^p \tag{3}$$

$$F_t=P_{w,t}-P_{c,t} \tag{4}$$

Where F is the fluctuation of the equivalent load of EV and wind power; C is the charging cost; α is the proportionality coefficient of optimization objective, $\alpha\in\left(0,1\right)$; t_{on} stands for the initial charging time of EV; t_{off} denotes the ending time of EV charging; F_t is the equivalent load of EV and wind power; $P_{w,t}$ represents the wind power at time t; $P_{w,t}$ is the power output by the charging pile at time t; and E_t^p is the electricity price at time t.

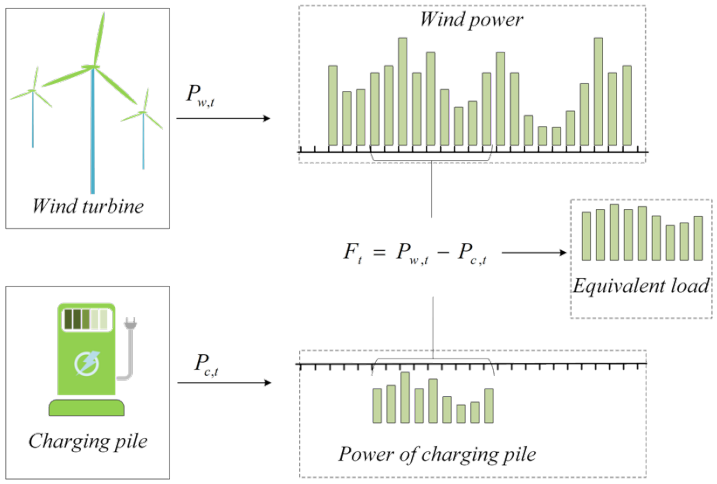

Fig. 1. Conceptual diagram of charging control based on minimum fluctuation of equivalent load

3.2 EV charging control model

In this study, the EV charging control model was formulated as a MDP with discrete time step. In short, the next state only depended on the present state and action, which was the precondition for using DRL.

The model included five elements $\{S, A, P, R, \gamma\}$, where S represents the state set, A is the action set, P stands for the rule and probability for transferring the action a under the present states to the next state s', R is the return function of environment; and γ is the discount factor of return function, where:

(1) $S: s_t = \{P_t^w, V_t^w, C_t, V_t^c, E_t^p, P_t, T_t^n, T_t^a\}$. The state space s_t includes the wind power P_t^w and change rate V_t^w, charging power C_t and change rate V_t^c of charging pile, electricity price E_t^p, and state of charging ((SOC) P_t) of EV battery at the present time t, as well as the time T_t^n needed to meet the user demand and user permitted charging time T_t^a. Here, the state space contains a characteristic quantity—user permitted time T_t^a, indicating that users can change the charging time requirement whenever possible so that the model can conform more to the uncertainty of actual users.

(2) $A: a_t$, where the action a_t is the variable quantity of power output by the charging pile. In this study, the MDP of EV charging control model consisted of 96 time steps $t = \{1, 2, \cdots, 96\}$. When the EVs were charged by connecting to the charging pile and the charging power was kept unchanged within each time step, the mathematical charging model is expressed as follows:

$$\begin{cases} P_t = P_{t-1} + \eta C_t \Delta_t & C_{\min} \le C_t \le C_{\max} \\ P_t \le P_{\max} & \\ C_t = 0 & P_t = P_{\max} \end{cases} \tag{5}$$

where $C_{\min}$ is the minimum value of charging power; $C_{\max}$ is the maximum value of charging power; P_{t-1} is the SOC at the previous time; η denotes the charging efficiency; Δ_t stands for the unit time step, namely, 15 min; and $P_{\max}$ is the maximum value of SOC.

In addition, the action is subjected to the following limitations because charging piles have many types:

$$a_t^{i,\min} \le a_t \le a_t^{i,\max} \tag{6}$$

where $a_t^{i,\min}$ and $a_t^{i,\max}$ represent the minimum and maximum values permitted by the charging power of the charging pile number i.

