



Article

An Optimal Scheduling Strategy of a Microgrid with V2G Based on Deep Q-Learning

Yuxin Wen ¹, Peixiao Fan ^{1,*} , Jia Hu ², Song Ke ¹ , Fuzhang Wu ¹ and Xu Zhu ¹¹ School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China² State Grid Hubei Electric Power Co., Ltd., Wuhan 430072, China

* Correspondence: whufpx0408@163.com; Tel.: +86-177-0599-0685

Abstract: In recent years, the access of various distributed power sources and electric vehicles (EVs) has brought more and more randomness and uncertainty to the operation and regulation of microgrids. Therefore, an optimal scheduling strategy for microgrids with EVs based on Deep Q-learning is proposed in this paper. Firstly, a vehicle-to-grid (V2G) model considering the mobility of EVs and the randomness of user charging behavior is proposed. The charging time distribution model, charging demand model, state-of-charge (SOC) dynamic model and the model of travel location are comprehensively established, thereby realizing the construction of the mathematical model of the microgrid with EVs: it can obtain the charging/discharging situation in the EV station, so as to obtain the overall output power of the EV station. Secondly, based on Deep Q-learning, the state space and action space are set up according to the actual microgrid system, and the design of the optimal scheduling reward function is completed with the goal of economy. Finally, the calculation example results show that compared with the traditional optimization algorithm, the strategy proposed in this paper has the ability of online learning and can cope with the randomness of renewable resources better. Meanwhile, the agent with experience replay ability can be trained to complete the evolution process, so as to adapt to the nonlinear influence caused by the mobility of EVs and the periodicity of user behavior, which is feasible and superior in the field of optimal scheduling of microgrids with renewable resources and EVs.

Keywords: renewable energy; electric vehicles; deep Q-learning; microgrid scheduling; V2G

Citation: Wen, Y.; Fan, P.; Hu, J.; Ke, S.; Wu, F.; Zhu, X. An Optimal Scheduling Strategy of a Microgrid with V2G Based on Deep Q-Learning. *Sustainability* **2022**, *14*, 10351. <https://doi.org/10.3390/su141610351>

Academic Editors: Xiaoqing Bai, Chun Wei, Peijie Li and Dongliang Xiao

Received: 19 July 2022

Accepted: 18 August 2022

Published: 19 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

A microgrid is a micropower system oriented to terminal energy users such as buildings, communities, industrial parks or towns and is one of the main forms of energy in the future human society [1]. Its operation stability is usually maintained by various microsources [2,3]. In recent years, the application of distributed power sources in power systems has become more and more extensive, which brings more and more randomness and uncertainty to the operation and regulation of microgrids [4].

Nowadays, a large number of scholars at home and abroad have completed mature research on the optimization and dispatching technology of a microgrid. In [5], the uncertainty of photovoltaic power generation is modeled based on probabilistic constraints, so that an optimal scheduling method using chance-constrained programming to minimize the operating cost of microgrids is proposed. In [6], a genetic algorithm based on a memory mechanism is proposed to solve the problem of minimizing the operating cost of a microgrid. In [7], for a microgrid system combining renewable energy and traditional power generation, an energy management strategy for a hybrid thermoelectric island microgrid is proposed based on a multi-objective particle swarm optimization (MOPSO) algorithm. In [8], based on the chaotic search particle swarm optimization algorithm, with the goal of minimizing the total cost, the economic operation optimization model of the microgrid is constructed from three aspects: operating cost, environmental impact and system safety,

so as to effectively reduce the operating cost of the microgrid and ensure the safety and stability of the power supply and electricity consumption. However, the above research methods are prone to falling into a local optimum when dealing with nonlinear or nonconvex problems, their computation time is long, and the generalization learning ability of the algorithm is insufficient. Therefore, in the face of the strong uncertainty of distributed power and load, the existing traditional optimization algorithms have difficulties meeting the requirements of microgrid optimal scheduling due to the above limitations.

Meanwhile, with the continuous development of new energy vehicles, the EVs industry has gradually become large-scale and market-oriented, and V2G technology has become more mature [9,10]. Therefore, the research of EVs in participating in power grid peak shaving and valley filling, and smoothing power fluctuations has also become more and more in-depth [11], but their mobility and randomness of user behavior also bring greater challenges to maintaining the economic operation of microgrids. In [12], based on the MPC algorithm, EVs are used as a mobile energy storage to participate in microgrid regulation, but the output power in the control model is constrained to a fixed value. In [13], the randomness of user travel demand is considered in the V2G model, and the EVs state of charge is modeled, but the impact of EV mobility on the controllable capacity of EV stations is not considered. In addition, the above V2G models are all modeled with the EV station as a complete output, which cannot reflect the internal charging and discharging of the EVs. In practice, the controllable capacity of the EV station would change randomly due to the user's charging behavior and the mobility of EVs.

Therefore, in order to cope with the randomness and uncertainty caused by the access of EVs and various distributed power sources to the economic dispatch of microgrids, reinforcement learning algorithms have been applied in the field of power systems [14,15]. In [16], the reinforcement learning theory is introduced to construct a mathematical model suitable for microgrid energy management, which solved the economical optimal scheduling optimization problem of a microgrid better. In [17], the model of reinforcement learning agent is applied to a microgrid system with distributed energy, which can formulate the optimal strategy for energy management and load scheduling among the three main bodies of a power source, distributed energy storage and user. In [18], facing the economic dispatch problem of microgrids with distributed energy resources, based on the reinforcement learning framework, an optimal equilibrium selection mechanism is proposed, which can improve the operation performance of microgrids in terms of economy and independence. However, the above research does not focus on the V2G modeling of EVs and cannot truly reflect the process of EVs participating in the optimal scheduling of microgrids. Meanwhile, [16–18] are mainly based on traditional reinforcement learning algorithms, which cannot solve the dimensionality disaster of policy sets in the face of complex environments or continuous actions. It is difficult to deal with the influence of the random change of the controllable capacity of the EV station and the uncertainty of the distributed power and load on the economic dispatch of the microgrid.

In summary, an optimal scheduling strategy for microgrids with electric vehicles based on Deep Q-learning is proposed in this paper. The main contributions are as follows:

- (1) A V2G mathematical model considering the mobility of EVs and the randomness of user charging behavior is proposed. The user charging time distribution model, charging demand model, EV state-of-charge (SOC) dynamic model and the model of travel location are comprehensively established, so that the agent can obtain the charging/discharging situation in an EV station to obtain the overall output power of the EV station.
- (2) A microgrid optimization scheduling strategy based on Deep Q-learning is proposed. The strategy has the ability of online learning and can cope with the randomness of renewable resources better. Meanwhile, the agent with experience replay ability can be trained to complete the evolution process, so as to adapt to the nonlinear influence caused by the mobility of EVs and the periodicity of user behavior, which

is feasible and superior in the optimal scheduling of microgrids with renewable resources and EVs.

The remainder of this paper is organized as follows: In Section 2, the mathematical model construction of a microgrid with EVs is established. The microgrid dispatch model based on Deep Q-learning is introduced in Section 3. The simulation results are presented and analyzed in Section 4, and the conclusions are summarized in Section 5.

2. The Mathematical Model Construction of Microgrid with EVs

The high penetration of renewable energy into the power grid may affect a series of problems such as the balance of supply and demand in the system and the stable operation of the power grid. Additionally, EVs are able to support large-scale integration of renewable energy by absorbing excess energy and returning it to the grid when needed. V2G technology can use the mobile energy storage characteristics of EVs to reasonably adjust their charging/discharging behavior, thereby alleviating the impact of load fluctuations [19,20]. Therefore, based on the randomness of user behavior and the mobility of EVs, a charging/discharging model for EVs is constructed, and an optimal scheduling model for microgrids with EVs is established.

2.1. The V2G Model of EVs

Firstly, it is assumed that the electric vehicle is fully charged before traveling, and the battery power consumption of the electric vehicle has a linear relationship with the daily mileage [21]. That is, after obtaining the probability distribution of the daily mileage of the electric vehicle, the probability distribution of the battery state of charge SOC_0 of the electric vehicle when it returns to the charging station can be obtained. The return time of different EVs within a day and the corresponding charging time are also important components of the V2G model of the EVs station.

In the existing EV model, due to the regular travel behavior of users, the arrival time and location of electric vehicles are relatively fixed. However, in the actual situation, EVs have mobility due to the randomness of real-time road network. For example, in the case of traffic congestion, users will adjust charging route decision, which will affect the arrival time and location of EVs and then affect the power consumption of EVs when they enter the station.

