Deep Q-Network Based Energy Scheduling in Retail Energy Market

TIANXIN LIN, ZHOU SU, QICHAO XU, RUI XING, AND DONGFENG FANG

Abstract

As a significant component of electric energy trades, retail electric market (REM) can effectively alleviate the pressure of load demand from the power grid. However, the load demand uncertainty of customers becomes a nodus because retail electricity providers (REPs) should predict the load demand when trading with wholesaler electricity provider (WEP) based on the interaction. Therefore, in this paper, we propose an optimal energy scheduling scheme in REM with consideration of the influence of decisions made in pre-purchase stages to situations in real-time stages. Firstly, we present a trading framework to analyze the strategies of REPs in REM, in which REPs conduct both pre-purchase trading with WEP, and real-time trading with customers. Then, to solve the scheduling problem caused by the demand uncertainty of customers, we design a power allocation mechanism based on the charging demand degree, by which REPs can minimize the operating cost while ensuring that each electric vehicle can be charged with sufficient energy. Next, to minimize the cost of REPs in the pre-purchase stage, we adopt deep Q-network (DQN) algorithm to implement the pre-purchase schedule. The charging station adjusts the pre-purchased schedule for each period through Q-learning and utilizes the optimal strategy to design the electricity schedule. Finally, simulation experiments show that the proposal can obtain the optimal strategy to significantly reduce the operating costs of REP.

Index Terms
Retail electricity market (REM), plug-in electric vehicles (PEVs), deep Q-network (DQN).

I. INTRODUCTION
As the global energy is gradually showing a shortage, energy conservation and environmental protection become hot topics that people pay attention to [1]. Emission from fossil energy consumption, which is regarded as the most important cause of climate change and air pollution, is mainly generated by the transportation and electric power produce sectors. In the past, cars were only driven by fuel, which caused severe emissions from fossil energy consumption. Therefore, plug-in electric vehicles (PEVs) are regarded as an efficient solution for alleviating the intensified crisis of the fossil energy consumption. Energy can be saved by using electricity instead of oil, and the environment can be protected by zero emissions of PEVs. Moreover, as a component of distributed energy resource (DER), PEVs are considered as an important solution to manage uncertainties of power generation [2]. With the development of the internet of things (IoT), the information of PEVs can be uploaded to the smart grid, which provides the charging station with plenty of details to analyze [3], [4]. However, PEVs can bring other issues into the power grid. The power grid needs to deal with the deviations of voltages because of PEVs’ mobility and charging uncertainty [5]. If the power grid cannot handle those issues well [6], [7], the PEVs cannot deviate the stress of load demand from users [8], [9], and even damage the stability of the power grid.

The renewable energy source (RES) is considered as another effective option to mitigate energy shortage and environmental pollution [10]. Developing wind power plants and hydropower plants instead of thermal power plant can reduce the pressure of fossil energy mining [11]. However, since the energy generation of RES is fluctuant, it is arduous to predict the precise information of power generation. Moreover, how to integrate RES into the grid becomes an non-trivial issue [12]. Thus, how to balance the power generation...
and load demand with the optimal strategies [13], [14] still remains a topic worthy of research.

The development of the electricity retail market forms a new structure between the wholesale electricity providers (WEPs) and the customers, which becomes an effective distributed method to solve the uncertainty of RES and the impact of PEVs. WEPs control a number of RES plants and set a time-of-use (TOU) power price by analyzing the uncertainty of RES. The retail electricity providers (REPs) purchase energy from WEPs at a pre-scheduling stage while selling energy to PEVs at real-time stage. By predicting the trading amounts at real-time stages, REPs purchase corresponding energy from the WEPs. In real-time operation, REPs set price for PEVs, and endeavour to reduce the deviation between pre-scheduling and real-time trading by taking several measures.

Current works [15], [16] to explore the architecture of REM mainly focus on the integration of DER and the demand responses of customers. However, current methods cannot provide an efficient solution when different types of trading are taken into account at the same time. On one hand, most of them just analyze the interactions either between REPs and WEPs or between REPs and PEVs, where the influence between those three parties has not been concerned. On the other hand, even some works have studied the pre-purchase stage, only single type of trading is considered, such as day-ahead trading. The different types of pre-purchase trading interact and impact the real-time market have not been mentioned. Due to the fluctuation of WEP’s electricity price and the uncertainty of load demand of PEVs, how to formulate a charging schedule for REPs is still a crucial issue with consideration of different trading types.

In this paper, a novel trading mechanism is proposed, in which multiple trading types are considered. Specifically, when REPs conduct a real-time trading with PEVs, REPs can obtain PEVs’ requests and draw up an optimal schedule for allocating power to the parking PEVs based on the charging urgency of each vehicle and the total power to be allocated. When trading with WEP, REPs can determine the optimal power schedule by both future trading and day-ahead trading. Future trading is to make a contract which ensures the trading amount in certain days, and day-ahead trading is to purchase extra amount on the day before real-time trading. In order to achieve a satisfied schedule, the operation cost of REP should be minimized. As such, we further model interactions among different trading types as a Markov decision process (MDP). In addition, a deep Q-network (DQN) based algorithm is presented to obtain the optimal strategy corresponding to different prices.

