IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 28, 2020, accepted March 12, 2020, date of publication March 17, 2020, date of current version March 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981543

Energy Storage Arbitrage in Grid-Connected Micro-Grids Under Real-Time Market Price Uncertainty: A Double-Q Learning Approach

YUNJUN YU[1,2], (Member, IEEE), ZHENFEN CAI[1], AND YUSHUI HUANG[1]
[1]School of Information Engineering, Nanchang University, Nanchang 330031, China
[2]School of Artificial Intelligence, Nanchang University, Nanchang 330031, China

Corresponding author: Yushui Huang (huangyushui@ncu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61563034 and Grant 51967013, and in part by the International S&T Cooperation Program of China under Grant 2014DFG72240.

ABSTRACT Energy storage plays a significant role in improving the stability of distributed energy, improving power quality and peak regulation in the micro-grid system, which is of great significance to the sustainable development of energy. In grid-connected mode, energy storage is mainly used to reduce the operating costs of micro-grid. Real-time price arbitrage is an important source of energy storage revenue. It is feasible to design arbitrage strategies using Q-learning algorithm. Due to the overestimation of the Q learning algorithm, this paper proposes an arbitrage strategy method based on Double-Q learning. Compared with Q-learning algorithm, Double-Q learning can avoid overestimation and provide more stable and accurate arbitrage strategy for energy storage systems. Since the source of arbitrage in previous studies was limited to electricity prices alone, this paper considers joint arbitrage of electricity and carbon prices. The simulation results show that if adding fluctuate carbon prices to arbitrage sources, the arbitrage profits will increase by more than 110%.

INDEX TERMS Energy storage, micro-grid system, double-Q learning, carbon prices.

I. INTRODUCTION

Compared with large power systems, micro-grid refers to a small power distribution system consisting of distributed power sources, energy storage devices, energy conversion devices, and loads. Since the power supply in the micro-grid is mostly a distributed power supply with a small capacity, the micro-grid does not have a certain anti-interference capability [1]. Energy storage devices in the micro-grid can help quickly track fluctuations in renewable energy generation. And it can not only reduce the need for large grid power, but also provide the flexibility needed to integrate renewable energy generation into the power system [2]. Energy storage is an important part of micro-grid with functions such as load transfer, energy management and frequency regulation. More and more people are paying attention to its economic viability [3]. Real-time market price arbitrage is one of the important ways of energy storage revenue. Energy storage devices discharge at higher prices and charge at lower prices, making use of the price difference in the real-time market to make profits. As the amount of renewable energy generation continues to increase, market prices fluctuate more. Arbitrage income also increases and more and more research on energy storage arbitrage [4].

In recent years, more and more researchers have focused on the arbitrage of energy storage in the electricity market. Analysis by Rahul *et al.* showed that there are strong economic reasons for energy storage arbitrage in the New York area, and a large number of regulatory services throughout New York State [5]. Sioshansi *et al.* analyzed the arbitrage benefits of an energy storage device in PJM during the six years from 2002 to 2007 [6]. Considering the impact of fuel prices, transmission constraints, energy storage capacity and fuel structure, arbitrage gains increased significantly. Therefore, it is necessary to design a feasible arbitrage strategy. Connolly *et al.* found that the 24-optimal operation strategy

The associate editor coordinating the review of this manuscript and approving it for publication was Behnam Mohammadi-Ivatloo.

was the most profitable practical method of dispatching a typical PHES facility [7]. It can get very substantial profits for an existing PHES facility. Abdulla *et al.* proposed a stochastic dynamic programming approach [8]. This method uses the available predictions and considers the factors of affect system degradation to optimize the operation of the energy storage. Krishnamurthy *et al.* proposed a stochastic formulation of arbitrage profit maximization in the case of uncertain electricity market prices [9]. But this method need to forecast electricity prices and its performances has a strong correlation with the quality of the forecast. However, due to the highly stochastic of real-time market prices, it is notoriously difficult to forecast well [10]. Jiang and Powell used an approximate dynamic programming approach to derive the energy storage bidding strategy for the NYISO real-time market without need to predict price information [11]. But the main disadvantage of this method is that it is computationally expensive. Qin *et al.* proposed an energy storage control strategy based on online modified greedy algorithm under uncertainty [12]. Although this method does not need to predict the price, there is a problem of storage space limitation in practice. Wang and Zhang optimized the real-time electricity price arbitrage strategy based on Q-learning [13].

