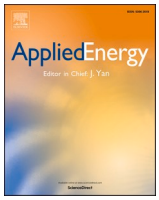
Applied Energy 328 (2022) 120140

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Deep reinforcement learning-based strategy for charging station participating in demand response

Ruiyang Jin [a], Yuke Zhou [a], Chao Lu [b], Jie Song [a,*]

[a] Department of Industrial Engineering & Management, Peking University, 100871 Beijing, China
[b] Department of Electrical Engineering, Tsinghua University, 100084 Beijing, China

HIGHLIGHTS

- The charging station is considered as a load aggregator and the participant of incentive-based demand response.
- A deep reinforcement learning-based decentralized framework is innovatively applied to charge multiple electric vehicles.
- The proposed approach trades off the revenue from DR and the user satisfaction well.
- The peak load of the charging station is shaved significantly.

ARTICLE INFO

Keywords:
Demand response
Charging station
Electric vehicle
Deep reinforcement learning

ABSTRACT

The trend of zero-carbonization has accelerated the prevalence of electric vehicles (EVs) owing to their advantages of low carbon emissions and high energy efficiency. The stochastic and high charging load of EVs results in a non-negligible challenge that may cause grid overload. A promising approach is the participation of charging stations in demand response as load aggregators by coordinating the charging power of electric vehicles. However, improper coordination of charging load may lead to unfulfilled charging demand, which would cause dissatisfaction on the demand side. In this study, the incentive-based and time-varying demand response mechanism is considered when charging stations coordinate charging of multiple EVs. A decentralized decision-making framework is innovatively applied to provide charging power of each EV. The charging process is modeled as a Markov decision process, and a virtual price is designed to help decide the charging power. Deep reinforcement learning algorithms such as deep deterministic policy gradient are applied to determine the charging strategy of multiple and heterogeneous EVs. Numerical experiments are performed to validate the effectiveness of the proposed method. A comparison with an optimal charging strategy and a heuristic rule-based method shows that the proposed method can trade off the revenue from demand response and user satisfaction, as well as reduce the peak load of the charging station. Furthermore, a test with inaccurate departure information indicates the robustness of the proposed method.

1. Introduction

Electricity is a promising alternative to fossil fuels because of the limited reserves and pollution caused by traditional energy resources. In the field of transportation, the trend of low pollution and low carbon emissions has triggered the blooming of electric vehicles (EVs) and the corresponding charging infrastructure [1]. EVs have a variety of advantages, including high energy efficiency and environment-friendliness. Global EV sales reached 6.75 million in 2021, 108% more than that in the previous year.[1] Thus, massive supporting infrastructures, such as charging poles, are being built. According to the National Energy Administration of China, the number of public charging poles in China would have reached 558,000 by the middle of 2020. In particular, in the context of carbon neutralization, EVs play an essential role in the global transportation system.

However, the high penetration and charging load of EVs pose significant challenges to the power grid [2]. The increasing number of EVs is expected to change the load profile significantly in distribution networks, especially when high numbers of EVs are charged

* Corresponding author.
E-mail address: jie.song@pku.edu.cn (J. Song).

[1] Data obtained from the website http://www.ev-volumes.com/country/total-world-plug-in-vehicle-volumes/

https://doi.org/10.1016/j.apenergy.2022.120140
Received 14 September 2021; Received in revised form 9 July 2022; Accepted 10 October 2022
Available online 29 October 2022

Nomenclature

Abbreviations

CS	Charging Station
DCS	Deep reinforcement learning-based Charging Strategy
DDPG	Deep Deterministic Policy Gradient
DR	Demand Response
DRM	DSR-Ranking Method
DSR	Demand Satisfaction Rate
EV	Electric Vehicle
GO	Grid Operator
OCS	Optimal Charging Strategy
PHEV	Plug-in Hybrid Electric Vehicle
ReL	Reference Load
RL	Reinforcement Learning
VP	Virtual Price

Parameter and variables

α	DR incentive per kWh shaved load
L_t	Total charging load of the CS without charging coordination
$L_{ref,t}$	Reference load
$L_{real,t}$	Real load of the CS after charging management
$L_{ave,t}$	Average charging load at time t before load management
$P_{i,k}^{max}$	Rated charging power of the k th EV connected to the i th charging pole
$d_{i,k}$	Charging demand of the k th EV connected to the i th charging pole
$t_{i,k}^{arr}$	Arrival time of the k th EV connected to the i th charging pole
$t_{i,k}^{dep}$	Departure time of the k th EV connected to the i th charging pole
Δt	Time interval
$R_{DR,t}$	Revenue obtained by the CS from GO for DR at time t
VP_t	Virtual charging price provided by the CS at time t
$F_{cs,t}$	Ratio of the total fulfilled charging demand in the CS
n	Number of charging poles
$P_{i,s}^{real}$	Real charging power of the i th EV at time s
$DSR_{i,t}$	Fraction of total required kWh of charging of the EV connected to the i th charging pole that has been realized by time t
$c_{i,t}$	Duration of parking time up to time t for the i th charging pole
$l_{i,t}$	Left time for charging for the i th charging pole
$v_{i,t}$	Average charging intensity of the i th charging pole at time t
$J_{i,t}$	Total number of EVs connected to the i th charging pole by time t
$K_{i,t}$	Index for the EV connected to the i th charging pole at time t
$a_{i,t}$	The action of the i th agent
$is_{i,t}^{end}$	Indicator of whether the charging process of the i th EV is completed at time t
$cost_{i,t}$	Cost function of the i th agent at time t
$r_{i,t}$	Reward function of the i th agent at time t
β	Price coefficient
μ	Current actor network
μ'	Target actor network
Q	Current critic network
Q'	Target critic network
θ^{μ}	Parameters of current actor network
θ^{Q}	Parameters of target actor network
$\theta^{\mu'}$	Parameters of current critic network
$\theta^{Q'}$	Parameters of target critic network
s_h	Current state
a_h	Current action
r_h	Current reward
s'_h	Next state
τ	Target smoothing coefficient
γ	Discounted factor
σ_e	Variance of exploration noise
T	Total training steps

Indices

i	Index for charging poles
j,k	Index for EVs connected to charging pole
t,s	Index for time

simultaneously, which can cause voltage stability and power equality problems, transformer losses, and reduced operation lifespan of the grid [3-5] . Therefore, uncontrolled EV charging may place the power grid in a critical situation.

It is more efficient to improve the security of the electricity supply by managing te demand-side load rather than extending the capacities of the power grid [6]. Owing to the development of new technologies such as 5G/6G and IoT [41], demand response (DR) is a potentially attractive approach to deal with the critical situation posed by EV charging by adjusting the electricity price or providing incentives to regulate the charging load and alleviate the pressure on the power grid [7,8]. DR plays an important role in the electricity market to maintain the balance between supply and demand by introducing load flexibility instead of adjusting only the generation levels [9]. More detailed information about the type and mechanism of DR can be found in [7,10]. The prototype application of the problem investigated here is the charging of EVs at a charging station (CS). In such applications, customers have their EVs charged during parking time in the CS. The CS acts as a load aggregator and regulates the charging load of each EV to fulfill the charging demand and participate in DR. The vital fact enabling the CS to regulate the charging of EVs is that most customers leave their EVs in CS longer than the time needed for charging, which provides flexibility to change the load profile of the CS, as confirmed on real-world dataset in [11].

The objective of the EV charging problem can be categorized into three types, that is, maximizing the benefit or minimizing the effect of EV charging for the grid[12-14], maximizing the benefit or promoting service satisfaction for the CS [15-20], and minimizing charging costs or maintaining the health of batteries for the EV [21-24]. Generally, there are two types of solutions, which are respectively based on traditional optimization methods and learning-based optimization methods. The authors of [8,25] summarized recent works in the area of traditional optimization algorithms for charging EVs in a smart grid. Various approaches and applications have been investigated, where linear programming, quadratic programming, dynamic programming, game theory, and so on are applied to solve the problem [26-29]. In addition, many heuristic intelligent optimization methods, such as genetic algorithms [30], evolutionary algorithms [31], particle swarm optimization algorithms [32], and simulated annealing algorithms [33], have been investigated. However, most of these studies assume that certain knowledge such as future EV arrivals and electricity prices are already known or obey some distribution. Nevertheless, such assumptions cannot be satisfied in most practical scenarios. Even though the knowledge of distribution can be obtained, the distribution of future

events may vary with time, which makes it challenging to capture the uncertainties. Learning-based methods, especially reinforcement learning (RL) methods, have emerged as effective tools to cope with such challenges. The RL algorithms and modeling techniques used for DR are reviewed in [34]. More details about RL techniques can be found in [35].

RL-based methods have been widely applied to the EV charging problem for DR. The EV charging problem is formulated as a Markov decision process (MDP) in most related studies. The literature closely related to this study using RL techniques to solve the EV charging problem is listed in Table 1 including our work. The deep deterministic policy gradient (DDPG) method was used in [12] to find the optimal EV charging strategy that maximizes the profit of the grid operator (GO) and satisfies all the physical constraints. The authors of [13] proposed a charging coordination system based on RL to create charging schedules for an EV fleet to avoid grid overload. The authors of [15,16] applied a fitted Q-learning algorithm to learn the optimal charging policy based on the MDP formulation and a new state representation method. In [17], an RL-based framework was proposed to learn the optimal charging price that obtains the maximum long-term revenue of the CS as well as social welfare. An RL approach was proposed for optimizing the charging scheduling and pricing strategies that maximize the objective of the CS in [18]. In[21], the problem was cast as a daily charging decision-making problem for planning the energy to be charged in the plug-in EV battery within a day to reduce the charging cost based on the forecasted price.