(3) $P: P(s_{t+1}|s_t, a_t)$, the transfer probability decides how the environment skips to the state s_{t+1} at the next time step.

(4) $R: R_t$, which represents the cumulative reward of the state s_t transformed into the state s_{t+1} after the action a_t is completed.

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{7}$$

where r_t is the immediate reward of the state s_t transformed into the state s_{t+1} by executing the action a_t at time t.

(5) $\gamma: \gamma \in [0,1]$, and the discount factor facilitates the cumulative reward to trade off the future reward.

3.3 DRL architecture

DRL aims to acquire the maximum cumulative reward, but the immediate reward function in cumulative reward has a bearing on the quality of agent training. A good instant reward function that fully include environmental factors and comprehensively describe the objective optimization degree can contribute to the fast convergence of neural network in the training process.

3.3.1 Neural network of instant reward function

Considering that the neural network had the potential of fitting any functions, a neural network of immediate reward function (RN) was designed in this study. The nerve cell structure adopted at the hidden layer of this network is shown in Figure 2.

The activation function of nerve cell is a rectifying linear function (RELU), as shown in Formula (8), and the output function F_k is as seen in Formula (9). If introduced into the neural network, then $RELU$ could promote the high sparsity of network to improve its temporal and spatial efficiency and avoid gradient vanishing.

The neural network of immediate reward function was set as the multilayer perceptron type, and the hidden layer was of (768,96,8) structure. The feature r_c of reward was learned from the characteristic state S_c using the network, so as to provide an appropriate immediate reward value in any states.

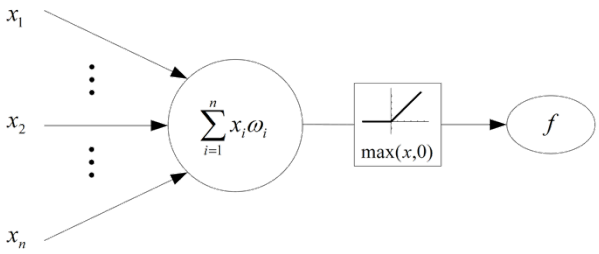

Fig. 2. Neuron diagram of reward function

$$RELU(x)=\max(x,0)=\begin{cases}x, x\geq 0\\ 0, x<0\end{cases} \tag{8}$$

$$F_K = RELU(\omega_0 x_0+\omega_1 x_1+\cdots+\omega_n x_n) \tag{9}$$

As shown in Figure 3, the designed neural network was combined with the EV system to obtain a complete EV interaction environment. The EV environment received and executed the action a_t to obtain the state s_{t+1} at the next time step,and the present state obtained the immediate reward r_t through RN. The obtained data could be trained by DRL as samples.

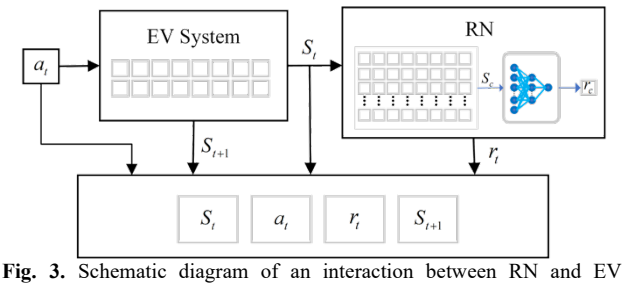

Fig. 3. Schematic diagram of an interaction between RN and EV environment

3.3.2 Particle turbulence operation

In this study, DDPG was used as the RL method, which not only absorbed the single-step updating framework of "Actor-Critic" (AC) but also took full advantage of the fixed network mode and experienced the replay technique of "Deep Q-Learning" (DQN). Hence, it had the more effective learning ability on continuous actions. In Figure 4, four neural networks, namely, Actor-Network, Critic-Network, Actor-Target-Network, and Critic-Target-Network, were reported. The improved DDPG was obtained by adding RN.