In the nutshell, the main contributions of this paper are threelfold.

1) Considering the interaction between pre-purchase trading and real-time trading, we present an energy scheduling model to show the trading modes among WEP, REP, and PEVs. Two trading modes are provided for REPs to pre-purchase power from WEP as future trading and day-ahead trading, respectively. Moreover, real-time trading with WEP is available for REPs in case that the pre-purchase schedule cannot match the selling amount to PEVs.

2) A charging schedule model is designed for each PEV in the real-time trading. All PEVs are required to submit their charging messages to REPs. REPs can delay the charging schedules for several PEVs under the condition that the charging demand of all PEVs could be satisfied, by which the operation cost can be significantly reduced. Furthermore, a mechanism to decide the charging amount for each PEV is presented to obtain its maximum utility.

3) A DQN based algorithm is proposed to schedule the future trading and day-ahead trading for REPs, respectively. Simulation results demonstrate that the proposed algorithm for a pre-purchased schedule can obtain the optimal strategies for both future trading and day-ahead trading.

The rest of this paper is organized as follows. The related work is reviewed in Section II. The system model is introduced in Section III. A power scheduling strategy for REPs and a strategy for PEVs to dynamically determine the charging amount are introduced in Section IV. In Section V, DQN is utilized to decide on the electricity pre-purchase schedule for REPs. The performance evaluations are shown in Section VI. The conclusion is given in Section VII.

II. RELATED WORK

A. INTEGRATION OF REM

An overview is given to describe the development of algorithms used in REM. REM is composed of three parts, which are WEP, REP, and PEVs, respectively. Existing works concerning about REM almost ignore the demand uncertainty. Sortomme and El-Sharkawi [17] utilized an aggregator to develop an algorithm for unidirectional regulation. Han and Sezaki [18] established an aggregator that efficiently utilizes the distributed energy source for PEV to produce the required electricity power. Ghazvini et al. [19] accessed the electricity market in Portugal and concerned about how retail rates affected the market.

Vandael et al. [20] presented an allocation schedule in real-time trading and made a heuristic EV fleet by day-ahead trading, but the interaction between future trading and day-ahead trading is not considered. Samadi et al. [21] introduced a distributed methodology that effectively controls the interactions among the energy consumption controller units at automatical measurer and energy producer. Qian et al. [22] established a real-time pricing system which utilized demand response to mitigate the peak load demand in the smart grid.

The optimal charging control for PEVs has been investigated in various works. Guo et al. [23] presented a hybrid control scheme to achieve a high efficiency and designed...
a guide of the compensator to ensure a stable performance. Bo et al. [24] researched on a charging control strategy in the residential area, which considers driving characteristics and peak load constant. Shuanglong et al. [25] studied on a charging controller cluster system and utilized the cluster control technology for connecting charging facilities. Khazraei et al. [26] established an average switch model to achieve the high utility of large multi-module converters. Kim et al. [27] proposed a pre-training framework for autonomous vehicles concerning the policy gradient of agents.

Although there are a plenty of works on REM, these works have not studied the design of multiple trading systems. Distinguished from existing works, in our work, REPs should determine corresponding strategies that could influence strategies in some future trading stages.

B. APPLICATION OF DQN

DQN is a promising technology that combines Q-learning and convolutional neural networks (CNN). Q-learning, as a component of reinforcement learning, is used to calculate a Q-matrix, in which we can select the most rewarding action in different states. However, if the division of states is too large to store in the Q-matrix, it will be a dimensional disaster [28], [29]. CNN is a proper method to solve this problem. In DQN, CNN is utilized to get a low-dimensional action output with a high-dimensional state input. In this way, the dimension of state into reinforcement learning can be expanded in a large dimension, and the corresponding output can be calculated by CNN. It will be more convenient to select actions in Q-learning and update the Q-matrix.

Reinforcement learning has been applied on the smart grid. Dimitrov and Lguensat [30] presented an online reinforcement learning methodology which optimized the utility of each PEV charging station which controls several renewable energy sources. Chis et al. [31] proposed a batch reinforcement learning algorithm in which an optimal charging policy is learned from a batch processing of training samples and the optimal charging decisions is designed when dealing with new situations. Radhakrishnan et al. [32] utilized techniques of agent-based computational to overcome the unpredictability in electricity prices. Zhou et al. [33] established a piece-wise regulation in order to simulate continuous and discrete regulation demand loads. In contrast to the above works, we introduce reinforcement learning to predict the charging behavior of electric vehicles and combine deep learning to solve the dimensional disaster problem of reinforcement learning.

III. SYSTEM MODEL

A system model of retail electricity market is shown in Fig.1. The component of system model includes a WEP, a REP and PEVs. The microgrid operator (MGO) provides WEP with distributed energy resource. REP purchases power from WEP by two options, which are pre-purchase trade and real-time trade, respectively. REP controls several charging station and sells power to PEVs. The pre-purchasing model and the charging model are going to be introduced as follows. The notations are summarized in TABLE 1.