Multiagent reinforcement learning (MARL) is a decentralized model to solve the problem involving multiple agents. In [14], reducing the energy costs and avoiding transformer overload were both considered, and a multiagent RL architecture was proposed to balance these two objectives. The authors of [20] proposed a novel multiagent deep RL method for the energy management of distributed EV CSs with a solar photovoltaic system and an energy storage system.

In this study, the CS is considered as an independent aggregator for DR and determines the charging of multiple EVs. The CS receives a series of time-varying DR signals of reference load (ReL) from the GO at the beginning of discrete time intervals. The CS can obtain incentives if it adjusts the aggregated charging load under the time-varying ReL. Otherwise, the incentive is reduced due to insufficient regulation [36]. The key problem is to maximize the revenue of the CS by obtaining the incentive from the GO while fulfilling the charging demand of EVs.

To the best of our knowledge, there have not been RL-based works considering that CS participates in the incentive-based and time-varying DR as an independent aggregator. The applied DR mechanisms in related works are price-based or relatively simplified, such as flattening the aggregated load, adjusting demands by dynamic prices. However, the incentive-based DR and time-varying signals pose new challenges for CS to coordinate charging of multiple EVs. The CS needs to coordinate charging of multiple heterogeneous EVs to satisfy a global time-varying objective. Most related works apply a centralized RL-based coordinator to decide the charging power of all EVs. The coordinator can provide charging power of EVs to satisfy the time-varying signal but can hardly work due to the curse of dimensionality caused by continuity and scale of the state and action spaces [15,37]. MARL is an attractive decentralized framework for learning the charging decisions of multiple agents. However, MARL is still limited by the scalability of EVs because dynamic environment may lead to unstable training process [13]. Furthermore, a large number of agents requires too much space and more samples to compute a good policy [14].

To solve the EV charging problem in CS to participate in incentive-based and time-varying DR, a decentralized decision-making framework is innovatively applied in which individual charging pole decides the charging power of connected EV based on virtual price (VP). The MDP model of the EV charging process is formulated, and the deep RL algorithm, DDPG, is applied to learn the charging strategy and decide the charging power at fixed time intervals, where each charging pole is seen as an agent to charge the connected EV. All the agents share the same parameters of single-agent RL algorithm to obtain a linear increase in computational complexity with the increasing scale of EVs. A VP is proposed to realize the consistency of the targets of the CS and all agents. The contributions of this study can be summarized as follows.

1) The incentive-based and time-varying DR mechanism is considered when CS coordinates the charging of multiple EVs as a load aggregator. The objective of CS is to maximize the revenue from DR and the user satisfaction.

2) A decentralized decision-making framework is innovatively applied to provide charging decisions by each charging pole. The framework can enable the scalability and the ability to deal with large scale of EVs. The VP is designed to realize the coupling of the reward of individual agent and the objective of CS, as well as the trade-off between revenue from DR and the user satisfaction.

3) Deep RL algorithm is applied to learn the charging strategy of EVs without the information of future arrivals of EVs and ReL. A numerical experiment based on real-world data is carried out to validate the effectiveness of the proposed method.

The remainder of this paper is organized as follows. The decentralized decision-making framework and MDP formulation of EV charging is introduced in Section 2. The algorithm based on DDPG is proposed to solve the charging problem in Section 3. In Section 4, a numerical experiment is performed to validate the effectiveness and robustness of the proposed method. Finally, the conclusions are drawn in Section 5.

2. Model formulation

2.1. Demand response formulation

For the GO, it is necessary to manage the high load of EVs, especially during the high-load period in case of the overload of the power grid. One of the potential approaches is to offer incentives to the CS to shave the peak load, as shown in Fig. 1. Note that the GO requires the load of the CS to stay below a ReL $L_{ref,t}$ at time t by offering a fixed incentive which is α per kW for the shaved load. The CS acting as a load aggregator can determine the charging load of each EV to track the DR signal. Before introducing the DR mechanism, once an EV arrives at the CS, the EV is charged with a fixed charging power. However, some EVs are left in the CS for a longer time than they need to be fully charged, which enables the CS to coordinate the charging load to participate in the DR.

2.2. Charging station formulation

The additional income of incentives from participating in DR encourages the CS to apply charging management. However, it is challenging to optimize the charging strategy without sufficient information, such as future EV arrival, parameters of future EVs, and future ReL. Thus, the CS can only make charging decisions based on historical and present information. For a CS with n charging poles, the load at time t without any management is the sum of the rated charging powers of the EVs, which is written as

$$L_t = \sum\nolimits_{i=1}^{n} P_{i,K_{i,t}}^{max} \tag{1}$$

where L_t is the total charging load of the CS without charging coordination, $P_{i,k}^{max}$ is the rated charging power of the k th EV connected to the i th charging pole, and $P_{i,0}^{max} = 0$. $K_{i,t}$ is the index for the EV connected to the i th charging pole at time t, which means that the $K_{i,t}$ th EV is connected to the i th charging pole at time t and $K_{i,t} = 0$ if the i th charging pole is unoccupied at time t. Similarly, $J_{i,t}$ is defined to represent the total number of EVs connected to the i th charging pole by time t. From the perspective of the CS, the objective is to coordinate the charging load of all the EVs to track the DR signal delivered by the GO and obtain the highest revenue while considering the user satisfaction. The incentive is decided by the real load after management, the ReL, and the uncontrolled load without management. The CS obtains nothing if

Table 1
Existing related works using reinforcement learning for EV charging.

Related work	Perspective / Objective'	Method	Demand response form	Experimental settings	Decision framework
[12]	GO / Max. profit	DDPG	No	Based on real-world data	Centralized
[13]	GO / Min. load variance	DQN	No	Using manually defined data	Decentralized
[14]	GO / Min. energy cost and avoid overload	MARL	No	Using manually defined data	Decentralized
[15,16]	CS / Min. load variance	Fitted Q-iteration	Flatten charging load	Based on real-world data	Centralized
[17]	CS / Max. revenue and social welfare	Q-learning	Adjust demand by price	Based on real-world data	Centralized
[18]	CS / Max. profit	SARSA	No	Using manually defined data	Centralized
[19]	EV / Min. cost	Fitted Q-iteration	Adjust demand according to price	Using manually defined data	Centralized
[22]	EV / Min. cost	Q-learning	Adjust demand according to price	Using manually defined data	Centralized
[23]	EV / Min. cost	DDPG	No	Using manually defined data	Centralized
Our work	CS / Max. revenue and user satisfaction	DDPG	Respond to incentive-based time-varying DR signal	Based on real-world data	Decentralized

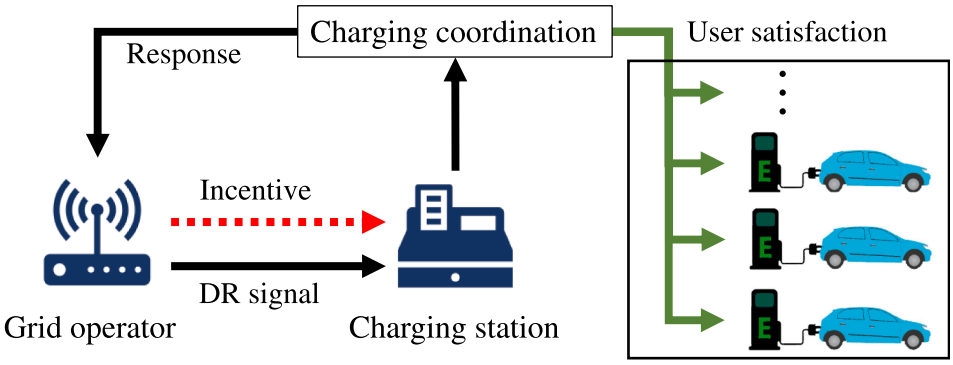


Fig. 1. Structure of the demand response (DR) model.

its uncontrolled load is lower than the ReL. If the ReL is lower than the uncontrolled load and the CS manages to shave the load to no higher than the ReL by coordinating the charging of EVs, it can obtain all the incentives. Otherwise, only the incentives for the shaved load can be obtained by the CS. Furthermore, the CS is punished if the real load is above the ReL, whereas the uncontrolled load is lower than the ReL. The detailed revenue mechanism is shown in Fig. 2. The reward of the CS participating in DR can be written as

$$R_{DR,t} = \alpha\left(\max\left(L_t, L_{ref,t}\right) - \max\left(L_{ref,t}, L_{real,t}\right)\right) \quad (2)$$

Here, R_{DR} is the direct revenue obtained by the CS from the GO for participating in DR without the loss of user satisfaction, and $L_{real,t}$ is the real load of the CS after charging management. From (1) and (2), the management of load L_t has a significant influence on the following charging load profile by delaying or moving up the charging load, probably changing the calculation of revenue significantly. In reality, the uncontrolled load L_t is usually estimated with the historical data in non-DR days before time t. Therefore, the L_t in (2) is generally replaced by the average charging load at time t before load management, written as $L_{ave,t}$. It is also noteworthy that the revenue mechanism used here can be expanded to other forms related to the curtailed load.