The input and output of Actor-Network were the present state S and action A, respectively. After receiving the output of Actor-Network and the state S, the Critic-Network outputs the function values that correspond to the output state and action, and the values were applied to the parameter updating of Actor-Network and calculation of loss function (TD-error). The Actor-Target-Network was just slightly different from the Critic-Target-Network.

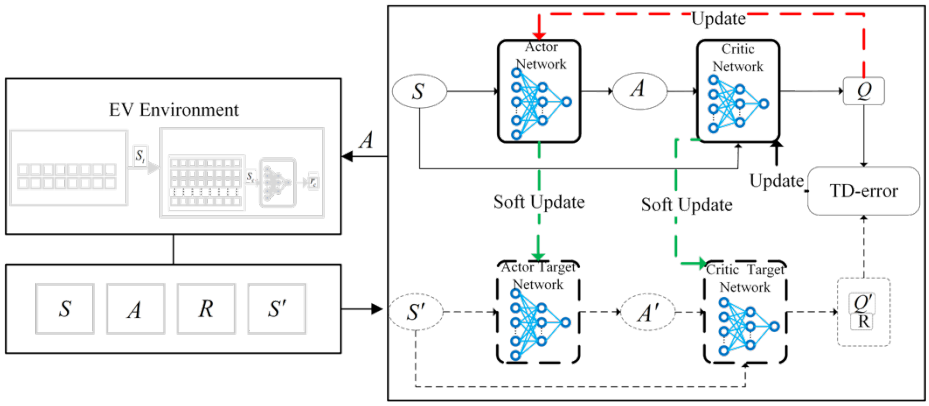

Fig. 4. Stucture of reinforcement learning

The concrete training process of optimized charging control is presented as follows:

The parameters $\theta^{\mu},\theta^{Q},\theta^{\mu'},\theta^{Q'}$ of neural network were initialized, and each parameter included the weight $\omega_\theta \sim N(0,0.3)$ and bias $b_\theta = 0.1$.

$s_t=\{P_t^w,V_t^w,C_t,V_t^c,E_t^p,P_t,T_t^n,T_t^a\}$ the present state was obtained by the interaction with the environment. The action V_c was acquired on the basis of the present state through the Actor-Network.

The action V_c was taken as the expectation and ε_V as the variance to construct a normal distribution $V_c \sim N(V_c,\varepsilon_V)$ followed by the random processing of output action. The value of ε_V depended on the exploration degree, and it would gradually decline with the training process, that is, it was finally subjected to the action output by the network.

The action was executed by interacting with the environment to obtain the next state and reward. The state and action at time t and those ($\{s_t,a_t,r_t,s_{t+1}\}$) at time $t+1$ were put into the experience replay buffer. The batch training and learning began after the data in the experience replay buffer reached the maximum capacity.

The Actor-Target-Network began receiving the state at the next time and outputted the actual executed action.

The Critic-Network evaluated the present state and action to obtain the value function, whose value $Q(s_t,a_t|\theta^Q)$ had two functions. On the one hand, it was used to calculate the loss TD-error, and on the other hand, it updated the Actor-Network together with the output of this network. The loss function L is listed as follows:

$$L=\frac{1}{N}\sum t(y_t-Q(s_t,a_t|\theta^Q))^2 \tag{10}$$

$$y_t=r_t+\gamma Q'(s_{t+1},\mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'}) \tag{11}$$

where y_t is acquired by calculating the immediate reward r_t and output value Q' of Critic-Target-Network at time t. $\mu'(s_{t+1}|\theta^{\mu'})$ is the output value of the Actor-Target-Network.