Uncertain influences such as road network congestion are closely related to the type of user distribution. For example, electric vehicles that are distributed in commercial areas for a long time are more likely to experience congestion during their journey, and electric vehicles in public areas have higher driving speed and lower unit power consumption. Therefore, the main travel behaviors and the proportion of each activity trip of electric vehicles can be obtained in this paper, as shown in Table 1 and Figure 1. Among them, various distribution areas can be divided into residential areas, commercial areas and public areas, abbreviated as R, C and P, respectively.

Table 1. Electric private car travel chain.

The User Types	The Chain of Travel	The Proportion/%
1	R→C→R	52.8
2	R→P→R	24.1
3	R→C→P→R	23.1

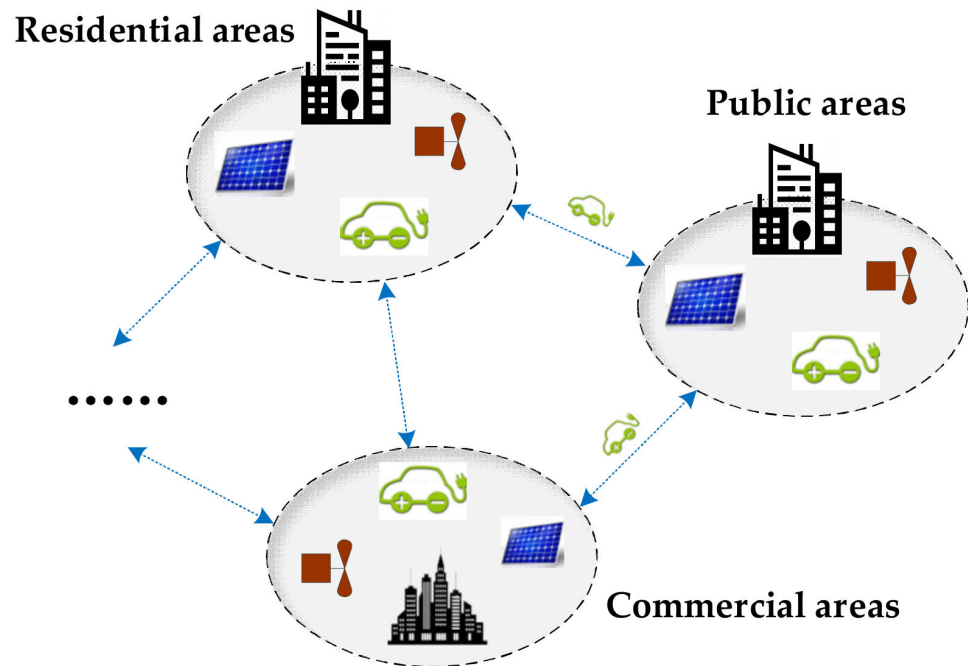


Figure 1. Mobility of EVs and randomness of user habits.

Therefore, the daily mileage obeys a log-normal distribution $L \sim \text{Log} - N(\mu_L, \sigma_L^2)$, and its probability density function is shown in Formula (1):

$$f(L) = \frac{1}{L\sigma_L\sqrt{2\pi}} e^{-\frac{(\ln L - \mu_L)^2}{2\sigma_L^2}} \tag{1}$$

where μ_L and σ_L represent the mean and variance, respectively, which is determined by different types of user behavior.

In addition, it is worth noting that according to the behavior data of electric vehicle users, the vehicle owner charges every ϵ days on average, so the probability density function of the total mileage of the electric vehicle when it enters the charging station can be obtained as shown in (2) and obeys a log-normal distribution $L \sim \text{Log} - N(\mu_L, \sigma_L^2)$.

$$f(L) = \frac{\epsilon}{L\sigma_L\sqrt{2\pi}} e^{-\frac{(\ln L - \mu_L)^2}{2\sigma_L^2}} \tag{2}$$

Secondly, it can be assumed that the EV returns at time t , obeying the normal distribution $t_0 \sim \text{Log} - N(\mu_s, \sigma_s^2)$, and its probability density function is Formula (3).

$$f(t) = \begin{cases} \frac{1}{\sigma_s\sqrt{2\pi}} e^{-\frac{(t-\mu_s)^2}{2\sigma_s^2}} & \mu_s - 12 < t < 24 \\ \frac{1}{\sigma_s\sqrt{2\pi}} e^{-\frac{(t+24-\mu_s)^2}{2\sigma_s^2}} & 0 < t < \mu_s - 12 \end{cases} \tag{3}$$

where μ_s and σ_s represent the mean and variance, respectively, which is determined by different types of user behavior as well.

Furthermore, it can be assumed that the charging power of the EVs after entering the charging station is constant. When the state of charge of the EV battery reaches SOC_m , the driving process expected by the user after the EV leaves the charging station can be satisfied. Therefore, according to the daily driving mileage, the time for the battery capacity to be charged to SOC_m after the EV enters the station can be calculated as T_c :

$$T_c = \frac{LQ_{100} - W_{total} + W_m}{100P_c} (T_c \geq 0) \tag{4}$$

where L is the daily driving distance of the EVs, P_c is the charging power and Q_{100} is the power consumption per 100 km, W_{total} is the full power of the EVs, and W_m is the power of the EVs when the state of charge is at SOC_m .

The duration of EV stagnation in the electric vehicle station can be defined as ΔT , and the departure time is defined as T_{leave} . It is easy to know that $\Delta T \geq T_c$. Therefore, ΔT and T_{leave} satisfy the following formula.

$$\Delta T = (1 + \sigma_T)T_c \quad (5)$$

$$T_{leave} = T_{enter} + \Delta T \quad (6)$$

where σ_T is a positive random number, and its value would be selected according to the user's travel habits on weekdays; T_{enter} is the EV inbound charging time.

As shown in Figure 2, when the state of charge reaches SOC_m , or the state of charge is greater than SOC_m when entering the station, the EVs will be able to participate in the load distribution optimization scheduling process of the microgrid. That is, it can be discharged when the microgrid encounters peak power consumption, and this discharge process will not make the EV power lower than SOC_m . When the state of charge of the EV reaches SOC_{max} , the EVs will no longer be charged to ensure battery life. At this time, the EVs will automatically stop charging (maintain SOC_{max}) or discharge.

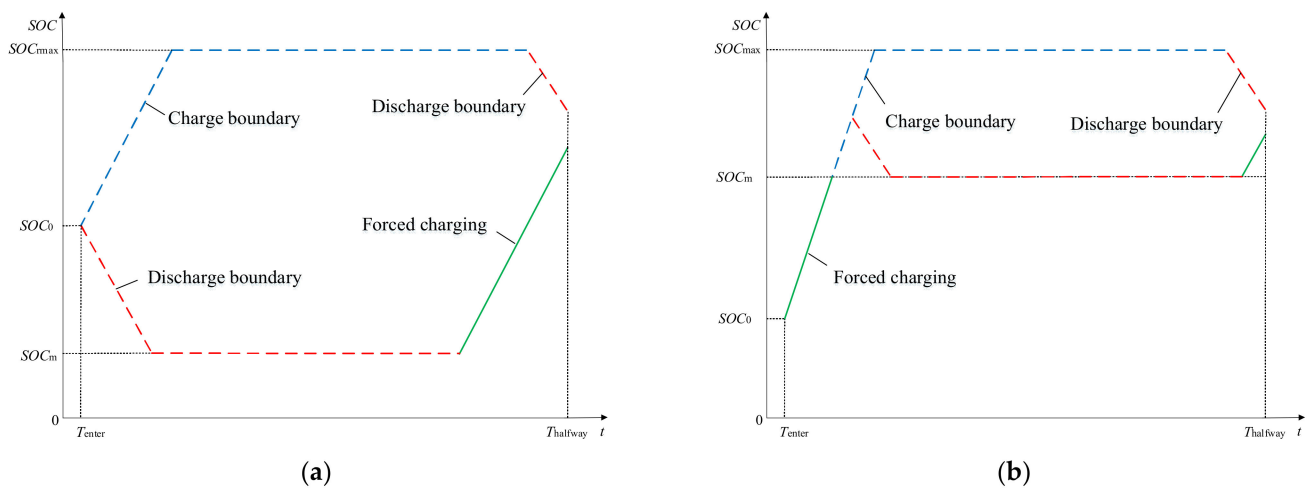


Figure 2. The charging/discharging constraint boundary of EVs. (a) The constraint boundary of EVs when $SOC_0 < SOC_m$. (b) The constraint boundary of EVs when $SOC_0 > SOC_m$.