A. PRE-PURCHASING MODEL

The WEP provides REP with two modes to pre-purchase power schedule, which are future trading and day-ahead trading, respectively. A trading day is divided into several time slots, and the electricity price is set as TOU price, which means that the price in different time slots will change as the supply changes. Two kinds of prices are set in considering the fluctuation of renewable resource. Future trading and day-ahead trading can be chosen by REPs. For future trading, REPs can make a contract with a WEP, which stipulates that in the future days, REPs will purchase a certain quantity of electricity at a certain price. \( P_{fu} = \{P_{fu}(1), \ldots, P_{fu}(t), \ldots, P_{fu}(T)\} \) is denoted as the power schedule REPs purchase from WEP by future trading and denote \( \beta_{fu} = \{\beta_{fu}(1), \ldots, \beta_{fu}(t), \ldots, \beta_{fu}(T)\} \) as the price of future trading in a period.

As for the day-ahead trading, WEP announces the price of day-ahead trading in the day before real-time trading day, which is denoted as \( P_{da} = \{P_{da}(1), \ldots, P_{da}(t), \ldots, P_{da}(T)\} \). Depending on \( \beta_{da} \), REPs decide the power that they will purchase in each time slot, which is denoted as \( P_{sch} = \{P_{sch}(1), \ldots, P_{sch}(t), \ldots, P_{sch}(T)\} \). Considering these two modes, the amount of pre-purchased power \( P_{sch}(t) \) in each time slot can be considered as

\[
P_{sch}(t) = P_{fu}(t) + P_{da}(t), \quad t \in \{1, \ldots, T\}.
\]

Furthermore, the power purchased from WEP in each time slot needs to comply with

\[
P_{sch} < P_{sch}^{max},
\]

where \( P_{sch}^{max} \) is the limit of maximum charging power. REPs can firstly conduct future trading by a certain contract, and
TABLE 1. Summary of notations.

| Notations | Description |
|-----------|-------------|
| \( P_{fu} \) | The power schedule REPs pre-purchase in the wholesale market by future trading. |
| \( P_{da} \) | The power schedule REPs that pre-purchase in the wholesale market by day-ahead trading. |
| \( P_{ch} \) | The total power REPs that pre-purchase in the wholesale market. |
| \( P_{\text{max}} \) | The maximum power REP can pre-purchase from WEP. |
| \( \beta_{fu} \) | The price of future trading. |
| \( \beta_{da} \) | The price of day-ahead trading. |
| \( \beta_{retail} \) | The price of real-time trading for which REPs sell electricity to WEP. |
| \( \beta_{self} \) | The price of real-time trading for which REPs sell electricity to PEVs. |
| \( t \) | The index of slot. |
| \( \Delta t \) | The length of time slot. |
| \( T \) | The total time slots in a day. |
| \( i \) | The index of PEV. |
| \( l_i \) | The PEVs parking in the charging station in the time slot t. |
| \( M^i \) | The charging message of PEV i. |
| \( t_{\text{arr}}^i \) | The arriving time of PEV i. |
| \( t_{\text{dep}}^i \) | The departure time of PEV i. |
| \( E_{\text{req}}^i \) | The total energy requirement of PEV i. |
| \( E_{\text{req}}^i \) | The total amount of electric that each PEV can get after allocation. |
| \( P_{\text{lim}}^i \) | The charging power limit of PEV i. |
| \( C_{da,\text{buy}}(t) \) | The cost caused by purchasing extra power in day-ahead trading in the time slot t. |
| \( C_{da,\text{sell}}(t) \) | The cost caused by selling extra power in day-ahead trading in the time slot t. |
| \( C_{da} \) | The extra cost of day-ahead trading. |
| \( C_{fu,\text{buy}}(t) \) | The cost caused by purchasing extra power in future trading in the time slot t. |
| \( C_{fu,\text{sell}}(t) \) | The cost caused by selling extra power in future trading in the time slot t. |
| \( C_{fu} \) | The extra future cost of future trading. |
| \( P_{\text{fl}}(t) \) | The power PEV i can get in the time slot t. |
| \( l_i^f \) | The PEVs parking in the charging station in the time slot t. |
| \( g_i^f \) | The urgency degree of PEV i in the time slot t. |
| \( P_{\text{new}}(t) \) | The new schedule after REPs do real-time trading with WEPs. |
| \( m_i \) | The marginal charging benefit of PEV i. |
| \( \omega \) | The maximum marginal charging benefit PEV i. |
| \( \xi \) | The charging efficiency reduction rate of PEV i. |
| \( \text{gain}^f \) | The total charging benefit of PEV i. |
| \( U^i \) | The total charging utility of PEV i. |
| \( e^i \) | The driving consumption for per unit distance of PEV i. |
| \( d^i \) | The distance that PEV i needs to drive to the charging station. |

conducted day-ahead trading in the day before real-time trading day as a supplement.

B. CHARGING MODEL

Each PEV charging in the station is required to submit its charging message \( M^i \left( t_{\text{arr}}^i, t_{\text{dep}}^i, E_{\text{req}}^i, P_{\text{lim}}^i \right) \), where \( i \) is the index of PEVs, \( t_{\text{arr}}^i \) and \( t_{\text{dep}}^i \) are the arriving time and the departure time of the PEV \( i \), respectively. \( E_{\text{req}}^i \) is the amount of energy requirement, and \( P_{\text{lim}}^i \) is the maximum limit charging power of PEVs. The submitted energy requirement should satisfy

\[
E_{\text{req}}^i \leq P_{\text{lim}}^i \left( t_{\text{dep}}^i - t_{\text{arr}}^i \right). \tag{3}
\]

PEVs can charge enough energy when it leaves the charging station. However, REPs can defer the charging schedule of those PEVs parked in the charging station for a long time in order to ensure each PEV can gain enough energy with the lowest operation cost.