The Actor-Network was updated through the following formula:

$$\nabla_{\theta^\mu}\mu\Big|s_t = \frac{1}{N}\sum_t \nabla_a Q(s,a|\theta^Q)\big|s=s_t, a=\mu(s_t)\nabla_{\theta^\mu}\mu(s|\theta^Q)\big|s_t \quad (12)$$

The updating was further implemented through the parameters of back propagation Actor-Network.

$$\theta^\mu = \theta^\mu - \beta^\alpha \nabla_{\theta^\mu}\mu|s_t \quad (13)$$

where β^α and β^c in Formula (14) correspond to the learning rates of respective networks.

Through the back propagation, the Critic-Network could be updated by the loss function L.

$$\theta^Q = \theta^Q - \beta^c \nabla_{\theta^Q} L \quad (14)$$

The soft updating mode was adopted for the target network.

$$\begin{cases} \theta^{\mu'} = \sigma\theta^\mu - (1-\sigma)\theta^{\mu'} \\ \theta^{Q'} = \sigma\theta^Q - (1-\sigma)\theta^{Q'} \end{cases} \quad (15)$$

where σ is the soft updating factor.

4 Result analysis and discussion

4.1 Data environment settings

As seen in Table 1, the wind power and its change rate, charging power of charging pile and its velocity, battery capacity of EV and permitted charging time all followed the uniform distribution. The initial wind power p^w was randomly generated in the uniform distribution of $U\sim(0,50)$. Similarly, the change rate v^w of wind power was randomly generated in $U\sim(-4,4)$. With the lapse of time, the wind power at the next time was decided by the wind power and rate at the present time. The initial charging power p^c of the charging pile followed $U\sim(0,30)$ distribution, the initial charging velocity v^c of charging pile was sampled from $U\sim(-5,5)$, and the v^c at the next time was generated by the RL network and restricted within $[-5,5]$. Similar to wind power, the caring power of charging pile at the next time depended on p^c and v^c. The electricity price E^p followed the normal distribution of $N\sim(0.75,0.45^2)$, and its sampling was restricted within $[0.3,1.2]$. The initial battery capacity E^t of EV was randomly generated from $U\sim(0,0.4)$. The user permitted charging time was sampled from $U\sim(2,6)$.

Table 1. Model Parameters Settings

Model Parameters	Distribution
Wind power p^w	$U(0,5)$
Wind power rate v^w	$U\sim(-4,4)$
Charging power p^c	$U\sim(0,30)$
Charging power rate v^c	$U\sim(-5,5)$
Energy price E^p	$N\sim(0.75,0.45^2)$
Battery energy level E^t	$U\sim(0,0.4)$
Battery capacity E^c	$U\sim(30,60)$
Allow charging time T^a	$U\sim(2,6)$

As it was assumed in this study that the users' charging demands must be satisfied, the user permitted time T^a also must be greater than the needed charging time T^n, which was calculated on the basis of time spent in fully charging the EV at the maximum power as follows:

$$T^n = \frac{E^c \cdot (1-E^t)}{p^c_{\max}} \quad (16)$$

where $p^c_{\max}$ is the maximum charging power of the charging pile; E^c is the battery capacity of EV; and E^t is the battery capacity of EV at time t.

The hyper-parameter settings of RL are listed in Table 2. The total number of training episodes was 40,000. Following the test, the number of episodes ep^2 in each test set was set as 200. Each step size represented 15 min, and the permitted maximum charging time was 5 h, so the highest number of steps in each episode was set as 20. Learning rates β^a and β^c of two networks in RL were set as 0.001, and the discount factor γ was set as 0.9. The capacity buf^s of experience replay buffer, batch training size, and soft updating factor σ were set as 20,000, 32, and 0.01.

Table 2. Hyperparameters Settings.