To sum up, different EVs will have different entry time $T_{enter,i}$ and the necessary charging time ΔT_i and will automatically participate in scheduling or continue charging according to the load status of the microgrid after the charge reaches SOC_m , and leave the charging station when $T_{leave,i}$. The specific scheduling process is shown in Figure 3: EV1 enters the station at time T_1 , and its state of charge is less than SOC_m at this time, so it enters the state of charge and participates in the scheduling of feeding at time T_3 until T_4 , at which time the state of charge of the vehicle is greater than SOC_m ; EV2 enters the station at time T_2 , and its state of charge is greater than SOC_m at this time, so it can immediately participate in dispatching and distribution when the microgrid is at peak load power consumption until T_4 ; EV3 is extremely low in battery power when entering the station, so it is always kept charged, and it is always kept charged until time T_5 .

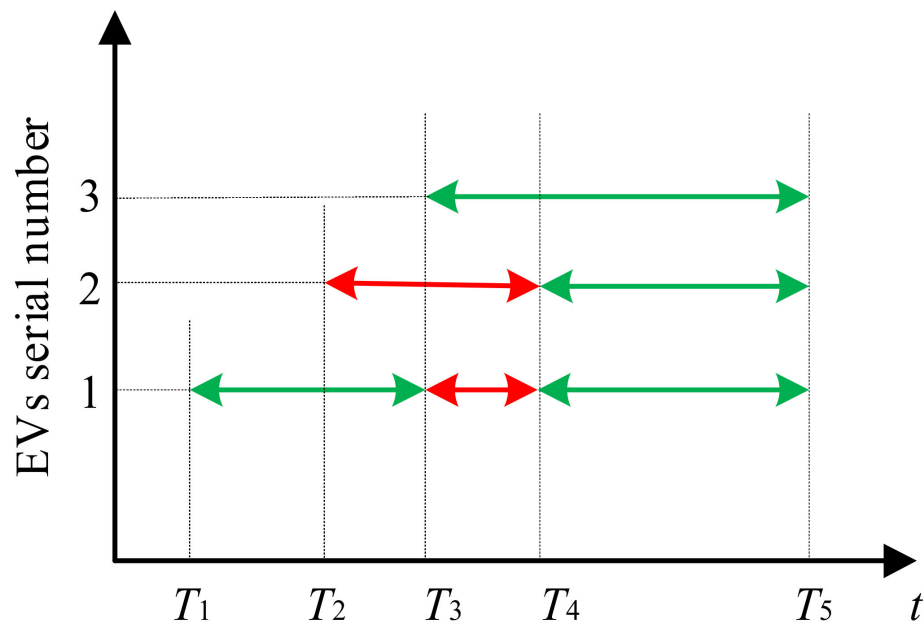


Figure 3. The specific scheduling process of V2G.

Therefore, it can be assumed that there are n EVs in the station at time t , including i vehicle in a state of nonchargeable and dischargeable ($SOC = SOC_{max}$), including j vehicle in a state of rechargeable and non-dischargeable ($SOC < SOC_m$), and the remaining vehicles in both charging and discharging states ($SOC_m < SOC < SOC_{max}$). It can be obtained that at time t , the boundary of the overall charging power of the EV station is shown in (7):

$$\begin{cases} P_{EV}^+(t) = (n - i) \cdot P_{ch} \\ P_{EV}^-(t) = (n - j) \cdot P_{dis} \\ 0 \leq \Delta P_{EV}^+(t) \leq P_{EV}^+(t) \\ 0 \leq \Delta P_{EV}^-(t) \leq P_{EV}^-(t) \end{cases} \quad (7)$$

where $P_{EV}^+(t)$ and $P_{EV}^-(t)$ are the boundaries of the interactive power output by the charging station. $\Delta P_{EV}^+(t)$ and $\Delta P_{EV}^-(t)$ represent the interactive power between the electric vehicle charging station and the microgrid at time t , which are determined by the agent: the agent can select the optimal action according to the actual situation and economic benefits of the microgrid and obtain the charging/discharging situation in the EV station at the current moment, so as to obtain the overall output power of the EV station, as shown in (8):

$$\Delta P_{EV}(t) = -\Delta P_{EV}^+(t) + \Delta P_{EV}^-(t) \quad (8)$$

2.2. The Optimal Dispatching Model of Microgrid

The microgrid structure considered in this paper is shown in Figure 4. The microgrid consists of wind turbines, photovoltaics, micro-gas turbines, an EVs station and other units. Therefore, the optimal dispatching model of a microgrid including EVs is constructed.

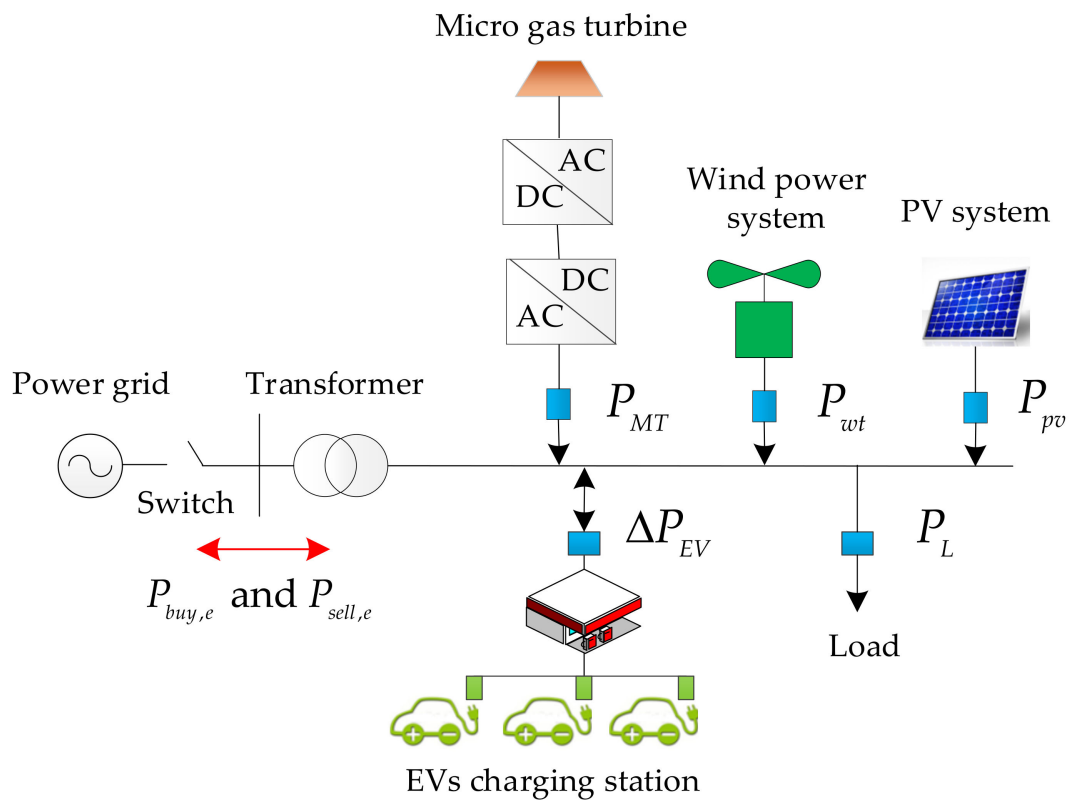


Figure 4. The structure of the microgrid with an EVs station.

Where P_L is the load disturbance power, P_{wt} is the wind disturbance power, P_{pv} is the photovoltaic power generation power, P_{MT} is the power variation of MT, ΔP_{EV} is the power variation of EVs and P_e is the power between the microgrid and the large grid.

2.2.1. Objective Function

Considering the randomness of wind and solar loads, an optimization model is established with the goal of minimizing the expected total economic operating cost during the optimization period. Its objective function is shown in (9):

$$F = \min \sum_{t=1}^T (C_{gas}(t) + C_e(t) + C_{EV}(t)) \quad (9)$$

where C_{gas} represents the cost of purchasing natural gas for the micro-gas turbine, C_e represents the cost of purchasing and selling electricity generated by the interaction between the microgrid and the power grid and C_{EV} represents the cost of purchasing and selling electricity generated by the charging and discharging of EVs.

$$\begin{cases} C_{gas}(t) = c_{gas} \frac{P_{MT}(t)}{\eta_{MT} q_{NG}} \\ C_e(t) = e_b P_{buy,e}(t) - e_s P_{sell,e}(t) \\ C_{EV}(t) = c_b \Delta P_{EV}^+(t) - c_s \Delta P_{EV}^-(t) \end{cases} \quad (10)$$

where P_{MT} represents the output of the micro-gas turbine at time t , η represents the conversion efficiency of the micro-gas turbine, q_{NG} represents the low calorific value of natural gas and c_{gas} represents the gas purchase cost coefficient of the micro-gas turbine. $P_{buy,e}$, $P_{sell,e}$ represent the power purchase and sale between the microgrid and the large grid, e_b and e_s represent the cost coefficient of purchasing and selling electricity. ΔP_{EV}^+ and ΔP_{EV}^- represent the charging power and discharging power of the electric vehicle charging station, and c_b and c_s represent the cost coefficient of charging and discharging.