Furthermore, the user satisfaction is measured by demand satisfaction rate (DSR). The DSR here refers to the fraction of total required kWh of charging of the connected EV that has been realized when EV leaves. The key problem is how to trade off the revenue from DR and the user satisfaction. To achieve this, a virtual charging price is proposed, which is used for coordinating the charging of multiple EVs. Assume that the CS provides a VP at time t, denoted as VP_t, when a ReL is received from the GO, and each charging pile provides charging power according to VP_t to achieve the minimum charging cost. Note that the charging price and cost are virtually set up. They are introduced only for converting the incentive-based DR problem to a virtual price-based DR problem and to help derive the charging strategy.

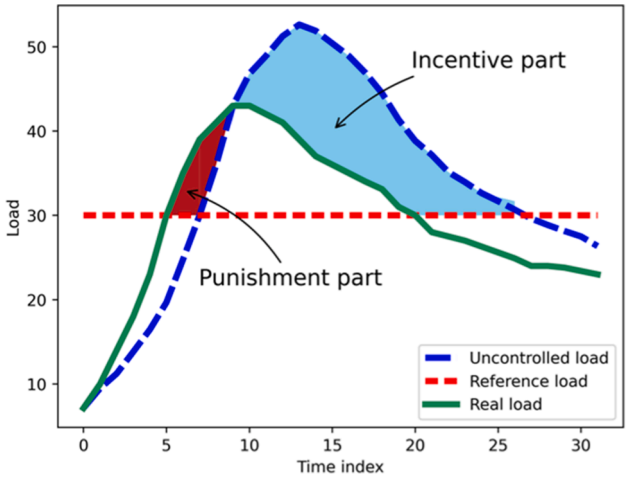


Fig. 2. Revenue that the charging station (CS) obtains from DR.

The objective is to reach a balance between DR revenue and the user satisfaction, so VP_t should include information related to both. The revenue from DR can be reflected from the ratio of real load to ReL, while the satisfaction of the demand side is related to the ratio of total fulfilled charging demand.

When the uncontrolled load of the CS is lower than the ReL, no regulation is needed and $VP_t = 0$. On the contrary, the VP should be larger to restrain the charging load when the CS needs to curtail the charging load. Thus, the VP at time t is given as

$$VP_t = \begin{cases} 0, \ if \ L_t \le L_{ref,t} \\ \dfrac{L_t \bullet F_{cs,t}}{L_{ref,t}\left(2 - F_{cs,t}\right)}, \ otherwise \end{cases} \tag{3}$$

$F_{cs,t}$ is the ratio of the total fulfilled charging demand in the CS, and it can be expressed as

$$F_{cs,t} = \frac{\sum_{i=1}^{n}\sum_{s=t_{i,K_{i,t}}^{arr}}^{t} P_{i,s}^{real} \bullet \Delta t}{\sum_{i=1}^{n} d_{i,K_{i,t}}} \tag{4}$$

where $P_{i,s}^{real}$ is the real charging power of the i th EV at time s and Δt is the time interval. $d_{i,k}$ is the charging demand of the k th EV connected to the i th charging pole and $d_{i,0} = 0$. $t_{i,k}^{arr}$ is the arrival time of the k th EV connected to the i th charging pole and $t_{i,0}^{arr} = 0$.

Obviously, the VP is higher when the ReL is lower, given EVs in the CS. Furthermore, the VP is lower when $F_{cs,t}$ is lower, which indicates a higher risk of failure to meet the charging demand, and EVs are encouraged to charge with higher power.

2.3. Markov decision process model of electric vehicle charging

When an EV arrives at the CS, it will get connected to the charging pole if there is unoccupied charging pole. When the EV is connected to a charging pole, we assume that the charging pole will get the information about the EV, i.e., arriving time, departure time, rated charging power and charging demand. However, the charging station cannot gain information about the future arrivals of EVs. In this section, the charging process of EV is formulated as an MDP model. The charging poles are regarded as agents that coordinate the charging power of the connected EVs. The MDP formulation is composed of four parts: state, action, reward, and transition function. The framework of the MDP model is shown in Fig. 3.

2.3.1. State

The state of the agent describes the current information of the connected EV and the VP from the environment, which consists of the current time t, virtual price VP_t, DSR of i th charging pole at time t $DSR_{i,t}$, duration of parking time up to time t in the CS $c_{i,t}$, left time for charging $l_{i,t}$, and average charging intensity $v_{i,t}$. $DSR_{i,t}$ is defined as

$$DSR_{i,t} = \frac{\sum_{s=t_{i,K_{i,t}}^{arr}}^{t} P_{i,s}^{real} \bullet \Delta t}{d_{i,K_{i,t}}}, \ for \ K_{i,t} \neq 0 \tag{5}$$

The duration of parking time and time left for charging are defined as

$$c_{i,t} = t - t_{i,K_{i,t}}^{arr}, l_{i,t} = t_{i,K_{i,t}}^{dep} - t, \ for \ K_{i,t} \neq 0 \tag{6}$$

where $t_{i,k}^{dep}$ is the departure time of the kth EV connected to the i th charging pole. Furthermore, $v_{i,t}$ is used to describe the relative average charging power compared with the maximum charging level, which is expressed as

$$v_{i,t} = \frac{\sum_{s=t_{i,K_{i,t-1}}^{arr}}^{t-1} P_{i,s}^{real}}{P_{i,K_{i,t-1}}^{max} \bullet c_{i,t-1}}, \ for \ K_{i,t-1} \neq 0 \tag{7}$$

To deal with heterogeneous EVs, the normalization in (5) and (7) can model the charging intensity of EVs with different demands and charging magnitudes. The entire state of the i th agent at time t can be denoted as $S_{i,t} = [t, VP_t, DSR_{i,t}, c_{i,t}, l_{i,t}, v_{i,t}]$. Among all the information, VP_t is the global information observable by all agents. Other than VP_t, the local EV information is only available to the connected agent. All the information is attainable and accurate, except for the departure time of EVs. The departure information can be set beforehand by the EV owners or predicted by contextual information, but it may be inaccurate owing to the uncertainty of user behavior. The effect of inaccurate departure time is analyzed later in the case study.

2.3.2. Action

The action of the i th agent $a_{i,t}$ is the charging intensity coefficient given by the agent, ranging from 0 to 1. The real charging power for the connected EV is a continuous variable and is expressed as

$$P_{i,t}^{real} = a_{i,t} \bullet P_{i,K_{i,t}}^{max} \tag{8}$$

Although practical charging poles may only provide discrete charging power, it is straightforward to expand the continuous action to a discrete one.

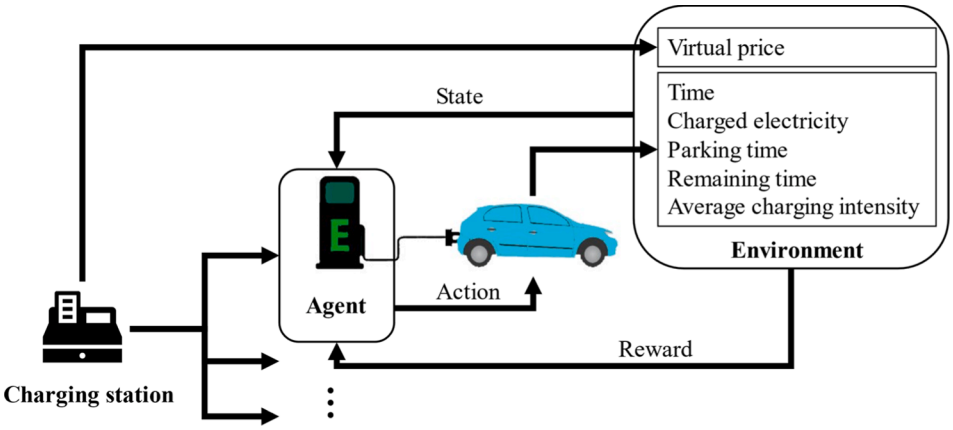


Fig. 3. Framework of the Markov decision process (MDP) model of the charging process.

2.3.3. Reward

The design of the reward function is the key factor that enables the agents to cooperate to achieve the goal of the CS. Here, the objective of the agent is to have the connected EV charged at the lowest cost without affecting user satisfaction. Once $DSR_{i,t} = 1$, say, the EV connected to the i th pole is fully charged at time t or it departs at time t, a cost will be assigned at time t. The cost contains (i) the average VP during its charging process, and (ii) unsatisfied charging demand upon departure. Consequently, the cost of the i th charging pole at time t can be designed as

$$cost_{i,t} = \begin{cases} 0, \ if \ is_{i,t}^{end} = 0; \\ \beta \bullet \dfrac{\sum_{s=t_{i,K_{i,t}}^{arr}}^{t} VP_s \bullet P_{i,s}^{real}}{\sum_{s=t_{i,K_{i,t}}^{arr}}^{t} P_{i,s}^{real}} + \left(1 - DSR_{i,t}\right), \ if \ is_{i,t}^{end} = 1; \end{cases} \tag{9}$$

Therefore, the reward function of the i th agent is designed as

$$r_{i,t} = \begin{cases} 0, \ if \ is_{i,t}^{end} = 0; \\ -\beta \bullet \dfrac{\sum_{s=t_{i,K_{i,t}}^{arr}}^{t} VP_s \bullet P_{i,s}^{real}}{\sum_{s=t_{i,K_{i,t}}^{arr}}^{t} P_{i,s}^{real}} - \left(1 - DSR_{i,t}\right), \ if \ is_{i,t}^{end} = 1; \end{cases} \tag{10}$$

where $is_{i,t}^{end}$ is an indicator of whether the charging process is completed. $is_{i,t}^{end} = 1$ when the i th connected EV is fully charged or the EV departs, and $is_{i,t}^{end} = 0$, otherwise. This definition indicates that the agent obtains nothing when the charging process is incomplete. The final reward when charging is finished consists of two parts. The former part is proportional to the average price, and the latter is the opposite of the ratio of the unfinished charging demand. $\beta \geq 0$ is the price coefficient for trading the two parts.