As one of the most popular reinforcement learning algorithms, Q-learning is a model-free learning method and provides a learning ability for intelligent systems to select the optimal action. It is currently applied in a variety of applications. Energy storage arbitrage strategy based on Q-learning is learned through an action-value function, and the value matrix is continuously updated during the learning process. Hasselt formally articulated and proved that there is an overestimation in using Q-learning algorithms. Therefore, to overcome this shortcoming, they proposed a Double-Q learning algorithm [14]. It can effectively avoid the disadvantages of overestimation in Q-learning. Huang *et al.* used Double-Q learning algorithm to design the DVFS selection method, and ultimately achieves the goal of reducing the energy consumption of real-time multi-core systems [15]. The results show that it is better than the Q-learning scheme. Liu *et al.* proposed a machine learning framework based on Double-Q learning algorithm [16]. Simulation results show that the proposed algorithm can achieve up to 19.4% and 6.7% gains in terms of the number of satisfied users compared to Q-learning algorithm. Therefore, we propose a strategy based on Double-Q learning for energy storage arbitrage which uses two estimators for approximation. This algorithm fundamentally reduces the overestimation existing when approaching the actual action value during the training process. It can obtain higher profit and provide more stable and accurate arbitrage strategy for energy storage system.

Some study results suggest that the revenue of energy storage has dropped in most European markets. Energy storage requires revenue from other markets [17]. The rapid development of the world economy has led to huge energy consumption, and the speed and total amount of greenhouse gas emissions have increased year by year. According to the analysis of relevant data by WRI (The World Resources Institute), the world's carbon dioxide emissions have increased at an annual growth rate of 2.4% in the past ten years. For this reason, all countries in the world are trying to find ways to mitigate the impact of the greenhouse effect on human beings. And reducing carbon emissions has become a major concern of the international public [18]. With the increase of environmental pollution pressure, more and more people are paying attention to the low-carbon economy with low emissions, low power consumption, and low pollution. Carbon trading is currently the main method to reduce carbon emissions and mitigate climate change. And it is an emerging product of the international environmental protection cooperation mechanism. As a large carbon dioxide emitter, the electricity industry has a profound impact on carbon trading especially in the optimization of power system scheduling. Therefore, we can consider adding carbon market to the arbitrage source of energy storage [19].

The comprehensive management of greenhouse gas emissions represented by carbon dioxide will cause cost differences to enterprises [20]. The difference can be measured as the general equivalent by the transaction amount per unit of carbon emission equivalent. Therefore, carbon trading has the function of exchange and value. With the function of value, carbon trading is a carbon currency in nature. On the one hand, emission reductions from different projects and companies can enter the carbon market for trading. On the other hand, the implementation of carbon currency can be developed into standard financial instruments, and even financial derivatives such as options and futures. There are many carbon trading markets in the world, of which the EU is the world's largest carbon demander, and its trading price largely reflects the trading price trend of the global carbon market [21]. The electricity industry accounts for 42% of global carbon dioxide emissions [22]. Many studies have focused on finding factors that affect carbon prices and the relationship between carbon trading markets and electricity markets. Many people have found that the impact of electricity and fuel prices on carbon pricing is significant [23]–[26]. If energy storage can participate in the trading of the carbon market, it will promote the investment development of the energy storage system.