2.2.2. Constraints

(1) Power Balance Constraints:

$$\begin{aligned}
 P_{wt}(t) + P_{pv}(t) + P_{MT}(t) + P_{buy,ev}(t) + P_{buy,e}(t) \\
 = L(t) + P_{sell,ev}(t) + P_{sell,e}(t)
 \end{aligned}
 \tag{11}$$

where $P_{wt}(t)$, $P_{pv}(t)$ represent the output power of winds and photovoltaics in the t period, and $L(t)$ represents the load in the t period.

(2) Micro gas turbine operating constraints:

$$\begin{aligned}
 -R_d \Delta t \leq P_{MT}(t) - P_{MT}(t - \Delta t) \leq R_u \Delta t \\
 P_{MT,min} \leq P_{MT}(t) \leq P_{MT,max}
 \end{aligned}
 \tag{12}$$

where P_{MT} represents the output power of the micro-gas turbine, R_d and R_u represent the downward and upward ramp rates of the micro-gas turbine and $P_{MT,min}$, $P_{MT,max}$ represent the lower and upper output limits of the micro-gas turbine.

(3) Grid interaction power constraints:

$$\begin{aligned}
 0 \leq P_{sell,e}(t) \leq P_{ex,max} \\
 0 \leq P_{buy,e}(t) \leq P_{ex,max} \\
 P_{sell,e}(t) \times P_{buy,e}(t) = 0
 \end{aligned}
 \tag{13}$$

(4) EV station constraints:

The constraints on the overall charge/discharge power of the EVs station have been given in Section 2.1, as shown in (7).

In addition, it can be considered that the main function of the EV station is to provide charging services for the users, and the priority of ensuring that the user's EVs is sufficient is the highest. Therefore, the charging and discharging power constraints of each EV in the EVs station can be obtained:

$$P_{dis} \leq P_{ch}
 \tag{14}$$

Meanwhile, the interest relationship between EV users, microgrid operators and large grids is considered, and the price constraints can be obtained, as shown in Figure 5:

$$\begin{aligned}
 e_s < e_b \\
 e_s < c_s \\
 c_b < e_b \\
 c_s < c_b
 \end{aligned}
 \tag{15}$$

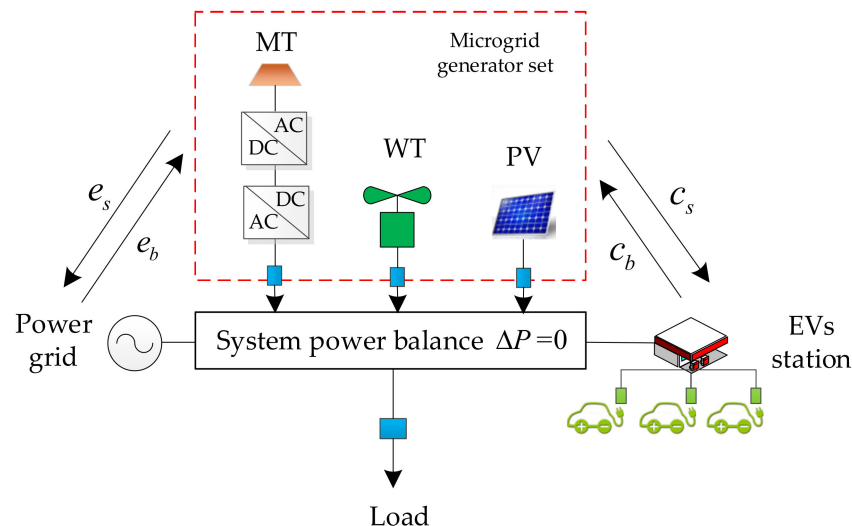


Figure 5. The relationship between microgrid transactions.

3. A Microgrid Dispatch Model Based on Deep Reinforcement Learning

3.1. Theory of Reinforcement Learning Algorithms

Reinforcement learning RL is a learning algorithm that maps from environmental states to actions, and its goal is to maximize the cumulative reward of an agent during trial and error with a given environment [22,23].

To achieve these functions, the reinforcement learning framework consists of agents, which are able to take certain actions a_t based on the current state s_t , as shown in Figure 6. After choosing an action at time t , the agent receives a scalar reward r_{t+1} and finds itself in a new state s_{t+1} , which depends on the current state and the chosen action.

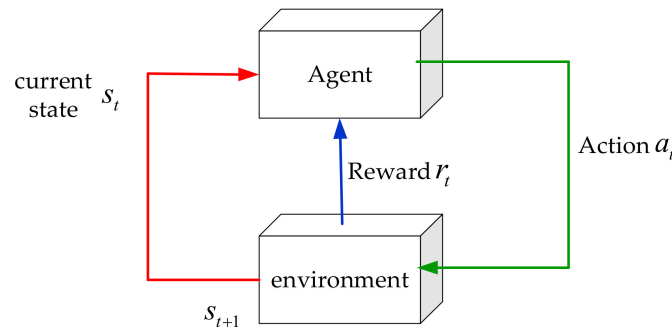


Figure 6. The schematic diagram of deep reinforcement learning.

As shown in Figure 7, the Markov decision process satisfies the Markov property and is the basic formalism of reinforcement learning, which can be described as:

$$P(s_{t+1}|s_0, a_0, \dots, s_t, a_t) = P(s_{t+1}|s_t, a_t) \quad (16)$$

where P is the state transition probability.

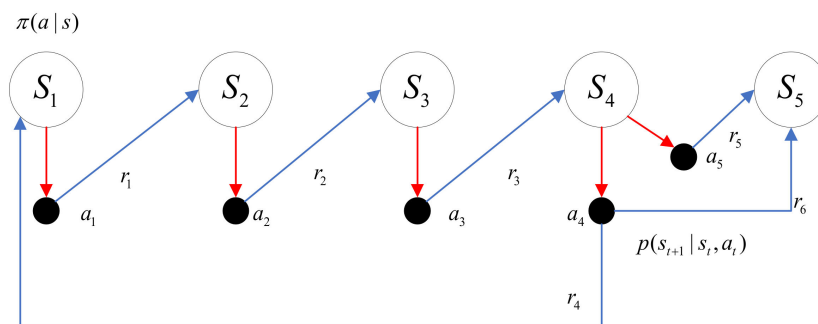


Figure 7. An illustration of a Markov decision process.

At each epoch, the agent takes actions to change its state in the environment and provide rewards. To further process the reward value, a value function and optimal policy are proposed. To maximize the long-term cumulative reward after the current time t , for a finite time horizon ending at time t , the payoff R_t is shown in (17):

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \quad (17)$$

where the Discount factor $\gamma \in [0, 1]$, and γ can take 1 only in intermittent MDP.

To find the optimal policy, some algorithms are based on a value function $V(s)$, which represents how beneficial the agent is to reaching a given state s . This function depends on the agent's actual policy π :

$$V^\pi(s_t) = E[R_t | s_t = s] = E \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s \right] \quad (18)$$

Similarly, the action-value function Q expresses the value of taking action a in state s under policy π as:

$$Q^\pi(s_t, a_t) = E[R_t | s_t = s, a_t = a] = E \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a \right] \quad (19)$$

In the Q-learning algorithm, the Q-function can be expressed in an iterative form by the Bellman equation:

$$Q^\pi(s_t, a_t) = E[r_{t+1} + \gamma Q^\pi(s_{t+1}, a_{t+1}) | s_t, a_t] \quad (20)$$

The optimal policy π^* is the policy that yields the largest cumulative reward in the long run:

$$\pi^* = \underset{\pi}{\operatorname{argmax}} V^\pi(s) \quad (21)$$

At this point, the optimal value function and action value function are shown in (22):

$$\begin{cases} V^*(s) = \max V^\pi(s) \\ Q^*(s, a) = \max_{\pi} Q^\pi(s, a) \end{cases} \quad (22)$$

3.2. Design of Optimal Scheduling Strategy for Microgrid Based on Deep Q-Learning

Deep Q-learning has the advantage of being suitable for solving optimal decision-making problems with uncertain factors and can be applied to solve the optimal scheduling problem of a microgrid considering intermittent renewable energy generation and charging uncertainty of EV users. Therefore, the established mathematical model of the optimal scheduling problem for microgrids is transformed into a Deep Q-learning framework in this section.