Owing to the structure of the reward function, the charging load of each EV can be coordinated by the CS through the VP. When the VP is high, the agent will provide a low charging power to maximize the former part related to the virtual charging price in the reward function. Thus, the total charging load of the CS can be reduced accordingly to track the ReL. With virtual price given by the charging station, each agent can gain access to the global information that indicates the ReL and the global charging state, which lays restriction on individual agent's decision. Since each agent wants to trade off the global DR revenue and user satisfaction, it combines both global and individual information to obtain optimal decision. The price coefficient β plays the role of balancing the revenue obtained from participating in DR and the satisfaction of the demand side. The agent prioritizes regulating the charging power to track the ReL and obtains more incentives when β is large. Through this reward design, the consistency of the agent's reward and the goal of the CS is achieved.

2.3.4. Transition function

After the agent decides the charging power of the i th connected EV, the EV is charged with $P_{i,t}^{real}$ for the time interval Δt from t to $t+1$, which drives the environment to transit into a new state. Despite the effects of actions executed by agents, the new state of the environment is also affected by the arrival and departure of EVs and the ReL.

At the beginning, all charging poles are empty thus $S_{i,0}$ could be initialized as $[t = 0, VP_0 = 0, DSR_{i,0} = 1, c_{i,o} = 0, l_{i,0} = 0, v_{i,0} = 0]$. The real power load of the charging pole will be 0 when it is unoccupied or the charging process is finished, and the DSR could be viewed as 1. Both $J_{i,0}$ and $K_{i,0}$ are 0 for there has been no EVs connected to the charging poles.

If an EV is disconnected from the charging pole or fully charged, the charging process is completed, and its information is removed. The elements of $DSR_{i,t+1}, c_{i,t+1}, l_{i,t+1}, v_{i,t+1}, J_{i,t+1}$ and $K_{i,t+1}$ are updated as

$$\begin{cases} DSR_{i,t+1} = 1 \\ c_{i,t+1} = 0 \\ l_{i,t+1} = 0 \\ v_{i,t+1} = 0 \\ J_{i,t+1} = J_{i,t} \\ K_{i,t+1} = \begin{cases} 0 \ , \ if \ EV \ is \ disconnected \\ K_{i,t} \ , \ if \ EV \ is \ fully \ charged \end{cases} \end{cases} \tag{11}$$

and $is_{i,t}^{end}$, which is usually defaulted to 0, will be set to 1.

If an EV arrives and is connected to the i th charging pole, the information of the EV is added to the observation as

$$\begin{cases} DSR_{i,t+1} = 0 \\ c_{i,t+1} = 0 \\ l_{i,t+1} = t_{i,K_{i,t+1}}^{dep} - t_{i,K_{i,t+1}}^{arr} \\ v_{i,t+1} = 0 \\ J_{i,t+1} = J_{i,t} + 1 \\ K_{i,t+1} = J_{i,t+1} \end{cases} \tag{12}$$

If the charging of the i th EV in the CS is not finished, the elements of $DSR_{i,t+1}, c_{i,t+1}, l_{i,t+1}$ and $v_{i,t+1}$ are updated as

$$\begin{cases} DSR_{i,t+1} = DSR_{i,t+1} + \dfrac{P_{i,t}^{real} \bullet \Delta t}{d_{i,K_{i,t}}} \\ c_{i,t+1} = c_{i,t} + 1 \\ l_{i,t+1} = l_{i,t} - 1 \\ v_{i,t+1} = \left(1 - \dfrac{1}{c_{i,t}}\right) v_{i,t} + \dfrac{1}{c_{i,t}} \dfrac{P_{i,t}^{real}}{P_{i,K_{i,t}}^{max}} \\ J_{i,t+1} = J_{i,t} \\ K_{i,t+1} = K_{i,t} \end{cases} \tag{13}$$

After receiving the new reference and information on EVs, VP_{t+1} is updated according to (3). The new VP_{t+1} is given by the CS such that the new state at time $t+1$ is $S_{i,t+1} = [t+1, VP_{t+1}, DSR_{i,t+1}, c_{i,t+1}, l_{i,t+1}, v_{i,t+1}]$.

3. Deep reinforcement learning-based charging strategy

DDPG is an actor-critic model-free RL algorithm based on a deterministic policy gradient and neural network function approximator. Many techniques are combined in the DDPG, such as replay buffer, soft target updates, and batch normalization. More details about the DDPG can be found in [38]. In this section, the deep reinforcement learning-based charging strategy (DCS) with DDPG applied is introduced to provide the charging power of each EV.

Generally, there are four neural networks used in the DDPG: current actor network μ, current critic network Q, target network μ', and Q'. Their parameters are $\theta^{\mu}, \theta^{Q}, \theta^{\mu'}$ and $\theta^{Q'}$, respectively. The actor network is used to provide the charging action, and the critic network can estimate the value of the action. Thus, the input of the actor network is the state, and the output is the action. The input of the critic network is the state and the provided action, and the output is the estimated value of the state and action. The actor and critic both provide the action or evaluation by a nonlinear neural network.

The current actor is used for selecting the current charging action $a_{i,t} = \mu\left(s_{i,t}|\theta^{\mu}\right) + \mathcal{N}_{i,t}$ given the current state, and $\mathcal{N}_{i,t}$ is the noise sampled from a random process to construct exploration policy. In this

study, a normal distribution $\mathcal{N}(0,\sigma_e{}^2)$ is used to create noise. The current critic is set to judge the performance of the actor by evaluating the value of action provided by the current actor in the current environmental state, that is, $Q(s_{i.t}, a_{i.t}|\theta^Q)$. The evaluated value from the critic dominates the update of the actor network, and thus the training of the critic is a key factor affecting the performance of the DDPG. RL involves sequential samples generated from exploring the environment, so it may not satisfy the independent and identical distribution assumption which is important for most algorithms. A replay buffer is introduced in the DDPG to enable learning from uncorrelated samples. Moreover, the setting of target networks is in the case of unstable training, and their parameters are copied from current networks regularly. Each time the reward and new state are returned after the current actor provides the action, a sample of the transition is stored in the replay buffer, including the current state $s_{i,t}$, action $a_{i,t}$, reward $r_{i,t}$, next state $s_{i,t+1}$, and the indicator $is_{i.t}^{end}$. N transitions are sampled randomly from the replay buffer when updating the current networks, that is, $(s_h, a_h, r_h, s_h^{'}, is_h^{end}), h = 1,2,\cdots,N$, where the index h indicates the h th sample. Then, the action and evaluation of the action are derived from the target network regarding the next state $s_{i,t+1}$ as the current state. The loss function to be minimized to update the current critic network is

$$L = \frac{1}{N}\sum_h \left(y_h - Q(s_h, a_h|\theta^Q)\right)^2 \tag{14}$$

where y_h is the estimated value from the target networks and it is written as

$$y_h = r_h + \left(1 - is_h^{end}\right)\gamma Q^{'}\left(s_h^{'}, \mu^{'}(s'_h|\theta^{\mu^{'}})|\theta^{Q^{'}}\right) \tag{15}$$

where γ is the discount factor, which ranges from 0 to 1. $Q^{'}(\bullet|\theta^{Q^{'}})$ is the estimated value given by the target critic network, and $\mu^{'}(\bullet|\theta^{\mu^{'}})$ is the action given by the target actor network. The current actor network is updated using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_h \nabla_a Q(s, a|\theta^Q)_{s=s_h, a=\mu(s_h)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)_{s_h} \tag{16}$$

The target networks are updated by the current networks using "soft" update as

$$\begin{cases} \theta^{Q^{'}} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q^{'}} \\ \theta^{\mu^{'}} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu^{'}} \end{cases} \tag{17}$$

where τ is the smoothing coefficient for updating, which constrains the target values to change slowly, greatly improving the stability of learning.

The entire framework of the DDPG is shown in Fig. 4, where the red lines indicate the interaction with the environment and the blue lines indicate the updating process of the networks. The detailed procedure of the DCS is presented in **Algorithm 1**.