The above issues prompt us to use reinforcement learning (RL) to develop a viable arbitrage strategy. Arbitrage strategy requires no price distribution and is superior to existing strategies. Without a clear assumption of distribution, our arbitrage strategy can arbitrage between two prices which may be unstable and constantly change. During training, through repeatedly performing charging and discharging operations at different real-time market prices, using reinforcement learning methods to learn the best strategy for maximizing cumulative rewards. Design of reward and punishment function is the main problem. This function will guide energy storage to make the right decision. The main contributions of this paper are as follows: We transform energy storage operations into Markov Decision

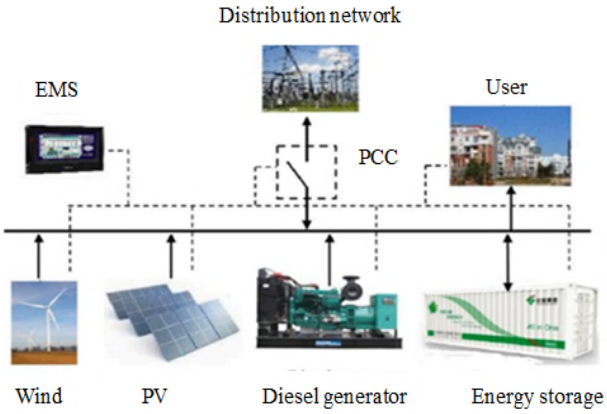


FIGURE 1. Structure diagram of micro-grid.

Process (MDP). An arbitrage strategy based on Double-Q learning is designed. The proposed arbitrage strategy based on Double-Q-earning uses two functions to decouple selection and evaluation, avoiding the overestimation caused by using actions in one function. Compared with the Q-learning algorithm, this method can get higher profits. After adding carbon price to arbitrage source, arbitrage profit increased by more than 110%.

II. ENERGY STORAGE ARBITRAGE MODEL

There are many types of structures in the micro-grid. This article presents a more common structure, as shown in Fig. 1. Each micro-grid contains multiple distributed power sources and energy storage systems that collectively power the load. The micro-grid as an external body is connected to the large grid or the upper-level substation through a PCC switch. From the perspective of load, micro-grid is an autonomous power system that can meet user's requirements for power quality and reliability. From a large grid perspective, micro-grid is equivalent to a generator or load in the grid [27]. The role of energy storage in micro-grid is mainly to improve the utilization of renewable energy, system stability, power quality and bring economic benefits. In the electricity market, when there is sufficient power in the micro-grid, the excess power can be sold to the large grid and gain corresponding profits. In the grid-connected mode, the energy storage device in the micro-grid can be discharged at high electricity prices and charged at low electricity prices to obtain profits. Price arbitrage is an important source of income for energy storage, but well-designed strategies are difficult due to the high degree of uncertainty in prices. Below we will build an energy storage arbitrage model.

We consider energy storage to operate within a limited time frame: T $\in (1 \ldots t)$. The purpose of energy storage is to obtain profits by discharged at high market prices and charged at low market prices. Assuming that the operation of energy storage will not affect market prices, we will express the problem of arbitrage maximization as follows:

$$max \sum_{t=1}^{T} P_t \cdot A_t \tag{1}$$

where T is the number of hours divided in the period. P_t is the real-time market price. A_t is the charge and discharge power of the energy storage system. The problem of arbitrage maximization is to maximize the sum of profits obtained from the charge and discharge actions of energy storage based on market prices.

The energy change and capacity limit of the energy storage system can be expressed as (2). E_t is the real-time energy level of the energy storage.

$$\begin{cases} E_t = E_{t-1} + A_t, & \forall t \in T \\ E_{min} \le E_t \le E_{max}, & \forall t \in T \\ A_{min} \le A_t \le A_{max}, & \forall t \in T \end{cases} \tag{2}$$

A_t can be further expressed as formula (3). μ is the charge and discharge efficiency of the energy storage system. D_t is the discharge power of energy storage. C_t is the charge power of energy storage.

$$A_t = \frac{1}{\mu} D_t - \mu C_t \tag{3}$$

The charge and discharge power of the energy storage system must meet the formula (4):

$$\begin{cases} C_t = \{0,\ I_c \cdot \min(C_{max}, E_{max} - E_{t-1})\} \\ D_t = \{0,\ I_d \cdot \min(D_{max}, E_{t-1} - E_{min})\} \\ I_d + I_c = 1 \end{cases} \tag{4}$$

The charge and discharge power of the energy storage system will follow the limits of Equation 2-4. Specifically, when energy storage decides to charge or discharge, it will charge and discharge at maximum charge rate or until it reaches the maximum or minimum energy level. I_d and I_c are 0-1 symbols that the energy storage system discharge or charges during time t. It can ensure that the energy storage cannot be charged and discharged at the same time.