The basic components of reinforcement learning include: the state space S representing the environment, the action space A representing the action of the agent, and the reward function r for training the agent.

(1) State space:

The state variables of the microgrid system include user electrical load demand, photovoltaic power generation power, wind turbine power generation power, charging and discharging power capability of EV stations and dispatching time period. Therefore, the state space can be expressed as:

$$S = [P_{PV}(t), P_{WT}(t), L(t), P_{EV}^+(t), P_{EV}^-(t), \Gamma^+(t), \Gamma^-(t), t] \quad (23)$$

(2) Action space:

After the agent observes the state characteristics of the environmental system, it generates actions based on the agent's own strategy π . Actions in the microgrid model with EVs can be represented by the output power of the micro-gas turbine, the interaction power between the EVs station and the microgrid and the power purchased and sold between the microgrid and the grid. Therefore, the action space can be expressed as:

$$A = [P_{MT}(t), \Delta P_{EV}^+(t), \Delta P_{EV}^-(t), P_{buy,e}(t), P_{sell,e}(t)] \quad (24)$$

In addition, when the power of MT and EV is known, the interaction power between the microgrid and the grid can be calculated by the power balance constraint. Therefore, the action space can be simplified as:

$$A = [P_{MT}(t), \Delta P_{EV}^+(t), \Delta P_{EV}^-(t)] \quad (25)$$

(3) Reward function:

In the optimal scheduling model of the microgrid proposed in this paper, the goal is to minimize the overall operating cost of the system, which includes the cost of purchasing and selling electricity between the microgrid and the grid, the cost of purchasing and selling electricity between the EVs station and the grid, and operating costs of micro-combustion

engines. Therefore, in this paper, the minimization problem is transformed into the form of reward value maximization under the reinforcement learning framework, and the reward function expression of the agent can be expressed as:

$$r(t) = -(C_{gas}(t) + C_e(t) + C_{EV}(t)) \quad (26)$$

In addition, when the agent is in the early stage of exploration, the policy model is not yet mature, and some actions may not meet the constraints. Therefore, it is necessary to set up an early termination mechanism to construct a penalty term to improve the training speed. From the action space (25), it can be known that the interactive power between the microgrid and the grid is solved by derivation of the balance constraint. Therefore, there is a problem that the interactive power crosses the line and the constraints cannot be satisfied. To sum up, the reward function is constructed by stacking penalty terms, as shown in Equation (27):

$$\begin{cases} r(t) = -(C_{gas}(t) + C_e(t) + C_{EV}(t)) - f_d |P_g(t) - P_{ex,max}| \\ P_g(t) = |P_{buy,e}(t) - P_{sell,e}(t)| \end{cases} \quad (27)$$

where f_d is the penalty term coefficient.

3.3. Neural Network Structure

In the optimization model of this paper, the random constraints of electric vehicles and the output of new energy are strongly nonlinear data. Deep Q-learning combines deep neural network and reinforcement learning, so it has the ability to effectively process large-scale data: agent training can be completed through a large amount of data, so as to output real-time decisions according to real-time state variables and obtain the optimal scheduling scheme. Therefore, this paper takes the state vector S as the input sequence through the neural network and finally gets the approximated Q value in the output layer. The corresponding network structure is shown in Figure 8, which has h layers of hidden layers, and each hidden layer is composed of u neurons, and the specific value of the (h, u) parameter is affected by the actual calculation example. In the optimization model of this paper, the neural network has a total of four hidden layers, and the ReLU (Rectified Linear Unit) function is used as the activation function.

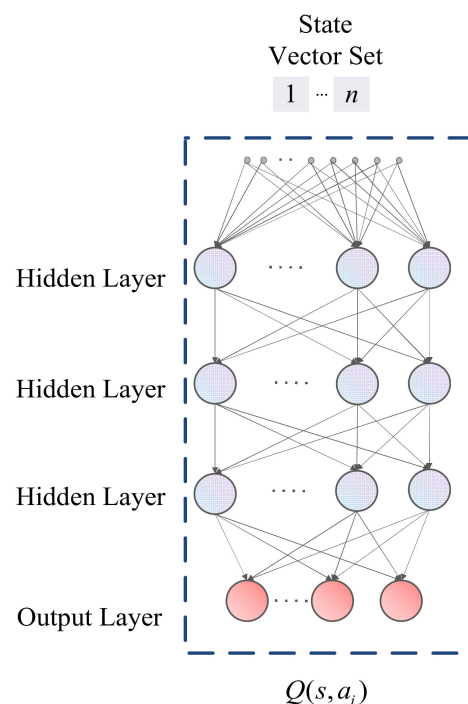


Figure 8. Schematic diagram of neural network structure based on Deep Q-learning.

3.4. The Flow Diagram of Deep Q-learning Algorithm

The dispatch strategy of this paper is carried out in the following steps:

First, determine the state set of the system as S . Furthermore, the action space can be defined as A .

Second, the parameters are adjusted according to the actual computing instance, and the values of the reward function coefficients and hyperparameters are obtained.

Finally, the agent is trained, and after convergence, the known information of the microgrid is input to the agent so as to obtain the optimal dispatch scheduling result of the next day.

In summary, after applying the deep neural network to Q-learning, Deep Q-learning introduces the experience playback mechanism and the freezing parameter mechanism in order to reduce the correlation between samples and improve the stability of training. Therefore, combined with the application scenarios of this paper, the training process of Deep Q-learning in the microgrid with EVs can be obtained as shown in Figure 9.

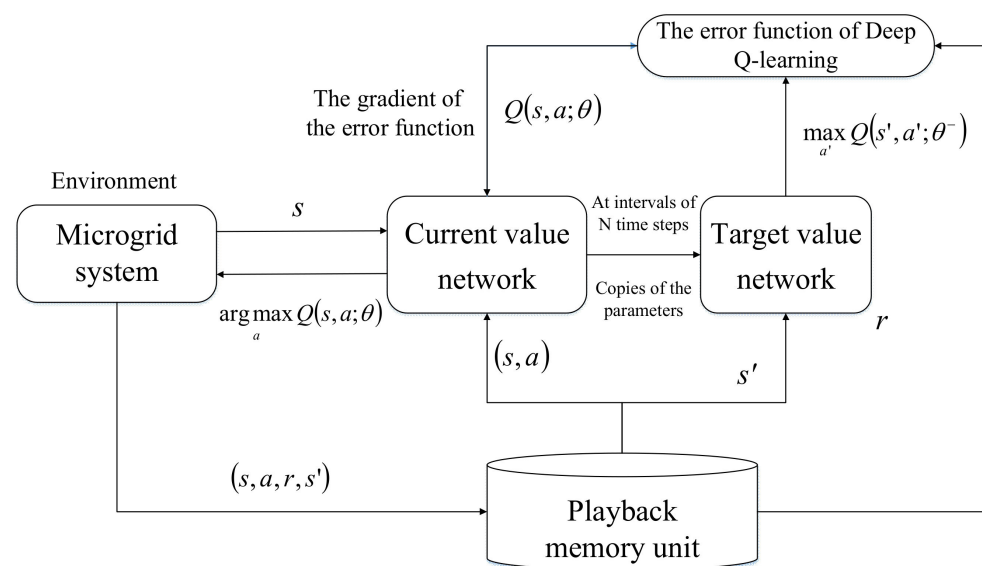


Figure 9. The training process of Deep Q-learning in the microgrid with EVs.

4. Simulation Results

In order to verify the effectiveness of the economic dispatch strategy for microgrids with electric vehicles based on Deep Q-learning proposed in this paper, the microgrid system with EVs shown in Figure 4 was used as an example for simulation research. The microgrid system includes a wind turbine WT, a photovoltaic PV, a micro-gas turbine MT and an electric vehicle charging station. The equipment abbreviations and working parameters in the system are shown in Table 2. During the operation of the system, the range of the interactive power between the system and the grid is $[-1000, 1000]$ kW, and the PV and WT output according to the real-time maximum power generation. In this paper, the purchase cost of natural gas is 0.059 USD/kWh, the electricity purchase price of the system is 0.074 USD/kWh and the electricity selling price is 0.044 USD/kWh.

Table 2. Abbreviations and working parameters of each device.