After accepting charging message from PEV \( i \), REPs create a charging schedule for PEV \( i \) as \( P_{\text{al}}^i(t) \). In the charging schedule, the charging power in each time slot can be different, but the total charging amount \( E_{\text{req}}^i \) should satisfy

\[
E_{\text{req}}^i = \sum_{t \in T} P_{\text{al}}^i(t) \Delta t, \tag{4}
\]

where \( P_{\text{al}}^i(t) = 0 \) if PEV is not parking at the charging station in the time slot \( t \).

IV. PROBLEM FORMULATION

In this section, we present a scheduling and pricing mechanism to allocate charging power to each PEV, aiming to minimize the total operation cost in a period. Moreover, we design a charging mechanism for PEVs to choose whether to charge or not and decide the amount to charge to maximize the largest utility. The following sections include REPs’ cost, power allocation and charging polices.
A. REPs COST

In the allocation stage, REPs have already pre-purchased power by pre-purchase trading. In this stage, REPs can improve revenue by reducing the cost of real-time trading to ensure that PEVs’ demand can be satisfied. Therefore, REPs should develop a dispatch plan about allocating the power that they purchase from a WEP to each PEV. The allocation mechanism is considered in conjunction with the amount of real-time trading and the urgency degree of PEVs.

On one hand, if a PEV cannot gain enough electricity when it leaves the station, REPs should purchase extra power by pre-purchase trading. In this stage, REPs can buy power at a lower cost, which will be the pre-purchased cost. If the pre-purchased power is larger than the total amount of real-time trading, REPs can purchase all the power by pre-purchase trading. If the pre-purchased power is less than the total amount, they need to purchase the real-time power. If the total pre-purchased power is larger than the total amount of real-time trading, REPs will not purchase any power by real-time trading.

REPs should develop a dispatch plan about allocating the power that they purchase from a WEP to each PEV. The allocation mechanism is considered in conjunction with the amount of real-time trading and the urgency degree of PEVs. On one hand, if a PEV cannot gain enough electricity when it leaves the station, REPs should purchase extra power by pre-purchase trading. In this stage, REPs can buy power at a lower cost, which will be the pre-purchased cost. If the pre-purchased power is larger than the total amount of real-time trading, REPs can purchase all the power by pre-purchase trading. If the pre-purchased power is less than the total amount, they need to purchase the real-time power. If the total pre-purchased power is larger than the total amount of real-time trading, REPs will not purchase any power by real-time trading.

Suppose that the power REPs pre-purchasing from a WEP is equal to the power that REPs sell to PEVs in each time slot. In this case, REPs will not purchase any power by real-time trading, so the total cost can be reduced. However, due to the charging uncertainty of each PEV, the power sold to PEVs cannot be predicted exactly, which results in the extra cost. We define the cost as the loss of deviation of prediction. When conducting day-ahead trading, the extra cost that REPs pay for real-time purchasing is

\[ C_{da, buy}(t) = P_{fu}(t)\beta_{fu}(t) + P_{da}(t)\beta_{da}(t) + P_{buy}(t)\beta_{buy}(t) - P_{fu}(t)(\beta_{fu}(t) + P_{buy}(t) + P_{da}(t))\beta_{da}(t) \]

Similarly, the extra cost that REP pays for real-time selling is

\[ C_{da, sell}(t) = P_{fu}(t)\beta_{fu}(t) + P_{da}(t)\beta_{da}(t) - P_{sell}(t)\beta_{sell}(t) - P_{fu}(t)(\beta_{fu}(t) + P_{da}(t) - P_{sell}(t))\beta_{da}(t) \]

C. The urgency degree is related to the remaining requirement of energy in time slot \( t \), the remaining time PEV parking in the charging station, and the charging limit of PEV. We define the urgency degree as follows

\[ \tau^i(t) = \frac{E_{req}^i - E^i(t)}{(t_{dep}^i - t) \cdot \left( P_{ach}^i \right)}, \quad i \in I, \]

where \( I \) denotes those PEVs parked in the charging station in the time slot \( t \), and \( E^i(t) \) indicates the total energy which has been charged to PEV \( i \). Then, depending on the urgency degree, the power allocated to each PEV in each time slot is designed by

\[ P_{ach}^i(t) = \min \left( \frac{\tau^i(t) \cdot P_{ach}^i(t)}{\sum_{i \in I} \tau^i(t) \cdot P_{ach}^i(t), P_{ach}^i(t)}, \quad t \in T. \]

After allocation, we can directly calculate the power which should be resold to WEP in each time slot. The power resold
to WEP in the time slot \( t \) is denoted as

\[
P_{\text{sell}}(t) = \max \left( P_{\text{sch}}(t) - \sum_{i \in I} P_{\text{al}}^i(t), 0 \right), \quad t \in T. \tag{13}
\]