The energy storage arbitrage maximization problem has four typical characteristics: (i) the action of energy storage is related to the market price and the energy level of energy storage is related to the past actions of energy storage; (ii) the purpose of energy storage actions is to maximized cumulative profit; (iii) energy storage does not know future prices, but know the price data in the past; (iv) actual prices are constantly changing.

Energy storage arbitrage requires maximize accumulated profits at non-stationary and constantly changing prices market prices. Due to the high degree of uncertainty in prices, it is difficult to design arbitrage strategies. The energy storage system needs to learn the historical data to get the current action strategy. Therefore, charge and discharge decision (C_t, D_t) is a function of market price information, and market price information can be multiple prices such as electricity and carbon prices. Energy storage participation in carbon market trading may lead to higher profits. $\pi(.)$ is an arbitrage strategy to maximize profit (5).

$$\begin{cases} W_t = (P_1, P_2, \ldots\ldots P_t) \\ (C_t, D_t) = \pi\ (W_1, W_2, \ldots\ldots W_t) \end{cases} \tag{5}$$

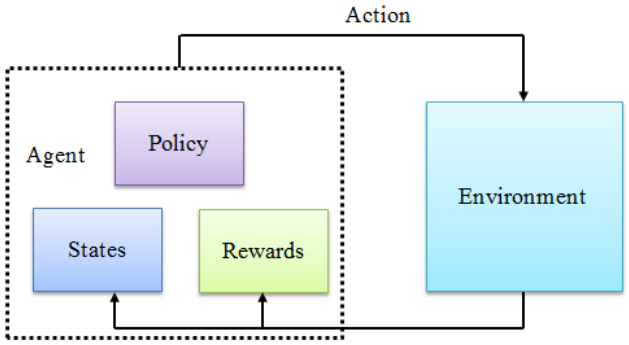


FIGURE 2. Base idea of reinforcement learning.

The problem of arbitrage maximization is a continuous decision problem under the constraints of market price and energy storage. This problem can be solved theoretically through dynamic programming or real-time price prediction. The high dimensionality of state space in dynamic programming makes the computational cost prohibitive so it is not suitable for real-time price arbitrage. At the same time, there are many factors affecting market prices, and it is difficult to predict prices [28], [29]. In order to solve the problem of maximizing energy storage arbitrage, arbitrage strategy is designed using reinforcement learning methods. We use Double-Q learning algorithm to solve the problem of maximizing energy storage arbitrage while avoiding the overestimation of Q-learning.

III. DOUBLE-Q LEARNING ALGORITHM FOR ENERGY STORAGE ARBITRAGE

Reinforcement learning (RL) is learning by continuously interacting with the dynamic environment so that the agent can obtain the maximum cumulative reward value from the environment. After the signal of reward and punishment have obtained, the agent will modify the action strategy to obtain a larger reward or a smaller penalty. The mechanism is shown in Fig. 2. Reinforcement learning is a way for agents to perform adaptive learning [30]. The problems that reinforcement learning can solve have the following characteristics: (i) the state of the system will affect the actions taken; (ii) maximize the cumulative reward through a series of actions; (iii) the system only knows the current and historical information; (iv) There may be factors of instability in the system. In order to solve the problem of maximizing energy storage arbitrage, we designed arbitrage strategy based on Double-Q learning algorithm.