Model Parameters	Value
Total number of episodes for training ep^1	40000
Total number of steps for each episodes ep^2	200
Test the model per episodes ep^3	20
Learning rate for actor β^a	0.001
Learning rate for critic β^c	0.001
Discount factor γ	0.9
Size of replay buffer buf^s	20000
Batch size for learning b^s	32
Soft update factor σ	0.01

4.2 Training results of EV charging optimization

First, Figure 5 displays the training results of Formula (1), which is an immediate reward function. To facilitate the observation, the average reward of 20 training episodes was selected. The reward value began converging at the 13,000th episode and tended to be steady at the 18,000th episode until all episodes were ended.

The training results of RN are presented in Figure 6. The reward value began converging at the 4,500th episode, reached the maximum value at the 5,500th episode, and kept steady fluctuation until all rounds were ended.

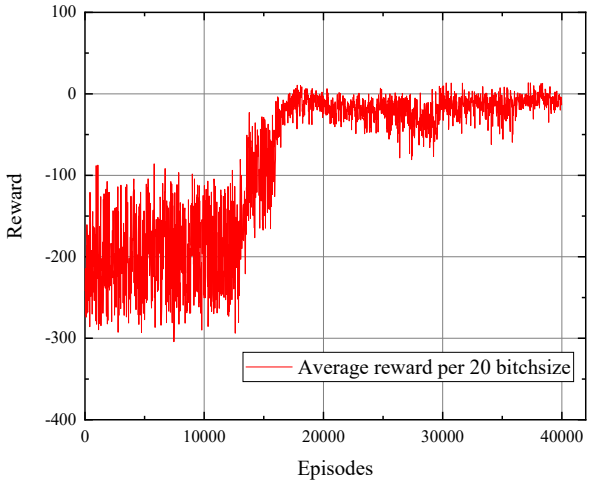

Fig. 5. The learning curve of DDPG with equation (1) as the immediate reward function

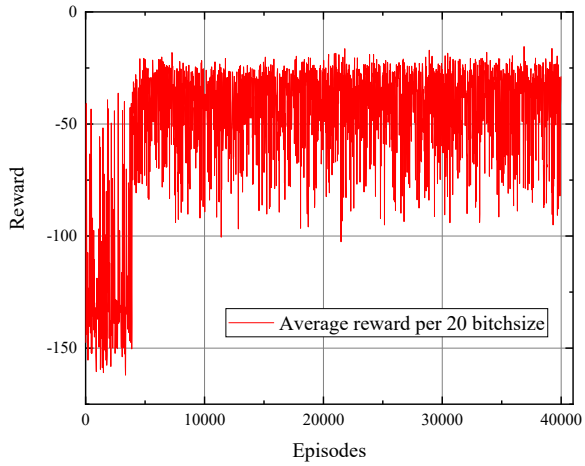

Fig. 6. The learning curve of DDPG using neural network as an immediate reward function

By comparing Figure 5 and Figure 6, the learning curve of the improved method could be converged fast, and the convergence process was reduced from 5,000 to 1,000 episodes. After the maximum reward was reached, the learning curve fluctuated greatly because the improved immediate reward contained more state quantities.

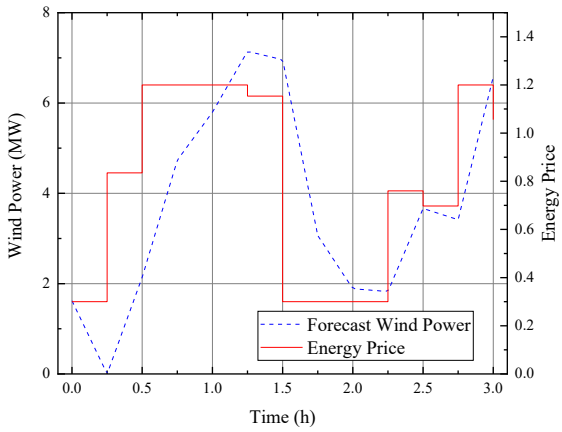

Fig. 7. Line chart of forecasted wind power output and energy price

The trained agent had good performance in the test set. Besides the current electric quantity of EV and user permitted charging time, the agent needed to acquire the predicted wind power quantity and electricity price, as shown in Figure 7, where the dotted line denotes the predicted wind power. Although the predicted value within 150 min was given, the agent could only acquire the predicted value within 15 min in the training process because of the uncertainty of wind power. The solid line represents the electricity price. Figure 7 also shows that the electricity price was highly correlated with wind power. In the test case, the initial SOC level was set as 0.06 and the battery capacity as 59 KW. The user permitted time was 5.75 h.