REPs should purchase extra power from a WEP by real-time trading to ensure that each PEV can get as much energy as it requires before leaving the charging station. The total amount of energy that each PEV can get after allocation can be calculated by

\[
E_{\text{al}}^i = \sum_{t=t_{\text{dep}}^i}^{t=t_{\text{arr}}^i} P_{\text{al}}^i(t) \cdot \Delta t. \tag{14}
\]

If PEVs cannot get enough energy charging after following the algorithm, REPs should purchase extra power at real-time price. The total amount of extra power for each PEV can be calculated by

\[
E_{\text{buy}}^i = E_{\text{req}}^i - \sum_{t \in T} P_{\text{al}}^i(t) \cdot \Delta t. \tag{15}
\]

The total energy that REPs should purchase from a WEP in real-time trading can be calculated by

\[
E_{\text{buy}} = \sum_{i \in I} E_{\text{buy}}^i = \sum_{i \in I} \left( E_{\text{req}}^i - \sum_{t \in T} P_{\text{al}}^i(t) \cdot \Delta t \right).
\]

To minimize the total cost of REPs, REPs would like to purchase power at time slot \( t \) when \( \beta_{\text{buy}}(t) \) is the lowest price among \( \beta_{\text{buy}} \). The purchasing power in real-time trading is then denoted by

\[
P_{\text{buy}}(t) = \min \left( \sum_{i \in I} \left( P_{\lim}^i - P_{\text{al}}^i(t) \right), \sum_{t \in T} E_{\text{buy}}(t)/\Delta t, P_{\text{max}} - P_{\text{sch}}(t) \right), \tag{17}
\]

where \( \sum_{i \in I} \left( P_{\lim}^i - P_{\text{al}}^i(t) \right) \) is the maximum extra power that PEV \( i \) can charge in the time slot \( t \), which ensures that each PEV can charge in their limited power. \( \sum_{t \in T} E_{\text{buy}}(t)/\Delta t \) is to ensure that the purchasing amount is equal to the amount of requirement and \( P_{\text{max}} - P_{\text{sch}}(t) \) is to ensure that the total charging amount is no more than the limitation of the power grid, respectively. After purchasing by real-time trading, the new schedule that station can allocate is denoted as

\[
P_{\text{sch}}^t = P_{\text{sch}}(t) + P_{\text{buy}}(t) - P_{\text{sell}}(t). \tag{18}
\]

Replacing \( P_{\text{sch}}(t) \) with \( P_{\text{sch}}^t \), we reallocate the power according to (10) and (11). Once a PEV arrives at the charging station and submits its message, the urgency degree for each PEV is changed, so formula (10)-(17) will be recalculated. After one day of trading, the total cost in the period can be calculated by (6) and (9).

### C. CHARGING POLICY

After DQN-training, the pre-purchasing price is announced to REPs, including \( \beta_{\text{fu}}, \beta_{\text{da}}, \beta_{\text{sell}} \) and \( \beta_{\text{buy}} \), and the amount of pre-purchasing is submitted to a WEP. The REPs should set a price mechanism for selling power to PEVs. We set that the retail price of electricity sold to electric vehicles is described as

\[
\beta_{\text{retail}} = \varrho \cdot \max \left( \beta_{\text{fu}}(t) \right), \quad t \in T, \tag{19}
\]

where \( \varrho \) is a parameter determined by the maximum value of \( \beta_{\text{buy}} \), i.e.,

\[
\varrho = \frac{\sum_{i \in I} \max \left( \beta_{\text{buy}}(t) \right) - \beta_{\text{buy}}(t)}{T \min \left( \beta_{\text{buy}}(t) \right)} + 1. \tag{20}
\]

The marginal benefit of PEVs charging is set as

\[
m^i = \omega^i - \xi^i E_{\text{req}}^i, \tag{21}
\]

where \( \omega^i \) is a parameter which refers to the maximum charging marginal benefit of PEV \( i \). \( \xi^i \) is the parameter which refers to the charging efficiency reduction rate of PEV \( i \). Then, the benefit of PEV \( i \) can be calculated as

\[
\text{gain}^i = \begin{cases} 
\omega^i E_{\text{req}}^i - \frac{\xi^i}{2} (E_{\text{req}}^i)^2, & 0 \leq E_{\text{req}}^i \leq \frac{\omega^i}{\xi^i} \\
\frac{(\omega^i)^2}{2 \xi^i} & E_{\text{req}}^i > \frac{\omega^i}{\xi^i}.
\end{cases} \tag{22}
\]

The utility of PEV \( i \) can be calculated as

\[
U^i = \begin{cases} 
\frac{(\omega^i)^2}{2 \xi^i} - \beta_{\text{retail}} E_{\text{req}}^i - \epsilon^i d^i, & E_{\text{req}}^i \leq \frac{\omega^i}{\xi^i} \\
\omega^i E_{\text{req}}^i - \frac{\xi^i}{2} (E_{\text{req}}^i)^2 - \beta_{\text{retail}} E_{\text{req}}^i - \epsilon^i d^i, & \text{else}.
\end{cases} \tag{23}
\]

where \( d^i \) is the distance that PEV \( i \) needs to drive to the charging station, \( \epsilon^i \) is the driving consumption per unit distance.