Unit	Parameter	Meaning	Value
MT	η	generation efficiency	0.85
	P_{MT}	capacity of MT	1000 kW
EV	P_{ch}	charge power for EV	5 kW
	P_{dis}	discharge power for EV	2.5 kW

Therefore, the hyperparameter settings of the Deep Q-learning agent can be obtained as follows: the discount factor γ is 0.9, the data sampling size is 256, the experience pool size is 10^6 , the network parameter learning rate α is 0.0001 and the Adam optimizer is used to update the network weights. The iterative training times are 5×10^5 times. In this paper, Python software and the computing unit of CPUi7-10700 are used in the simulation experiment platform to construct and verify the simulation model.

4.1. Case1: Analysis of Electric Vehicle Mobility and User Behavior Habits

From the model in Section 2.1, it can be seen that the controllable capacity of the EV station is affected by the mobility of EVs and the randomness of the user. Take the model constructed in this paper to generate the distribution of EVs in the microgrid on a certain day as an example, as shown in Figure 10.

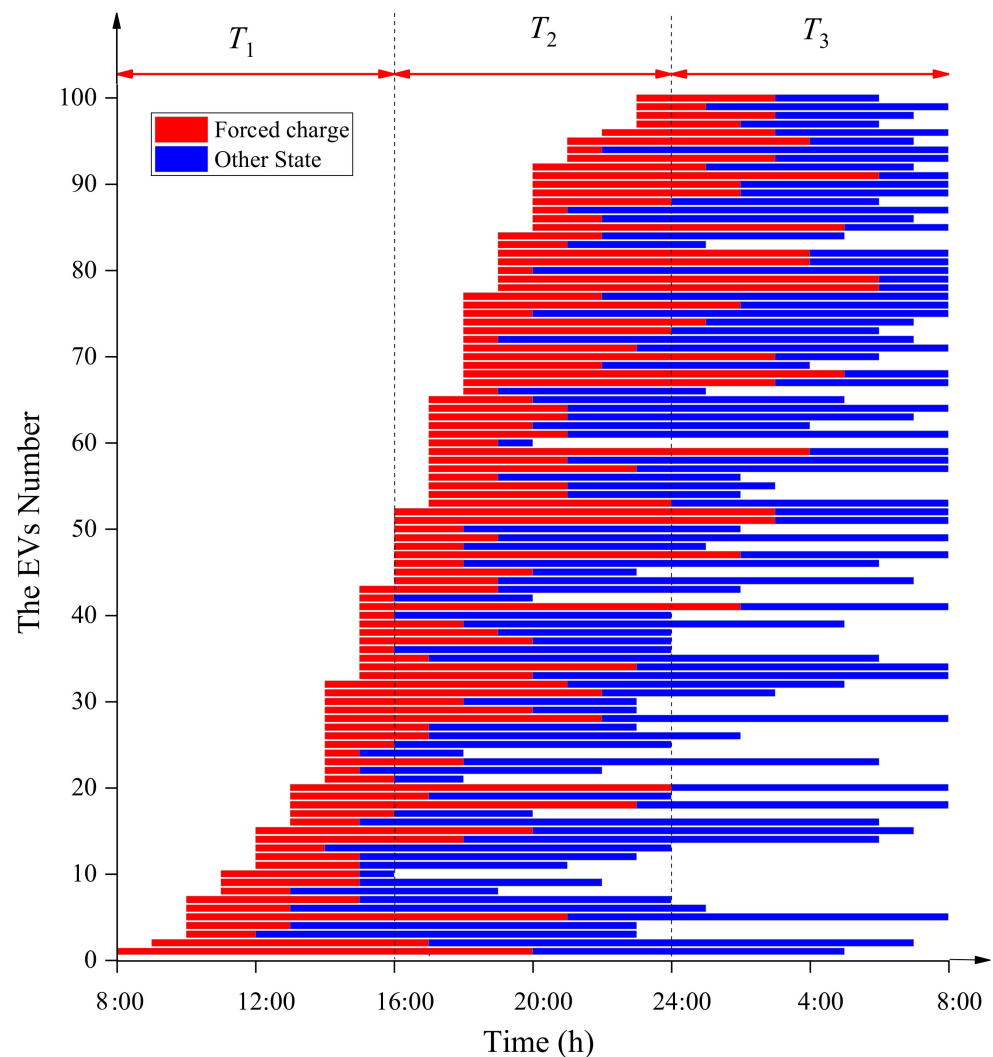


Figure 10. Parking situation in an EV station within 24 h.

It can be seen that the number of EVs (maximum controllable capacity) in the EV station during T1, T2 and T3 is quite different. Therefore, the ability of EV stations to participate in microgrid regulation will also show obvious time peaks and valleys during the day, which is closely related to the living habits of the user group: Between 8:00 and 12:00, a small number of EVs gradually entered the station. After 12:00, the number of EVs entering the station began to increase rapidly and reached saturation at 23:00. In addition, after 24:00, the power of most electric vehicles has exceeded SOC_m . At this time, the controllable capacity of the station reaches its peak and starts to gradually decrease

at 4:00, and a large number of EVs leave the charging station at 8:00, which makes the controllable capacity of the station plummeted.

The above situation has brought a strong nonlinear influence to the microgrid dispatching process, which makes the traditional algorithm without evolution ability unable to adapt, thus posing a challenge to the dispatching of the power grid. Therefore, in order to better reflect the superiority of the Deep Q-learning algorithm in the dispatching of microgrids with EVs, the PSO algorithm will be introduced in this paper as a comparison.

4.2. Case2: Energy Dispatching Results of a Microgrid

After the Deep Q-learning agent completes the training process, it accumulates enough experience to be able to complete the intelligent scheduling process of the microgrid [24]. The optimization comparison results of using Deep Q-learning and PSO algorithms solving the same scheduling day scenario are shown in Figures 11 and 12 and Table 3. Among them, the specific data of typical scheduling day scenario are shown in the yellow and pink histograms and blue lines in Figures 11 and 12. It can be seen that:

(1) Between 8:00–12:00, because most of the vehicles were stranded outside the station, the output of the EV station was small. Between 16:00–22:00, the EV station mainly acted as the load. At this time, most of the EVs were charged in the station. After 22:00, the controllable capacity of the EV station gradually reached its peak value and could be discharged to participate in dispatching.

(2) Compared with the PSO algorithm, the Deep Q-learning algorithm had the online learning ability and could adapt better to the staged mutation of the capacity of the EV station caused by the mobility of EVs and the randomness of user behavior based on the experience accumulated in the training process, which significantly enhanced the robustness and adaptability of the microgrid.

(3) The total operating cost of the microgrid under the Deep Q-learning algorithm was 801.07 USD, and the calculation time was 0.5 s. The total operating cost of the microgrid under the PSO algorithm was 814.57 USD, and the calculation time was 7 min 23 s. In detail, the natural gas cost of the microgrid under the PSO algorithm was 825.34 USD, which was smaller than the 897.7 USD obtained by the Deep Q-learning algorithm, because the total output of the MT in the solution result of the PSO algorithm was smaller than that of the Deep Q-learning. In fact, the unit cost of power generation of MT is the lowest, that is, the PSO algorithm needed to make up for the shortage of electricity by purchasing a large amount of electricity from the grid: the electricity purchase cost of the microgrid under the PSO algorithm was 159.07 USD, which was much greater than that of the Deep Q-learning algorithm.

(4) As shown in Figure 11, the PSO could not adapt to the nonlinear effects brought about by changes in the constraints of EVs, and its scheduling results were mostly in the charging state. Although a small number of EVs participate in the discharge, the charging station as a whole cannot discharge and is in a continuous charging state. As shown in Figure 12, Deep Q-learning could adopt the most economical charging and discharging strategy under the constraint conditions, could discharge properly to reduce the power supply pressure when the load was high, and acted as a power source at night to achieve economy.

In summary, it can be seen that the Deep Q-learning algorithm was better than the PSO algorithm in all aspects. Among them, the advantage in flexible handling of randomness of EV stations is particularly obvious.