Algorithm 1 DCS with DDPG

Randomly initialize actor network μ and critic network Q.
Initialize target network μ' and Q'.
Initialize replay buffer R.
Initialize a random process $\mathcal{N}$ for action exploration.
Receive initial observation of each charging pole in the CS, i.e., $s_{i,1}, i = 1,2,\cdots,n$.
for $t = 1 : T$ **do**
for $i = 1 : n$ **do**
Select charging action $a_{i.t} = \mu(s_{i.t}|\theta^\mu) + \mathcal{N}_{i.t}$ according to current policy and exploration noise. ①
Execute charging action $a_{i.t}$, observe reward $r_{i.t}$, new state $s_{i.t+1}$, and $is_{i.t}^{end}$.
Store transition $(s_{i.t}, a_{i.t}, r_{i.t}, s_{i.t+1}, is_{i.t}^{end})$ in R. ②
end for
Sample a minibatch of N transitions $(s_h, a_h, r_h, s_h^{'}, is_h^{end}), h = 1, 2, \cdots, N$ from R. ③
Compute y_h according to (15). ④
Update critic by minimizing the loss in (14). ⑤
Update the actor policy using the sampled policy gradient according to (16). ⑥
Update the target networks according to (17). ⑦
end for

(*continued on next column*)

(*continued*)

Each time the CS needs to coordinate the charging load, each agent provides the charging power according to the same algorithm and parameters. All updating procedures are executed with the same parameters when interacting with the environment. Such settings guarantee a linear computational burden with the scale of charging poles and the ability to deal with scalable and heterogeneous EVs. Besides, DDPG is only a basic algorithm that can deal with continuous states and actions, and can be embedded in the charging framework. Any other proper deep RL algorithm can also be applied to learn the charging strategy in the proposed method, which is addressed in the following analysis.

4. Numerical experiment

4.1. Simulation setting

In this section, a case study based on real charging data is conducted to validate the model and method proposed above. The dataset consists of over 28,000 charging records in Los Angeles, USA, from April 2018 to December 2020.[2] Each record includes the arrival time, departure time, charging time, and delivered electricity of the EV. The distribution of arrival rate and departure rate over time according to the dataset is shown in Fig. 5. A CS with 30 charging poles is simulated, and the time interval is set to 15 min. The arrivals of EVs are simulated according to the distribution of all arrival times in the charging records. The ReL from the GO is uniformly distributed in the simulation. The average of the ReL and the average uncontrolled load of the CS before participating in the DR in one day is shown in Fig. 6, where the red band represent the uniform distribution that the reference load is sampled from and the blue band manifests the variance of the uncontrolled load. This indicates that the curtailment of the charging load is mainly required from 15:00 to 20:00 in one day. It is also assumed that the rated charging power of all EVs is lower than the maximum output charging power of charging poles. Note that there is no assumption regarding the charging behavior and ReL in the model formulation. The distribution of the arrival/departure rate, future demand, and ReL are not known beforehand in the proposed method.

The neural networks of the actor and critic are set with two hidden layers and 64 neurons in each layer. The parameters of the DDPG algorithm and the scenario are set as presented in Table 2.

There is no acknowledged baseline method in such a scenario, and most of the existing methods cannot deal with the charging problem of multiple and heterogeneous EVs in the DR context. To validate the effectiveness of the proposed method, an optimal charging strategy (OCS) and a heuristic DSR-ranking method (DRM) are introduced. In the OCS, it is assumed that all the future information, that is, the arrival, departure, demand of EVs, and ReL, are perfectly known to derive an optimal charging scheduling. The charging problem to be solved is

$$\min \sum_i \sum_{j=1}^{J_{i,T_0}} \left(d_{i,j} - \sum_{s=t_{i,j}^{arr}}^{t_{i,j}^{dep}} P_{i,s}^{real}\Delta t\right) \tag{18}$$

$$s.t. \sum_{s=t_{i,j}^{arr}}^{s=t_{i,j}^{dep}} P_{i,s}^{real}\Delta t \le d_{i,j}, \forall 1 \le j \le J_{i,T_0}, \forall i$$

[2] The data can be accessed on the website: https://ev.caltech.edu/dataset

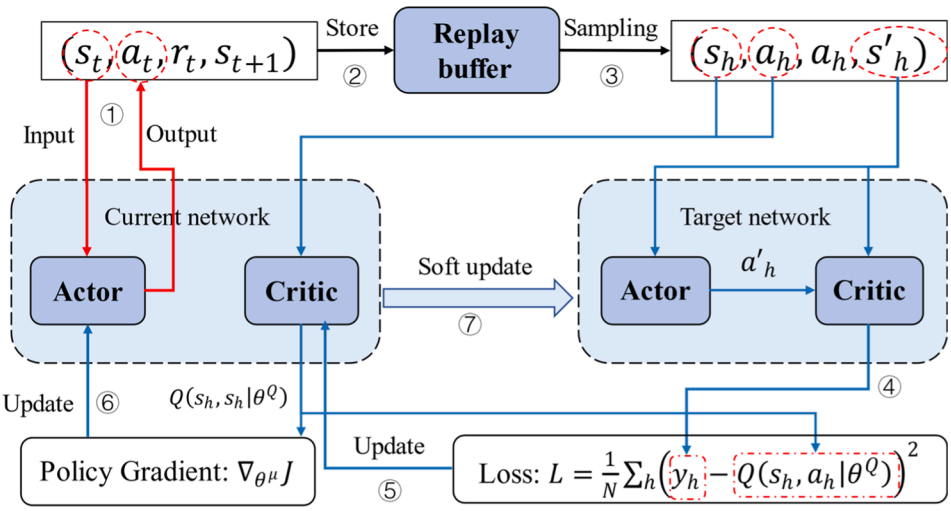


Fig. 4. Framework of deep deterministic policy gradient (DDPG) algorithm.

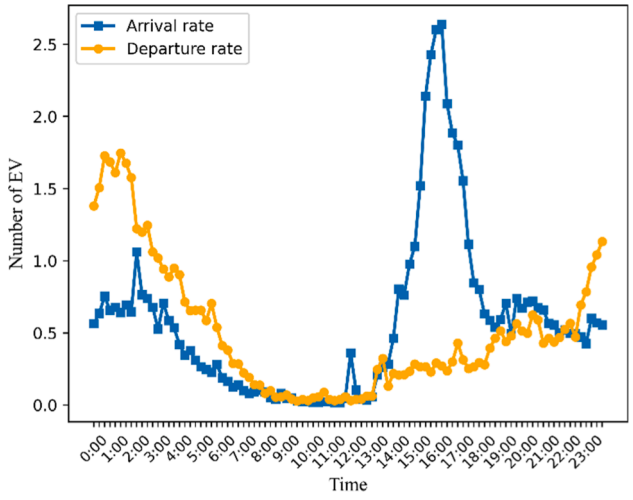


Fig. 5. Distribution of the arrival rate and the departure rate according to the dataset.

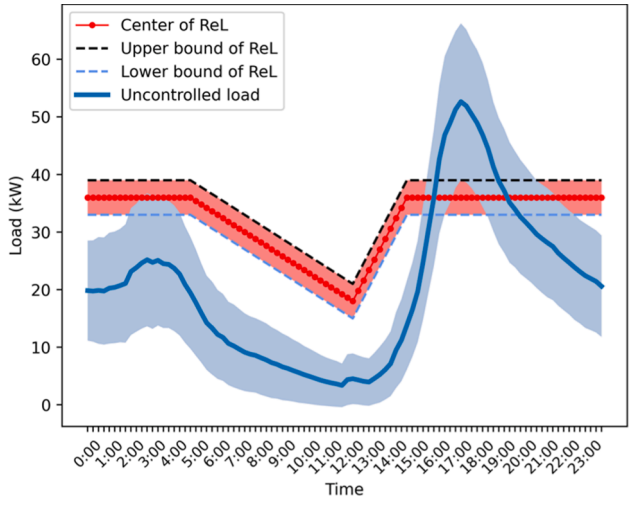


Fig. 6. Distribution of the ReL provided by the grid operator (GO) and uncontrolled average load of the CS.

Table 2
Setting of parameters in the DDPG algorithm and scenario.

Parameters (DDPG)	**Value**
Learning rate	0.0001
Target smoothing coefficient τ	0.95
Discounted factor γ	0.99
Replay buffer size	200,000
Batch size	512
Variance of exploration noise σ_e	0.05
Total training steps T	500,000
Parameters (Scenario)	**Value**
Incentive α	2
Price coefficient β	1, 2, 3
Number of charging poles n	30
Time interval Δt	15 min

$$\sum_i P_{i,t}^{real} \le L_{ref,t}, \forall t$$

$$0 \le P_{i,t}^{real} \le P_{i,K_{i,t}}^{max}, \forall i, t$$

where T_0 is the test steps. In problem (18), the ReL is seen as a constraint of the total charging load in the CS, and the objective is to minimize the unsatisfied charging demand. Problem (16) is solved using Gurobi 9.1.2.

In the DRM, the charging power of each EV is determined by its DSR, without future information. The strategy puts a higher charging priority on the EV with a lower DSR. The charging power of each EV is determined by the rank of its DSR.

The order statistics $DSR_{(1),t}, DSR_{(2),t}, \cdots, DSR_{(n),t}$ are defined as a permutation of $DSR_{1,t}, DSR_{2,t}, \cdots, DSR_{n,t}$ such that $DSR_{(1),t} \le DSR_{(2),t} \le \cdots \le DSR_{(m),t} < 1 = DSR_{(m+1),t} = DSR_{(n),t}$, which means that m EVs are in need of charging at time t. Accordingly, the rated charging power of the charging pole with j th lowest DSR will be defined as $P_{(j),t}^{max}$. The k EVs with the lowest DSRs can be charged with the rated power, wherein k satisfies

$$\sum_{j=1}^{k} P_{(j),K_{(j),t}}^{max} \le L_{ref,t}, \ \sum_{j=1}^{k+1} P_{(j),K_{(j),t}}^{max} > L_{ref,t} \ or \ k = m \tag{19}$$

Then the EVs connected to charging pole $(1)\cdots(k)$ are selected to be charged at the rated charging power.