In order to maximize the utility of PEV \( i \), the optimal charging amount of PEVs can be calculated as

\[
E_{\text{req}}^i = \frac{\omega^i}{\xi^i} - \beta_{\text{retail}}. \tag{24}
\]

The maximum of utility that PEV \( i \) can get can be obtained by

\[
U^i_{\text{max}} = \frac{(\omega^i)^2}{2 \xi^i} - \epsilon^i d^i. \tag{25}
\]

Only when the utility is positive can the PEVs decide to go to the charging station, which means that the relationship between \( \omega^i \) and \( \xi^i \) should satisfy \( \omega^i \geq \beta_{\text{retail}} + \sqrt{2 \xi^i \epsilon^i d^i} \). This mechanism is designed to update the training set after each period of DQN, which will be introduced in the next section.
V. DQN-BASED PREDICTIONS OF PRE-PURCHASE

To minimize the operation cost, we need to predict the pre-purchase trading as precisely as possible. In this section, a DQN-based algorithm is developed to calculate the amount of pre-purchased trading. When doing day-ahead trading, REPs get the message from a WEP about the price of day-ahead trading and real-time trading. However, due to the fluctuation and randomness of PEVs, it is inevitable to analyze how to make a precise schedule that matches the PEVs’ arriving information in the future.

A. DAY-AHEAD TRADING

The goal of the REPs strategy is to choose the daily amount of day-ahead trading to minimize the cost of deviation of $C_{da}$. We set that future trading has been done, which means that $P_{fu}$ is a certain schedule. The process of deciding the day-ahead schedule is defined as an MDP process where the state space is a certain schedule. The process of deciding the day-ahead trading to minimize the cost of deviation of $C_{da}$.

First of all, we initialize $P_{da}$ by a series of random numbers and initialize Q value by a zero matrix. In each episode $t$, REPs choose the action $P_{da}(t)$ for each time slot $t$ to update $P_{da}$. Then, the reward can be calculated by the allocation algorithm. The $\epsilon$ greedy method is utilized for REPs to choose actions, which means the probability of choosing the action which maps the largest Q value for REPs is $\epsilon$. On the contrary, the probability of choosing a random action to explore the new situation is $1-\epsilon$. According to [34], after choosing a certain policy, the reward is adopted to update the Q matrix by Bellman equation as

$$Q(s,a) = \alpha (r + \gamma (\max(Q(s',a')))),$$  (26)

where $r$ is the reward when choosing the policy $a$ in state, which is designed as the negative of $C_{da}$. $Q(s',a')$ is the Q reward that can be obtained when choosing next policy $a'$ in next state $s'$, and $\gamma$ is the reward diminishing parameter which shows the degree of emphasis on the reward in the future. The larger $\gamma$ is, the more attention pays to the future reward. Otherwise, the less $\gamma$ means REP cares more about the reward that can be obtained from current action.

However, the relationship between states and rewards cannot be correctly represented in the Q matrix because of the high division of states. A CNN is used to restore the relationship of states, actions, and rewards, in which we define the Q value as the actual output of the neural network. The network owns multiple input and a single output, which can solve the problem of dimension disaster. The states and the actions are the input of the neural network while the Q value is the output. The goal of the neural network is to minimize the loss of actual output and estimate output. The loss function applied is Mean Square Loss. We utilize mean square error to define the loss in the neural network as

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Q^i - \hat{Q})^2,$$  (27)

where $n$ is the total times of training, $Q^i$, which presents the actual Q-value, is equal to the result we calculate by the allocation algorithm. $\hat{Q}$ is the output that we get from the neural network, which means the target that we want to get from the neural network. According to the neural network, we can input a certain pre-purchasing schedule and a certain action. Then, the output is obtained which is regarded as the reward of Q-learning. Using $\hat{Q}$ to update the value in (22), we can choose the best input as a day-ahead trading schedule which matches the largest Q-value.

B. FUTURE TRADING

When conducting the future trading, REP can only get the price of future trading $\beta_{fu}(t)$. Similar to day-ahead trading, the process of future trading is also indicated as an MDP process. The state is defined as $S = \beta_{fu} \times T$. REPs need to use the data of past transactions for training. Each episode includes a certain number of days. In this paper, we set each period of future trading including 30 days. The day-ahead price and the real-time price are different every day, which leads to the reward of future trading is the average of the reward in 30 days. The action space is to choose the energy demand between $[0, P_{da}^{max}]$ in each time slot.

We also establish a neural network to store the Q values. Different from day-ahead trading, each episode includes 30 days, so the amount of pre-purchase electricity is composed of two parts. One is $P_{fu}$, which is decided by actions, and the other one is $P_{da}$, which is decided by day-ahead trading method introduced in the last section. After an episode, the output of the neural network is the average Q value in 30 days. After full training, the best strategy of future trading can be decided according to different future trading prices.

With the two neural networks used for future trading and day-ahead trading respectively, the best pre-purchase schedule can be determined. Firstly, REPs design $P_{fu}$ according to $\beta_{fu}$. Then, $P_{da}$ can be decided according to the purchased $P_{fu}$ and $\beta_{da}$, which is announced by a WEP every day. Finally, in the real-time trading, the allocation algorithm can make difference to save operation cost.