After being optimized through the improved DDPG, the charging pile, namely, the agent, could complete the EV charging within the user-permitted time and optimize it under different proportions of wind power fluctuation and electricity price. As shown in Figure 8, it took different time for the charging pile to complete EV charging under different α values, and the charging power was changing with the wind power fluctuations, thereby reducing the fluctuations of wind power. The charging power under $\alpha = 0.5$ experienced no obvious change in comparison with that under $\alpha = 0.9$. Figure 7 shows that the trend of wind power was approximate to that of electricity price. As shown in Figure 9, the change rate of charging power was also changed with the change rate of wind power, the functions of EVs, as generalized energy storage units, could be fully exerted, thereby buffering the wind power consumption.

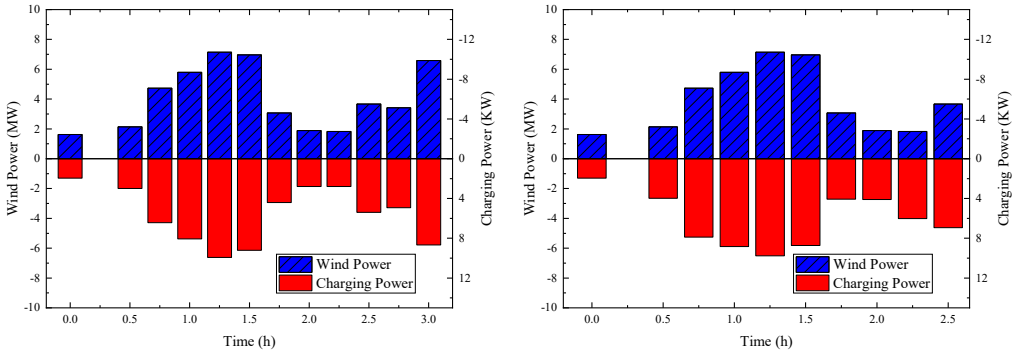

Fig. 8. Wind power and charging power in the proportion of: (a) $\alpha = 0.5$. (b) $\alpha = 0.9$

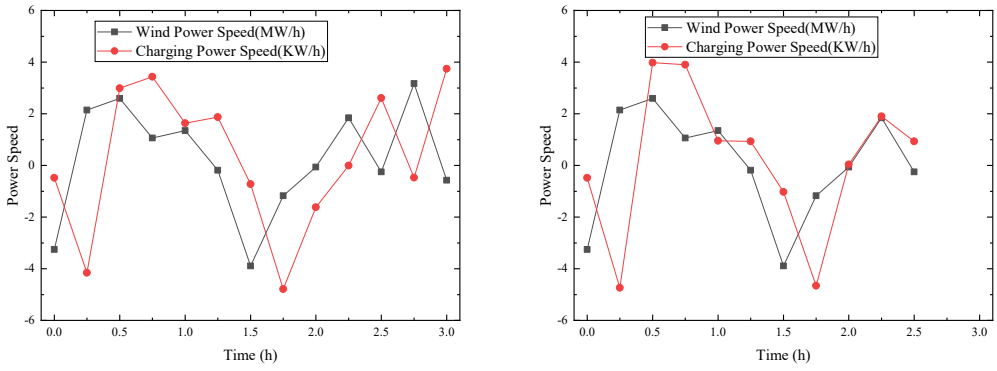

Fig. 9. Wind power rate and charging power rate in the proportion of: (a) $\alpha = 0.5$. (b) $\alpha = 0.9$

As it was difficult to predict user behaviors, the user uncertainty was simulated by changing the permitted charging time in the above case. To be more specific, the permitted time was altered into 2 h when the charging proceeded to 60 min, namely, 8 step sizes of agent. The optimization results are presented by the dotted line in Figure 10. The agent did not rely upon the complete charging period, but instead, they could perform appropriate action by only needing the data in similar cases. Therefore, the agent could serve individual users.