4.2. Results and analysis

With the parameters listed in Table 2, the algorithm is trained with a maximum of 500,000 steps. The results of DCS when the price

coefficient $\beta = 1$ are shown in Fig. 7, wherein the average DSR when the charging process is completed, reward of EVs, and revenue of the CS up to each stage are plotted. Each stage in the figures consists of 30 days. This shows that the performance of the proposed algorithm can converge with the learning steps. At the beginning of training, the DSR and reward are both low, and the revenue of the CS is high, which indicates that the CS obtains revenue by not fully satisfying the charging demand. As time progresses, the reward and DSR increase, and the revenue decreases. The evolution illustrates that the reward function of agents realizes the trade-off between the revenue from DR and the satisfaction of the demand side.

A series of experiments with different price coefficients are also carried out. The algorithms with $\beta = 1, 2, 3$ are trained for 500,000 steps respectively, and then the algorithms are tested for 10,0000 steps based on the same environment. For the solution time for OCS is too long with long test steps, the OCS is tested for 10,000 steps for each test. The training and testing process are repeated for 10 times. To validate that any proper deep RL algorithm that can deal with continuous states and actions can also be applied, here another two deep RL algorithms, proximal policy optimization (PPO) and twin delayed DDPG (TD3), are tested to learn the charging strategy as well. The details about PPO and TD3 can be found in [39,40]. The test results and performances of the baseline methods and DCS with different deep RL algorithms are shown in Fig. 8, including the average revenue for the CS, average DSR and standard deviation (Std) of DSRs.

The results show that the performances of DCS with different deep RL algorithms are similar in general. Furthermore, a higher price coefficient can generate higher revenue from DR for the CS and lower average DSR for EVs. A higher price coefficient results in a larger weight for revenue, that is, the agents may take more adventurous measures to track the DR signal by curtailing the charging load, which makes EVs more likely to be insufficiently charged. Therefore, the average DSR declines and the Std of the DSR increases with a higher price coefficient. The OCS achieves all the revenue, about 245, by tracking the DR signal perfectly, which is then the upper bound of revenue for the CS. The average DSR and revenue from DR in the OCS are both relatively high because OCS can make optimal decisions with perfect information. The DRM also achieves all the revenue for the CS owing to its absolute satisfaction of the DR signal. However, despite the priority charging of EVs with lower DSR, the average DSR is only about 93.5% and the Std of DSR reaches 0.090, probably damaging the satisfaction of EV owners severely. Compared with the DRM, the proposed DCS can significantly improve the satisfaction of the charging demand. Compared with the OCS, the proposed DCS can provide similar performance when β is large without the information of future arrivals, demands, departures, and DR signals, validating the near-optimal performance of the DCS.

The charging load profiles of the different methods are also investigated by calculating the average load at each time in one day, as shown in Fig. 9 and Fig. 10. The results show that the differences in load profiles occur mainly in the period from 15:00 to 3:00. There is a sharp peak load from 15:00 to 20:00 in the uncontrolled load profile. In addition, the peak load is smoother under the DCS, with the price coefficient being higher. The DRM, OCS, and DCS with $\beta = 3$ maintained the average peak load under the average ReL. The profiles illustrate that the DCS can improve the overload effect caused by a sharp charging load, reducing the risk of collapse of the grid.

The charging load of 10 days is sampled from the test results of different algorithms as shown in Fig. 11. Compared with the uncontrolled load, all the three methods can shave the peak load and reduce the variance of load. The variance of loads in one day of DCS and OCS are larger due to environmental uncertainties, and the load of DRM in one day is smoother for DRM is a rule-based method and relatively more stable. When taking the DRM strategy, the EVs at the valley time can share a large DSR, while the EVs at the peak-load time may share a low DSR, so the DRM strategy may result in unfairness over time. The load of DCS and OCS from about 5:00 to 10:00 is higher than that of uncontrolled load and DRM, which indicates some uncontrolled load during the peak-load time is transferred to be satisfied during the valley time. This may alleviate the unfairness over time although it could sacrifice the fairness for the EVs charging at the same time. As a result, the Std of DSR of DCS and OCS is lower than that of DRM, which can be seen from Fig. 8. It can be found that the charging load of DRM and OCS keeps staying below the ReL while the charging load of DCS cannot always keep staying below the ReL. That is because the DCS needs to trade off the revenue from DR and user satisfaction without future information so DCS may put more priority on DSR during some time intervals.

4.3. Analysis of departure information

In the proposed method, the departure time information is used as a variable in the state space. However, the departure time of EVs cannot be attained or accurately predicted in practical scenarios because the departure time is dependent on the EV owners' behaviors, which are difficult to capture. Therefore, further experiments are conducted to test the performance of the proposed method with inaccurate departure information. Different magnitudes of noise are added to the departure information. Here, the noise is normally distributed with a mean value $\mu = 0$ and variance $\sigma = 1, 3, 5$, respectively.

First, the test results of the algorithms trained and tested with inaccurate departure information are presented in Table 3, where the average DSR and revenue from DR are obviously lower than those without noise in departure information. More specifically, the decrease in average daily revenue is greater, and the Std of DSR is slightly higher with the noise of higher variance. This indicates that the accuracy of the departure information can affect the performance of the charging strategy. Without accurate departure information, the agent cannot properly evaluate the extent of charging urgency of connected EVs, resulting in more unfulfilled demand and less revenue.

The test results of the algorithms trained with accurate departure information and tested with inaccurate departure information are listed

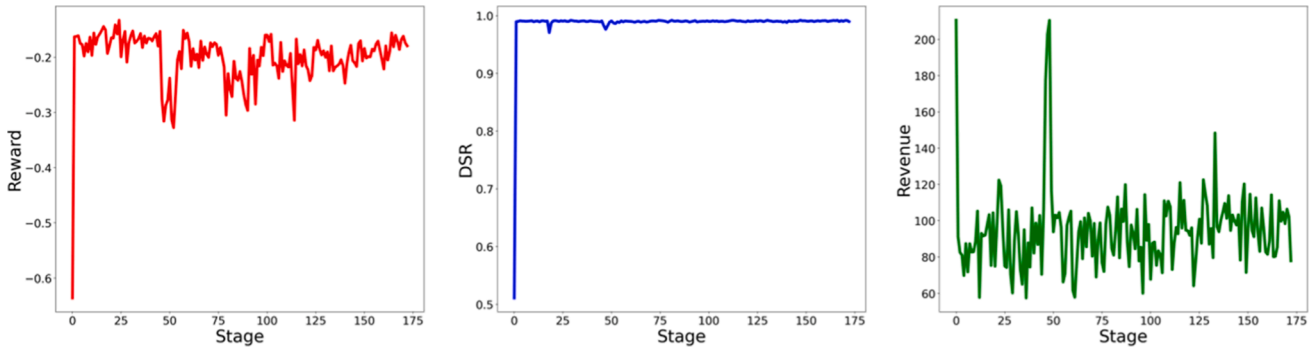


Fig. 7. Reward of electric vehicles (EVs), demand satisfaction rate (DSR), and revenue of the CS up to each stage when DCS with DDPG is applied and.$\beta = 1$

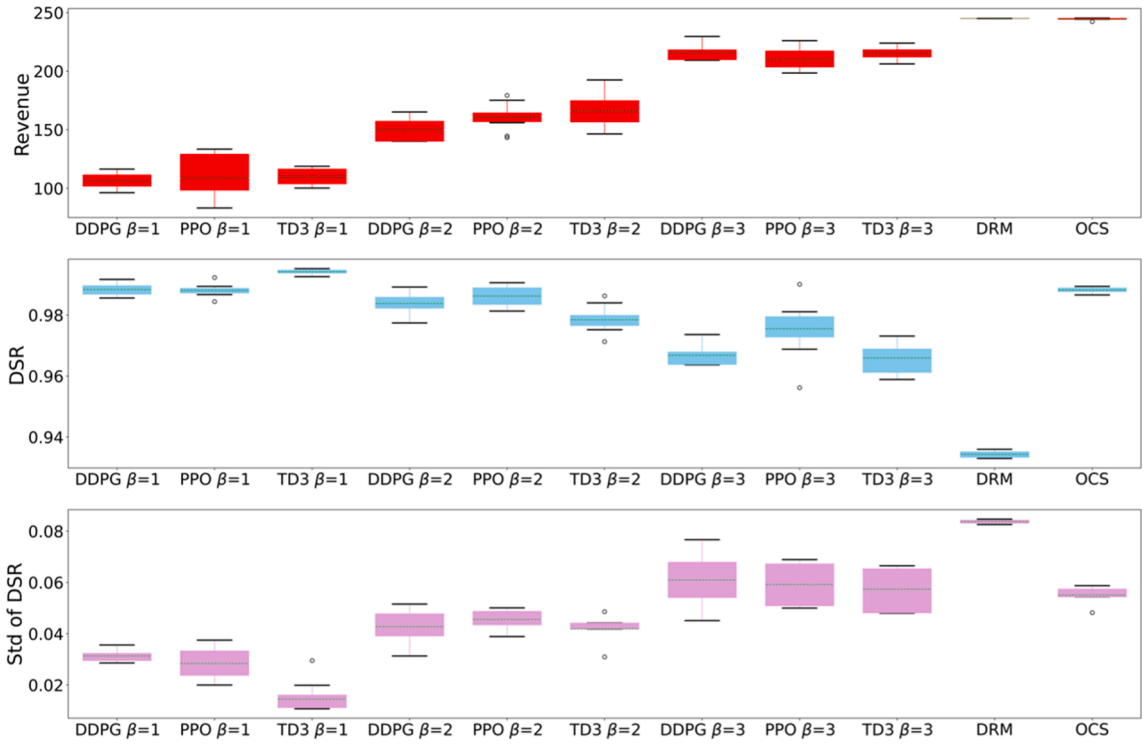


Fig. 8. Revenue from DR, average DSR and Std of DSR of DCS with DDPG, PPO, and TD3 and baseline algorithms.