VI. PERFORMANCE EVALUATION

In this section, we conduct a series of simulations to evaluate the effectiveness of the proposed mechanism. The simulation setup is first introduced, followed by numerical results.

A. SIMULATION SETUP

In the simulation, twenty time slots are available for power trading in a day and each time slot lasts for an hour. Sixteen levels of electric amounts are provided for REPs to choose to purchase from WEP in each time slot. As for the data
Algorithm 1 DQN Algorithm

1: Input: Episode $M$, reward function $R$, greedy degree $\varepsilon$, Study rate $\alpha$, reward diminishing parameter $\gamma$;
2: Initialize repetition memory pattern $D$ to store $N$;
3: Initialize the origin action-value function $Q(s,a)$ with weights $\theta$ by a random value;
4: Utilize weights $\theta = \theta - \theta$ to initialize objective action-value function $\hat{Q}$;
5: for episode = 1, ..., $M$ do
6: Initialize state sequence with $s_1 = x_1$ and $\phi_1 = \phi(s_1)$;
7: for $t = 1, ..., T$ do
8: Utilize probability $\varepsilon$ to choose action $a_t$ with a random value;
9: Otherwise set $a_t = \arg\max_a Q(\phi(s_t),a;\theta)$;
10: Choose the action $a_t$ and gain the reward $r_t$ by reward function $r_t$ and image $x_{t+1}$;
11: Set $s_{t+1} = s_t, a_t, x_{t+1}$ and $\phi_{t+1} = \phi(s_{t+1})$;
12: Store interim $(\phi_t, a_t, r_t, \phi_{t+1})$ in $D$;
13: Utilize a minimized random batch sample for interims $(\phi_t, a_t, r_t, \phi_{t+1})$ from $D$;
14: if episode stops at step $j + 1$ then
15: Set $y_j = r_j$;
16: else
17: Set $y_j = \alpha(r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-))$;
18: end if
19: Select a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ pertaining to the related network parameters $\theta$;
20: Reset $\hat{Q} = Q$;
21: end for
22: end for

of PEVs, we set that the maximum limit of PEVs’ charging power is 4 kWh. The time distribution when the PEVs arrive at the charging station follows a normal distribution with a variance of 5, and the time distribution when the PEVs leave the charging station satisfies a normal distribution with a variance of 15. As for the distance distribution between PEVs and charging station, it follows the even distribution in [0, 10] km. For all PEVs, the maximum marginal benefit is set to be $\omega = 60$, and the efficiency reduction rate of charging is set to be $\xi = 1$. Other related parameters in the simulation are summarized in TABLE 2.

B. CASE STUDY RESULTS

After the pre-purchase operation, the REP can obtain a certain power from a WEP. The total schedule REP can dominate is shown in Fig. 2. Then, the REP allocates the power schedule to PEVs according to the urgency degree allocation algorithm. The price of the pre-purchase trading and the real-time trading is shown in Fig. 3. Generally speaking, the price of the day-ahead trading is larger than that of future trading, and the price of pre-purchasing is larger than that of real-time selling price but smaller than that of the real-time purchasing price. Furthermore, because the price is set by TOU power price to adapt to the fluctuation of DER, the prices in different time slots have different values.

During allocation, REPs should purchase power from a WEP by real-time trading to ensure that each PEV can get the amount of electricity they required, and selling power to WEP if the power cannot be sold to PEVs in several time slots. The schedule of allocating power is shown in Fig. 4. Compared with Fig. 2, during the time slots from 0 to 5,

| Parameter | Value |
|-----------|-------|
| $T$       | 20    |
| $\xi$     | 1     |
| $\varepsilon$ | 10   |
| $P_{lim}$ | 4     |
| $\alpha$  | (0.01, 0.1) |
| $\gamma$  | (0.8, 0.9) |
| $\omega$  | Rand in [50, 100] |
| $d_i$     | Rand in [0, 100] |
| $E_{reg}$ | Rand in [20, 80] |

FIGURE 2. Pre-scheduling power purchasing from WEP in case study.

FIGURE 3. Price of pre-purchasing and real-time trading in case study.
the pre-purchased power still remains after REPs selling power to each PEV, for which the error of over-purchasing occurs. On the contrary, during the 13-17th time slots, REPs should purchase power in real-time trading, for which the error of under-purchasing occurs.

For comparison purposes, we set another simulation in which REPs do not adopt our allocation algorithm. Instead, PEVs can charge at their maximum power as soon as possible. The comparison of these two methods is shown in TABLE 3. As we can see, with our allocation algorithm, both the day-ahead trading cost and future trading cost can be reduced significantly.

In the pre-purchasing stage, DQN is utilized to provide the best strategy for REP to make a pre-purchasing schedule plan. In reinforcement learning, the pre-purchasing price and real-time trading price are regarded as the state, which is stored on CNN. The results we simulate in the day-ahead trading and future trading are shown in Fig. 5 and Fig. 6, respectively. The simulation is unstable at the beginning and tends to be stable after 5000 episodes. Sometimes DQN may lead to higher costs because new strategies are allowed in each time slot in DQN training. But overall, the results are stored within a stable range. Compared with that before DQN, the cost of day-ahead trading has been reduced by about 85%, and the cost of future trading has been reduced by about 61%. The cost of future trading we simulate is the average of the results in 30 days. The results prove that our mechanism can reduce the cost of pre-purchasing effectively.