The above experiments have proved the effectiveness of the proposed PSO-net algorithm. The turbulence operator and the local search strategy account for the good performance

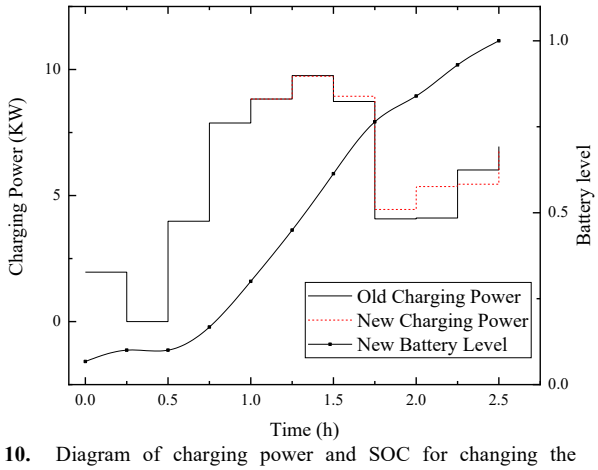

Fig. 10. Diagram of charging power and SOC for changing the allowable time during charging

5. Conclusions

Deep reinforcement learning requires an interactive environment to provide effective feedback. To improve the convergence of DRL training and obtain effective agent through training, an instant reward function was designed in this study, and an EV charging control strategy based on DRL was proposed for the real-time feedback of environmental state and overcoming the influences of uncertain factors on the charging process. The following conclusions were drawn:

(1) The instant reward function based on neural network can contain all state quantities of EV environment, and the network itself is of plasticity, so it can be used to improve the learning performance of DRL.

(2) The EV charging control strategy based on improved DDPG can change the control strategy according to the changes in optimization objectives and also adjust it according to the changes in EV environment.

The proposed charging control strategy based on DDPG can realize the optimized EV charging control under uncertain wind power, electricity price, charging pile, EV, and user requirements. However, the control strategy is not extended to multiple agents; hence, its extensibility remains to be further tested.

Acknowledgements

This work was supported in part by the Scientific and Technological Planning Project of Jilin Province (20180101057JC). A Project Supported by Scientific and Technological Planning Project of Jilin Province (20190302106GX).

References

1. Liu, D., Wang, Y., Shen, Y., "Electric Vehicle Charging and Discharging Coordination on Distribution Network Using Multi-Objective Particle Swarm Optimization and Fuzzy Decision Making". *Energies*, 9(3), 2016, pp.030186.
2. Usman, M., Knapen, L., Yasar, A., Bellemans, T., Wets, G., "Optimal recharging framework and simulation for electric vehicle fleet". *Future Generation Computer Systems*, 107, 2020, pp. 745-757.
3. Alonso, M., Amaris, H., Germain, J. G., Galan, J. M., "Optimal Charging Scheduling of Electric Vehicles in Smart Grids by Heuristic Algorithms". *Energies*, 7(4), 2014, pp.2449-2475.
4. Floch, L., Meglio, D., Moura, J., "Optimal charging of vehicle-to-grid fleets via PDE aggregation techniques". In: *2015 American Control Conference (ACC)*, Chicago, IL, USA: IEEE, 2015, pp.3285-3291.
5. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A., "Human-level control through deep reinforcement learning". *Nature*, 518(7540), 2015, pp.529-533.
6. Rigas, S., Ramchurn, D., Bassiliades, N., "Managing Electric Vehicles in the Smart Grid Using Artificial Intelligence: A Survey". *IEEE Transactions on Intelligent Transportation Systems*, 16(4), 2015, pp.1619-1635.
7. Silva, D., Nishida, H., Roijers, M., "Coordination of Electric Vehicle Charging Through Multiagent Reinforcement Learning". *IEEE Transactions on Smart Grid*, 11(3), 2020, pp.2347-2356.