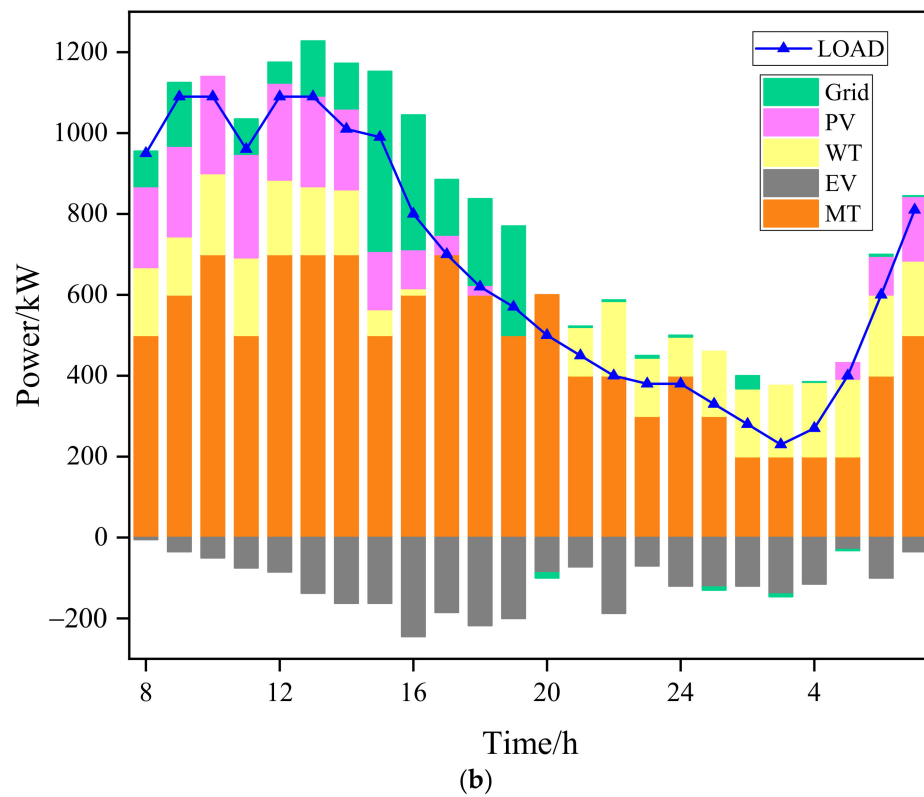
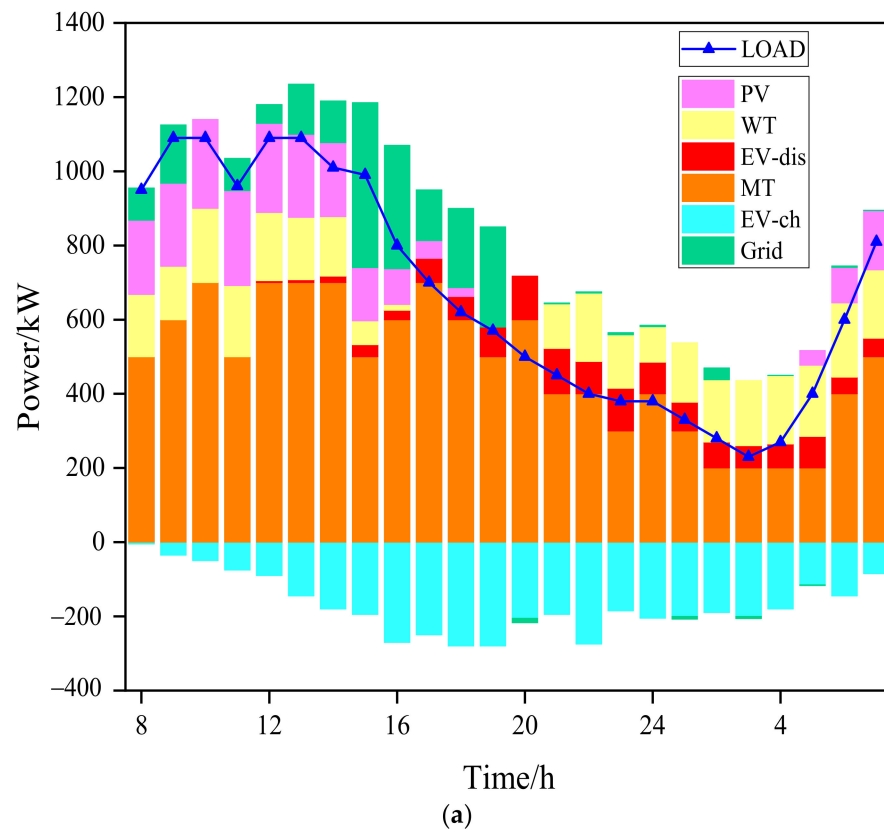
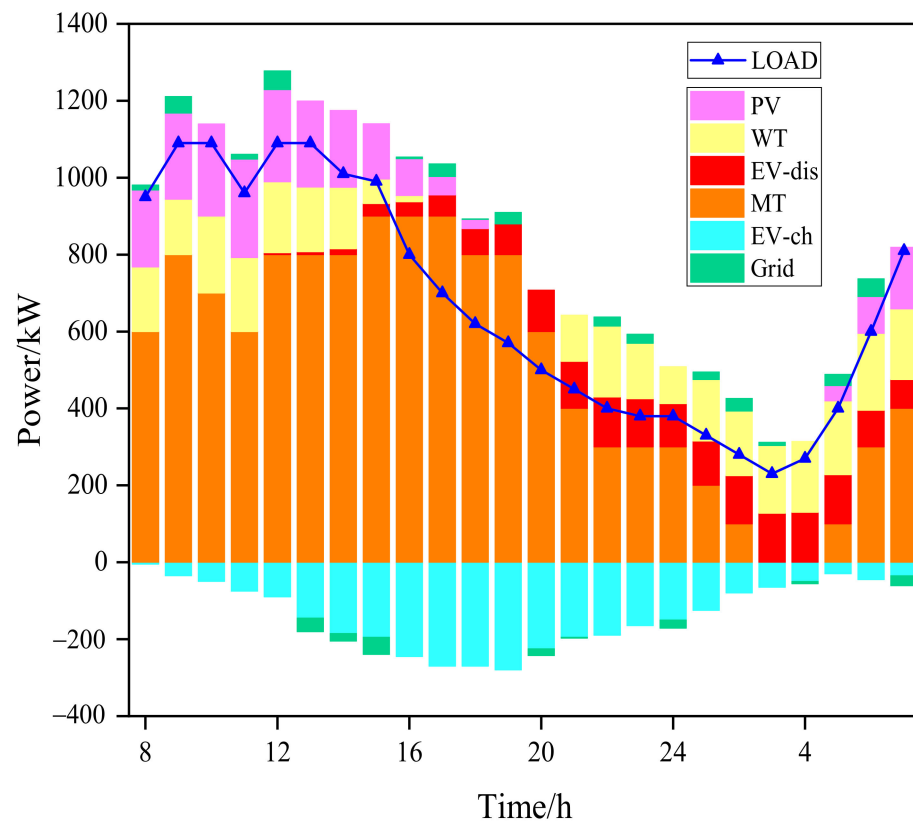
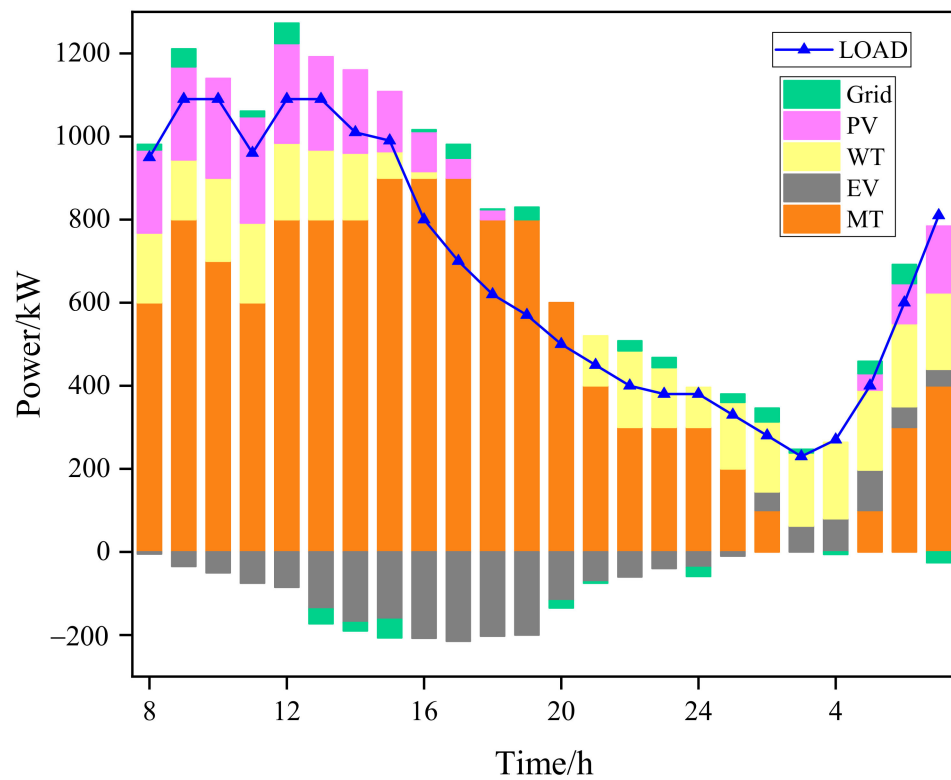


Figure 11. Microgrid scheduling results based on the PSO algorithm. (a) Microgrid schedule results when EVs output is divided into charging and discharging. (b) Microgrid schedule results when the output of EVs is the whole.