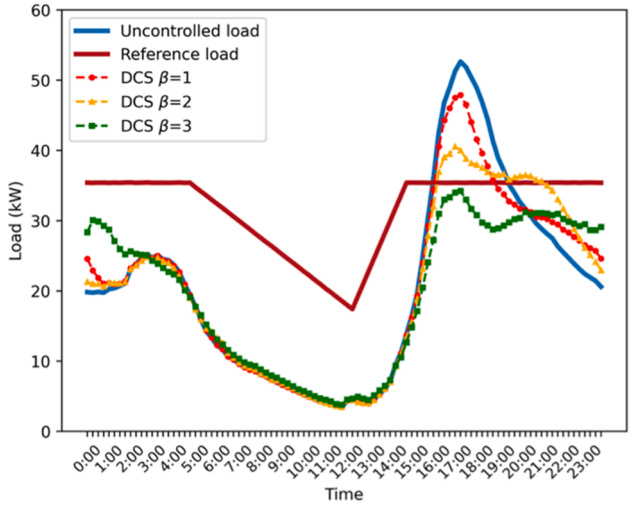


Fig. 9. Average load curve in one day of the DCS with DDPG and different price coefficients.

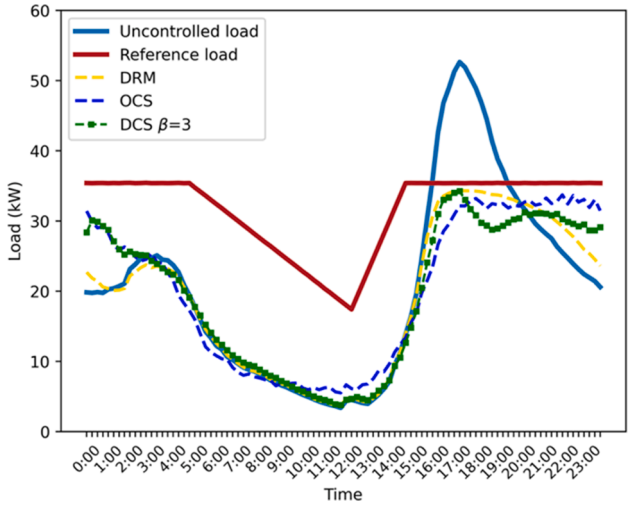


Fig. 10. Average load curve in one day of DCS with DDPG and baseline algorithms.

in Table 4. It is obvious that the effect of inaccurate departure information is slight, although the performance is indeed worse with a higher magnitude of noise, the difference is not very significant. This indicates that training the algorithms with historical accurate information can maintain the performance when interacting with an environment of inaccurate departure information. Thus, the results show that the DCS can still achieve an excellent performance when the departure information cannot be attained accurately, thus verifying its robustness.

5. Conclusion

In this paper, we propose a deep reinforcement learning-based charging strategy for the charging station to coordinate the charging of multiple EVs and participate in DR. The charging process of an EV is modeled as an MDP. In particular, the virtual price is introduced as a significant tool to trade off the revenue from DR and the satisfaction of the demand side. Each agent determines the charging power of the connected EV based on the same single-agent algorithm. All the agents share the same parameters to deal with scalable and heterogeneous EVs, resulting in a linear increase in computation when managing multiple EVs. In the case study, the DDPG algorithm as well as PPO and TD3 is trained and tested based on real charging data in a charging station with 30 charging poles. The results show that any proper reinforcement algorithm that can deal with continuous states and actions can be applied in the proposed strategy. The comparison with the demand-satisfaction-rate-ranking method and optimal charging strategy indicates that the proposed strategy can achieve near-optimality and trade off the revenue and risk well. The proposed strategy can significantly smooth the

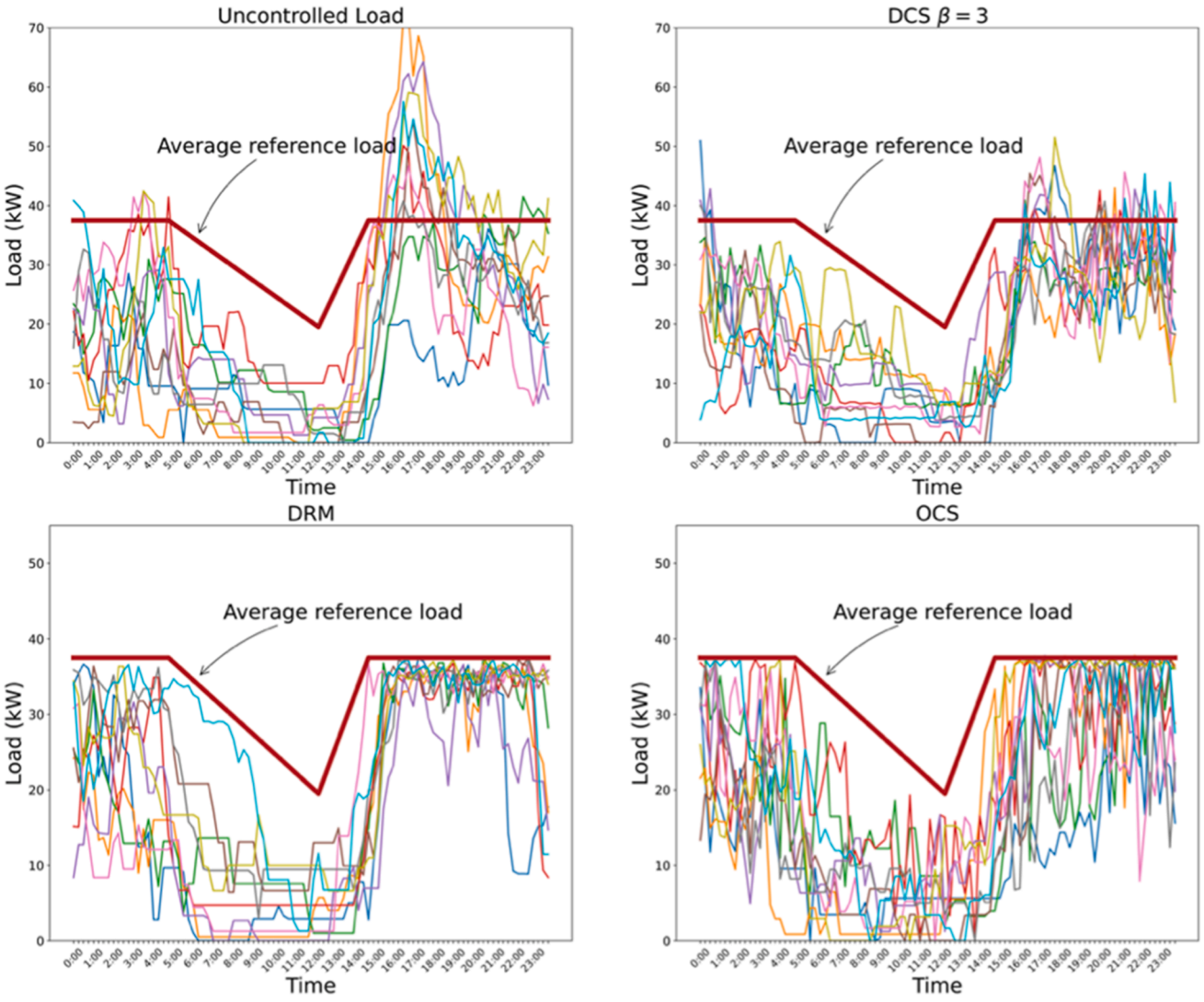


Fig. 11. Samples of daily charging load of DCS with DDPG and baseline algorithms.

Table 3
Test results of the DCS with DDPG when trained with inaccurate departure information.

Method		Average reward	Average DSR	Average revenue	Std of DSR
DCS (σ = 1)	β = 1	−0.167	98.73%	106.50	0.035
DCS (σ = 3)		−0.162	98.56%	103.71	0.036
DCS (σ = 5)		−0.176	98.79%	90.98	0.038
DCS (σ = 1)	β = 2	−0.328	98.47%	154.57	0.041
DCS (σ = 3)		−0.320	98.39%	140.05	0.047
DCS (σ = 5)		−0.349	97.74%	140.13	0.052
DCS (σ = 1)	β = 3	−0.485	96.37%	210.02	0.099
DCS (σ = 3)		−0.459	95.64%	209.71	0.116
DCS (σ = 5)		−0.442	93.85%	214.49	0.130

Table 4
Test results of the DCS with DDPG when trained with historical accurate departure information.