8. Wan, Z., Li, H., He, H., Prokhorov, D., "Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning". *IEEE Transactions on Smart Grid*, 10(5), 2019, pp. 5246-5257.
9. Lee, S., Choi, D., "Energy Management of Smart Home with Home Appliances, Energy Storage System and Electric Vehicle: A Hierarchical Deep Reinforcement Learning Approach". *Sensors*, 20(7), 2020, pp.072157.
10. Sedighizadeh, M., Shaghaghi-shahr, G., Esmaili, M., "Optimal distribution feeder reconfiguration and generation scheduling for microgrid day-ahead operation in the presence of electric vehicles considering uncertainties". *Journal of Energy Storage*, 21, 2019, pp. 3422-3432.
11. Rana, R., Mishra, S., "Day-Ahead Scheduling of Electric Vehicles for Overloading Management in Active Distribution System via Web-Based Application". *IEEE Systems Journal,* 13(3), 2019, pp. 82-97.
12. Shin, M., Choi, D., Kim, J., "Cooperative Management for PV/ESS-Enabled Electric Vehicle Charging Stations: A Multiagent Deep Reinforcement Learning Approach". *IEEE Transactions on Industrial Informatics*, 16(5), 2020, pp.3493-3503.
13. Kanellos D., "Optimal Scheduling and Real-Time Operation of Distribution Networks With High Penetration of Plug-In Electric Vehicles". *IEEE Systems Journal,* early access, 2020, pp.1-10.
14. Ali, M., Ghanbar, A., Soffker, D., "Optimal control of multi-source electric vehicles in real-time using advisory dynamic programming". *IEEE Transactions on Vehicular Technology*, 68(11), 2019, pp.10394-10405
15. Fettinger, N., Ten, C., Chigan, C., "Minimizing residential distribution system operating costs by intelligently scheduling plug-in hybrid electric vehicle charging". *2012 IEEE Transportation Electrification Conference and Expo (ITEC)*, Dearborn, MI, USA: IEEE, 2012, pp.1-8.
16. Chis, A., Lunden, J., Koivunen, V., "Reinforcement Learning-Based Plug-in Electric Vehicle Charging with Forecasted Price". *IEEE Transactions on Vehicular Technology*, 66(5), 2017, pp.3674-3684.
17. Zhang, F., Yang, Q., Dou, A., "CDDPG: A Deep Reinforcement Learning-Based Approach for Electric Vehicle Charging Control". *IEEE Internet of Things Journal*, 8(5), 2020, pp.3075-3087.
18. Qiu, D., Ye, Y., Papadaskalopoulos, D., Strbac, G., "A Deep Reinforcement Learning Method for Pricing Electric Vehicles with Discrete Charging Levels". *IEEE Transactions on Industry Applications*, 56(5), 2020, pp.5901-5912.
19. Zhang, Y., Zhang, Z., Yang, Q., An, D., "EV charging bidding by multi-DQN reinforcement learning in electricity auction market". *Neurocomputing*, 397, 2020,pp.404-414.
20. Du, G., Zou, Y., Zhang, X., Liu, T., Wu, J., He, D., "Deep reinforcement learning based energy management for a hybrid electric vehicle". *Energy*, 2020, 201, 2020 pp.117591-1117601.
21. Yue, H., Weimin, L., Kun, X., Zahid, T., Chenming, L., "Energy Management Strategy for a Hybrid Electric Vehicle Based on Deep Reinforcement Learning". *Applied Sciences*, 8(2), 2018, pp.020187.