(a)



(b)

Figure 12. Microgrid scheduling results based on the Deep Q-learning algorithm. (a) Microgrid schedule results when EVs output is divided into charging and discharging. (b) Microgrid schedule results when the output of EVs is the whole.

Table 3. Comparative analysis of dispatch results' data.

Index	Operating Costs (USD)	Gas Costs (USD)	V2G Costs (USD)	Grid-Connected Costs (USD)	Calculation Time
PSO	814.57	825.34	−169.84	159.07	7 min 23 s
Deep Q-learning	801.07	897.70	−92.79	−3.84	0.05 s

5. Conclusions

In summary, an optimal scheduling model for microgrids with electric vehicles based on Deep Q-learning is proposed in this paper. Through simulation analysis under various scenarios, the following conclusions are drawn:

- As a mobile energy storage component with V2G capability, EVs can participate well in the dispatching control of the microgrid, providing a more flexible dispatching scheme for the stable operation of the microgrid.
- Compared with traditional algorithms, Deep Q-learning with online learning ability can better adapt to the strong nonlinear effects caused by the mobility of EVs, randomness of user behavior and renewable resources based on the experience accumulated in the training process. The cost of the microgrid under Deep Q-learning was 801.07 USD, and the calculation time was 0.05 s, while the total operating cost of the microgrid under the PSO algorithm was 814.57 USD, and the calculation time was 7 min 23 s. Therefore, Deep Q-learning was better than the PSO algorithm in all aspects, such as operating total costs, micro-turbine output, V2G interaction situation, grid-connected costs and operating time, which is explained in great detail in Section 4.2.

Author Contributions: Y.W., P.F. and J.H. conceptualized the idea of this research; F.W., P.F. and S.K. performed the experiments and data analysis; Y.W., J.H., X.Z. and S.K. wrote the paper; X.Z., J.H. and F.W. provided supervision and reviewed the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Postdoctoral Science Foundation [grant number 2021M702511].

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Lee, E.-K.; Shi, W.; Gadh, R.; Kim, W. Design and Implementation of a Microgrid Energy Management System. *Sustainability* **2016**, *8*, 1143. [\[CrossRef\]](#)
2. Bevrani, H.; Feizi, M.R.; Ataei, S. Robust Frequency Control in an Islanded Microgrid: H_∞ and μ -Synthesis Approaches. *IEEE Trans. Smart Grid* **2015**, *99*, 1527–1532. [\[CrossRef\]](#)
3. Li, Q.; Gao, M.; Lin, H.; Chen, Z.; Chen, M. MAS-based distributed control method for multi-microgrids with high-penetration renewable energy. *Energy* **2019**, *15*, 284–295. [\[CrossRef\]](#)
4. Chu, S.; Majumdar, A. Opportunities and challenges for a sustainable energy future. *Nature* **2012**, *488*, 294–303. [\[CrossRef\]](#)
5. Ciftci, O.; Mehrtash, M.; Marvasti, A.K. Data-Driven Nonparametric Chance-Constrained Optimization for Microgrid Energy Management. *IEEE Trans. Ind. Inform.* **2019**, *99*, 2447–2457. [\[CrossRef\]](#)
6. Askarzadeh, A. A memory-based genetic algorithm for optimization of power generation in a microgrid. *IEEE Trans. Sustain. Energy* **2017**, *9*, 1081–1089. [\[CrossRef\]](#)
7. Anh, H.P.H.; Van Kien, C. Optimal energy management of microgrid using advanced multi-objective particle swarm optimization. *Eng. Comput.* **2020**, *37*, 2085–2110. [\[CrossRef\]](#)
8. Liu, J.; Xu, F.; Lin, S.; Cai, H.; Yan, S. A Multi-Agent-Based Optimization Model for Microgrid Operation Using Dynamic Guiding Chaotic Search Particle Swarm Optimization. *Energies* **2018**, *11*, 3286. [\[CrossRef\]](#)
9. Zhu, X.; Xia, M.; Chiang, H.D. Coordinated sectional droop charging control for EV aggregator enhancing frequency stability of microgrid with high penetration of renewable energy sources. *Appl. Energy* **2018**, *210*, 936–943. [\[CrossRef\]](#)
10. Rahimi, F.; Ipakchi, A. Demand Response as a Market Resource Under the Smart Grid Paradigm. *IEEE Trans. Smart Grid* **2010**, *1*, 82–88. [\[CrossRef\]](#)
11. Bremermann, L.E.; Matos, M.; Lopes, J.A.P.; Rosa, M. Electric vehicle models for evaluating the security of supply. *Electr. Power Syst. Res.* **2014**, *111*, 32–39. [\[CrossRef\]](#)

12. Yang, J.; Zeng, Z.; Tang, Y.; Yan, J.; He, H.; Wu, Y. Load Frequency Control in Isolated Micro-Grids with Electrical Vehicles Based on Multivariable Generalized Predictive Theory. *Energies* **2015**, *8*, 2145–2164. [[CrossRef](#)]
13. Fan, P.; Ke, S.; Kamel, S.; Yang, J.; Li, Y.; Xiao, J.; Xu, B.; Rashed, G.I. A Frequency and Voltage Coordinated Control Strategy of Island Microgrid including Electric Vehicles. *Electronics* **2022**, *11*, 17. [[CrossRef](#)]
14. Tang, Y.; He, H.; Wen, J.; Liu, J. Power system stability control for a wind farm based on adaptive dynamic programming. *IEEE Trans. Smart Grid* **2015**, *6*, 166–177. [[CrossRef](#)]
15. Ruelens, F.; Claessens, B.J.; Vandael, S.; De Schutter, B.; Babuška, R.; Belmans, R. Residential demand response of thermostatically controlled loads using batch Reinforcement Learning. *IEEE Trans. Smart Grid* **2017**, *8*, 2149–2159. [[CrossRef](#)]
16. Foruzan, E.; Soh, L.K.; Asgarpoor, S. Reinforcement learning approach for optimal distributed energy management in a microgrid. *IEEE Trans. Power Syst.* **2018**, *33*, 5749–5758. [[CrossRef](#)]
17. Kofinas, P.; Dounis, A.I.; Vouros, G.A. Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids. *Appl. Energy* **2018**, *210*, 53–67. [[CrossRef](#)]
18. Sun, J.Y.; Tang, J.M.; Chen, Z.R. Multi-agent Deep Reinforcement Learning for Distributed Energy Management and Strategy Optimization of Microgrid Market—Science Direct. *Sustain. Cities Soc.* **2021**, *74*, 103163.
19. Li, P.; Hu, W.; Xu, X.; Huang, Q.; Liu, Z.; Chen, Z. A frequency control strategy of electric vehicles in microgrid using virtual synchronous generator control. *Energy* **2019**, *189*, 116389. [[CrossRef](#)]
20. Zhong, W.; Xie, K.; Liu, Y.; Yang, C.; Xie, S. Topology-Aware Vehicle-to-Grid Energy Trading for Active Distribution Systems. *IEEE Trans. Smart Grid* **2018**, *10*, 2137–2147. [[CrossRef](#)]
21. Rao, Y.; Yang, J.; Xiao, J.; Xu, B.; Liu, W.; Li, Y. A frequency control strategy for multimicrogrids with V2G based on the improved robust model predictive control. *Energy* **2021**, *222*, 119963. [[CrossRef](#)]
22. Huang, L.; Fu, M.; Qu, H.; Wang, S.; Hu, S. A deep reinforcement learning-based method applied for solving multi-agent defense and attack problems. *Expert Syst. Appl.* **2021**, *176*, 114896. [[CrossRef](#)]
23. Yang, Q.; Zhu, Y.; Zhang, J.; Qiao, S.; Liu, J. UAV Air Combat Autonomous Maneuver Decision Based on DDPG Algorithm. In Proceedings of the 2019 IEEE 15th International Conference on Control and Automation (ICCA) IEEE, Edinburgh, Scotland, 16–19 July 2019.
24. Yu, T.; Zhou, B.; Chan, K.W.; Yuan, Y.; Yang, B.; Wu, Q.H. $R(\lambda)$ imitation learning for automatic generation control of interconnected power grids. *Automatica* **2012**, *48*, 2130–2136. [[CrossRef](#)]