Method		Average reward	Average DSR	Average revenue	Std of DSR
DCS (σ = 1)	β = 1	−0.190	99.75%	119.75	0.0357
DCS (σ = 3)		−0.189	98.73%	119.93	0.0370
DCS (σ = 5)		−0.189	98.63%	120.93	0.0408
DCS (σ = 1)	β = 2	−0.354	98.54%	164.52	0.0360
DCS (σ = 3)		−0.352	98.52%	163.63	0.0377
DCS (σ = 5)		−0.351	98.45%	162.21	0.0409
DCS (σ = 1)	β = 3	−0.481	97.34%	208.87	0.0657
DCS (σ = 3)		−0.480	97.31%	208.25	0.0668
DCS (σ = 5)		−0.479	97.30%	206.97	0.0673

charging load profile of the charging station, which alleviates the pressure on the power grid. Furthermore, the test with inaccurate departure information illustrates that the excellent performance of proposed strategy can be maintained when the algorithm is trained with historical accurate information.

There are still many issues that could be addressed in future work, which are summarized as follows:

1. A better mechanism to deal with unfulfilled charging demand should be explored. In this study, the effect of unfulfilled demand on the operation is not considered, but future work is necessary to take it into account. Compensation or insurance mechanisms may be an effective alternative.

2. The behavior of EV owners should also be considered. Information on EV behaviors is very important for the charging station to make better decisions. However, information on EV arrival, departure, and demand is unknown in practice. In future work, such information may be obtained by prediction based on additional context information of EVs or other data sources.

3. Work related to improvements in the DR mechanism is still in demand. In our work, we decide not to venture into the discussion of this mechanism. However, a DR mechanism that properly considers the interests of different groups is vital to its application.

CRediT authorship contribution statement

Ruiyang Jin: Conceptualization, Methodology, Validation, Visualization, Writing – original draft. **Yuke Zhou:** Methodology, Visualization. **Chao Lu:** Methodology, Investigation, Writing – review & editing, Supervision. **Jie Song:** Investigation, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors are unable or have chosen not to specify which data has been used.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grants 72131001 and U2066601. We also thank the support of High-Performance Computing Platform at Peking University. The data was processed and the algorithms were run in Weiming No.1 supercomputing systems of the platform.

References

[1] Zhang L, Li Y. Optimal management for parking-lot electric vehicle charging by two-stage approximate dynamic programming. IEEE Trans Smart Grid 2015;8(4):1722–30.
[2] Schey S, Scoffield D, Smart J. A first look at the impact of electric vehicle charging on the electric grid in the EV project. World Electric Vehicle Journal 2012;5(3):667–78.
[3] Ban D, Michailidis G, Devetsikiotis M. Demand response control for PHEV charging stations by dynamic price adjustments. 2012 IEEE PES Innovative Smart Grid Technologies (ISGT). IEEE, 2012: 1-8.
[4] Ferro G, Laureri F, Minciardi R, Robba M. An optimization model for electrical vehicles scheduling in a smart grid. Sustainable Energy Grids Networks 2018;14:62–70.
[5] Elma O. A dynamic charging strategy with hybrid fast charging station for electric vehicles. Energy 2020;202:117680.
[6] Lijesen MG. The real-time price elasticity of electricity. Energy Econ 2007;29(2):249–58.
[7] Albadi MH, El-Saadany EF. A summary of demand response in electricity markets. Electr Power Syst Res 2008;78(11):1989–96.
[8] Yao L, Lim WH, Tsai TS. A real-time charging scheme for demand response in electric vehicle parking station. IEEE Trans Smart Grid 2016;8(1):52–62.
[9] Balijepalli VSKM, Pradhan V, Khaparde SA, et al. Review of demand response under smart grid paradigm. In: ISGT2011-India. IEEE; 2011. p. 236–43.
[10] Siano P. Demand response and smart grids—A survey. Renew Sustain Energy Rev 2014;30:461–78.
[11] Sadeghianpourhamami N, Refa N, Strobbe M, Develder C. Quantitive analysis of electric vehicle flexibility: A data-driven approach. Int J Electr Power Energy Syst 2018;95:451–62.
[12] Ding T, Zeng Z, Bai J, Qin B, Yang Y, Shahidehpour M. Optimal Electric Vehicle Charging Strategy with Markov Decision Process and Reinforcement Learning Technique. IEEE Trans Ind Appl 2020;56(5):5811–23.
[13] Tuchnitz F, Ebell N, Schlund J, Pruckner M. Development and Evaluation of a Smart Charging Strategy for an Electric Vehicle Fleet Based on Reinforcement Learning. Appl Energy 2021;285:116382.
[14] Silva FLD, Nishida CEH, Roijers DM, Costa AHR. Coordination of electric vehicle charging through multiagent reinforcement learning. IEEE Trans Smart Grid 2020;11(3):2347–56.
[15] Sadeghianpourhamami N, Deleu J, Develder C. Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning. IEEE Trans Smart Grid 2019;11(1):203–14.
[16] Lahariya M, Sadeghianpourhamami N, Develder C. Reduced state space and cost function in reinforcement learning for demand response control of multiple EV charging stations. Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. 2019: 344-345.
[17] Hou L, Ma S, Yan J, et al. Reinforcement Mechanism Design for Electric Vehicle Demand Response in Microgrid Charging Stations. 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020: 1-8.
[18] Wang S, Bi S, Zhang YA. Reinforcement learning for real-time pricing and scheduling control in EV charging stations. IEEE Trans Ind Inf 2019;17(2):849–59.
[19] Yudovina E, Michailidis G. Socially optimal charging strategies for electric vehicles. IEEE Trans Autom Control 2014;60(3):837–42.
[20] Shin MJ, Choi DH, Kim J. Cooperative management for PV/ESS-enabled electric vehicle charging stations: A multiagent deep reinforcement learning approach. IEEE Trans Ind Inf 2019;16(5):3493–503.
[21] Chiş A, Lundén J, Koivunen V. Reinforcement learning-based plug-in electric vehicle charging with forecasted price. IEEE Trans Veh Technol 2016;66(5):3674–84.
[22] Mhaisen N, Fetais N, Massoud A. Real-time scheduling for electric vehicles Charging/Discharging using reinforcement learning. 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT). IEEE, 2020: 1-6.
[23] Zhang F, Yang Q, An D. CDDPG: A Deep Reinforcement Learning-Based Approach for Electric Vehicle Charging Control. IEEE Internet Things J 2021;8(5):3075–87.
[24] Ortega-Vazquez MA. Optimal scheduling of electric vehicle charging and vehicle-to-grid services at household level including battery degradation and price uncertainty. IET Gener Transm Distrib 2014;8(6):1007–16.
[25] Mukherjee JC, Gupta A. A review of charge scheduling of electric vehicles in smart grid. IEEE Syst J 2014;9(4):1541–53.
[26] Xu Z, Hu Z, Song Y, Zhao W, Zhang Y. Coordination of PEVs charging across multiple aggregators. Appl Energy 2014;136:582–9.
[27] Jian L, Zhu X, Shao Z, Niu S, Chan CC. A scenario of vehicle-to-grid implementation and its double-layer optimal charging strategy for minimizing load variance within regional smart grids. Energy Convers Manage 2014;78:508–17.
[28] Škugor B, Deur J. Dynamic programming-based optimization of charging an electric vehicle fleet system represented by an aggregate battery model. Energy 2015;92:456–65.
[29] Tushar W, Saad W, Poor HV, Smith DB. Economics of electric vehicle charging: A game theoretic approach. IEEE Trans Smart Grid 2012;3(4):1767–78.
[30] Fazelpour F, Vafaeipour M, Rahbari O, Rosen MA. Intelligent optimization to integrate a plug-in hybrid electric vehicle smart parking lot with renewable energy resources and enhance grid characteristics. Energy Convers Manage 2014;77:250–61.
[31] Su W, Chow MY. Performance evaluation of an EDA-based large-scale plug-in hybrid electric vehicle charging algorithm. IEEE Trans Smart Grid 2011;3(1):308–15.
[32] Suganya S, Raja SC, Venkatesh P. Simultaneous coordination of distinct plug-in Hybrid Electric Vehicle charging stations: A modified Particle Swarm Optimization approach. energy 2017;138:92–102.
[33] Sousa T, Morais H, Vale Z, Faria P, Soares J. Intelligent energy resource management considering vehicle-to-grid: A simulated annealing approach. IEEE Trans Smart Grid 2012;3(1):535–42.
[34] Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. Appl Energy 2019;235:1072–89.
[35] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.

[36] Peiran, S, Yanbin L, Changming J, et al. Rule Design and Practice for Third-party Independent Entities Participating in Electric Power Peak Regulation Auxiliary Service Market of North China. Automation of Electric Power Systems, http://doi.org/10.7500/AEPS20200609005.
[37] Powell WB. Approximate Dynamic Programming: Solving the curses of dimensionality. John Wiley & Sons; 2007.
[38] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2015.
[39] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.
[40] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods. International conference on machine learning. PMLR, 2018: 1587-1596.
[41] Song Jie, He Guannan, Wang Jianxiao, Zhang Pingwen. Shaping Future Low-Carbon Energy and Transportation Systems: Digital Technologies and Applications. iEnergy 2022. In press.