To test the effectiveness in DQN, we set several groups of scaling factors for comparison purposes, i.e., to analyze how the greedy degree $\varepsilon$ affects the DQN result, we set two groups where $\varepsilon = 0.9$ and $\varepsilon = 0.99$. The results of different greedy degrees are shown in Fig. 7. The cost reduces faster with the smaller greedy degree, because a smaller greedy degree means more possibility to adopt a new strategy, which results in faster influence at the beginning of the simulation. However, as the simulation goes, the larger greedy degree, the more possible that DQN will make mistakes, so the cost with the larger greedy is larger than that with a smaller greedy degree in the late simulation.

To analyze the study rate $\alpha$, we set two groups in which $\alpha$ is equal to 0.01 and 0.1, respectively. As the simulation episodes go on, the results in DQN in day-ahead trading costs are shown in Fig. 8. As it is shown in the simulation, the smaller study rate results in the curve with slow convergence, which means the tardiness of learning. On the contrary, the larger study rate results in the curve with oscillation, which means the instability of learning. Therefore, the study rate we adopted is related to the episodes. In the beginning, the study rate is $\alpha = 0.01$. As the simulation goes on,
the study rate changes to 0.005 when the episode is larger than 5000 and changes to 0.001 when the episode is larger than 10000.

For comparison, we set two control groups in which our DQN algorithm is not utilized in pre-purchasing trade. Instead, in the first control group, the last-day algorithm (LDA) algorithm is adopted, which is to pre-purchase the amount totally the same as the purchased amount yesterday. In the second control group, REPs utilize the minimum charging algorithm (MCA), which is to store the real-time trading schedule and pre-purchase the power in each time slot with the minimum power. The simulation results of day-ahead training and future training are shown in Fig. 9 and Fig. 10, respectively. Obviously, LDA cannot reduce the trading cost because of the uncertainty of PEVs, which results in the difference of PEVs’ charging message every day. With the simulation goes on, the performance of our algorithm can get less cost than LDA and MCA, even though the result seems not better at the beginning. It proves that with full training, our algorithm can get high performance.

VII. CONCLUSION
To improve the revenue of the charging station in the REM, we have presented a solution through an allocation mechanism in real-time trading and a scheduling mechanism in pre-purchasing trading. Considering the fluctuation of the TOU power price, we have designed an operation mechanism to allocate charging power for PEVs. Focusing on the uncertainty of PEVs’ charging message, we have utilized DQN to predict the charging amount in real-time trading to reduce the operation cost. A series of simulations have been conducted to validate the superiority of the proposed mechanism. The numerical results have demonstrated that REPs can choose the best strategies and reduce the operation cost substantially compared with other mechanisms. As for the future work, we will study the optimal pricing mechanism when multiple REPs compete with each other in the REM.

REFERENCES
[1] J. Guiot and W. Cramer, “Climate change: The 2015 Paris agreement thresholds and Mediterranean basin ecosystems,” Science, vol. 354, no. 6311, pp. 465–468, Oct. 2016.
[2] Q. Huang, Q.-S. Jia, Z. Qiu, X. Guan, and G. Deconinck, “Matching EV charging load with uncertain wind: A simulation-based policy improvement approach,” IEEE Trans. Smart Grid, vol. 6, no. 3, pp. 1425–1433, May 2015.
TIANXIN LIN is currently pursuing the master’s degree with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China. He received the B.E degree from Nantong University, Jiangsu, China, in 2016. His research interests are in the general area of smart grids and machine learning.

ZHOU SU received the Ph.D. degree from Waseda University, Tokyo, Japan, in 2003. His research interests include multimedia communication, wireless communications, and network traffic. He is currently a TPC Member of some flagship conferences, including the IEEE INFOCOM, the IEEE ICC, the IEEE GLOBECOM, and so on. He received the Best Paper Award of the International Conference CHINACOM2008 and the Funai Information Technology Award for Young Researchers, in 2009. He is also the Chair of the Multimedia Services and Applications Over Emerging Networks Interest Group (MENI) of the IEEE ComSoc Society and the Multimedia Communications Technical Committee. He has served as the Co-Chair of several international conferences, including the IEEE VTC Spring 2016, the IEEE CCNC2011, and so on. He is also an Associate Editor of IET Communications and the IEICE Transactions on Communications.
QICHAO XU is currently pursuing the Ph.D. degree with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China. His research interests are in the general area of wireless network architecture and vehicular networks.

RUI XING is currently pursuing the master’s degree with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China. His research interests are in the general areas of wireless network architecture and vehicular networks.

DONGFENG FANG received the B.S. degree from the School of Astronautic, Harbin Institute of Technology, the M.S. degree from the School of Mechatronic Engineering and Automation, Shanghai University, and the Ph.D. degree in electrical and computer engineering from the University of Nebraska-Lincoln. She is currently an Assistant Professor with the Computer Science and Software Engineering Department, Cal Poly, San Luis Obispo. She is also affiliated with the Computer Engineering Program. Her research interests include wireless communications and networks, public safety communications, cyber security (wireless security, cyber-physical security, critical infrastructure security, 5G security, the Internet of Things (IoT) security), and privacy.

* * *