The max operation in standard Q-learning uses the same value to select and measure an action. This is actually more likely to choose an overestimated value and can lead to suboptimal strategies. To avoid this situation, Hasselt proposed the Double-Q learning algorithm. The main idea of Double-Q learning is to use two estimators to decouple selection and measurement. Therefore, the algorithm can more accurately converge to the optimal operation value after iteratively updating. Q-learning will not be able to maximize profit because it will produce a sub-optimal arbitrage strategy. Therefore, the arbitrage strategy based on the Double-Q learning algorithm is more effective. Similar to standard Q-learning, Double-Q learning can interact with the environment and produce appropriate actions. In each process, after the agent perceives the complete state of the environment, it performs the corresponding operations. Then put the environment into a new state. The agent will receive a feedback to evaluate this state transition. Based on these previous experiences, the agent can easily determine the next proactive action with the maximum expected reward.

Applying the Double-Q learning method to energy storage arbitrage, we must first determine the state and action space of the system, design appropriate reward and punishment functions for the iterative algorithm.

A. STATE SPACE

The state space of the system includes the current market price P_t and the system's energy level E_t. We discretize the market price into M intervals and the energy level of the system into N intervals. Finally, the state input space contains a total of (M × N) states:

$$S = \{1, \ldots, M\} \times \{1, \ldots, N\} \tag{6}$$

B. ACTION SPACE

The action space of the system is mainly the charge and discharge operation of energy storage. Because energy storage cannot be charged and discharged at the same time, there are three main states in the system's action space:

$$A = \begin{cases} -\check{D}_{max} & if\ Discharge \\ 0 & if\ Idle \\ \check{C}_{max} & if\ Charge \end{cases} \tag{7}$$

$A = \check{C}_{max}$ express charge at the maximum charge rate C_{max} or until reach the maximum energy level E_{max}, $A = -\check{D}_{max}$ express discharge at the maximum discharge rate D_{max} or until reach the minimum energy level E_{min}.

C. REWARD

The setting of action rewards is especially important, which will directly affect the overall arbitrage results. At time t, the system will get a reward R after performing action A in state S. R can help the energy storage system measure how well its action is performing. We set the reward to:

$$R = \begin{cases} \left(\overline{P}_t - P_t\right)\check{C}_{max} & if\ Charge \\ 0 & if\ Idle \\ \left(P_t - \overline{\overline{P}}_t\right)\check{D}_{max} & if\ Discharge \end{cases} \tag{8}$$

$\overline{P}_t$ is the average price of the market price, which can be specifically expressed as:

$$\overline{P}_t = (1-\eta)\,\overline{P}_{t-1} + nP_t \tag{9}$$

η is a smoothing parameter. We use a moving average in (9) instead of simply calculating the average. This method can

not only use past price information, but also adapt to current price changes. When the current market price is lower than the average price, the action of charge will get a positive reward, and discharge will get a negative reward. Similarly, when the current market price is higher than the average price, the action of discharge will receive a positive reward, and charge will receive a negative reward. Therefore, energy storage will learn the ability to distinguish good or bad actions under the guidance of reward function. Our reward function can explore more arbitrage opportunities and get higher profits. It eases price instability because it weights current prices much more than historical prices.

D. DOUBLE-Q LEARNING ALGORITHM FOR ENERGY STORAGE ARBITRAGE

The basic settings of Double Q-learning algorithm have been given. The specific training process of energy storage arbitrage based on Double Q-learning algorithm will be explained below. In order to find the best arbitrage strategy, the proposed Double-Q learning algorithm needs to allow the agent to continuously interact with the environment, and finally obtain a table of Q values.

At time t, the energy storage determines the action of system based on the current status information. After interacting with the environment, energy storage can receive the reward R, and then observe the next state S_{t+1}. In the standard Q-learning, the update formula of the Q value can be expressed as:

$$Q_{t+1}(s,a) = Q(s,a) + \alpha\left(R + \gamma Q(s',a') - Q(s,a)\right) \quad (10)$$

Unlike standard Q-learning with only one update function, there are two update functions in Double-Q learning. Double-Q learning will use two functions Q_A and Q_B (corresponding to two estimators). And each function will update the next state with the value of the other function. It is important that both functions learn from different sets of experiences. You can use two value functions at the same time to choose the action to perform. Hence, the proposed algorithm can avoid the overestimation in Q-learning. To enable energy storage to record the values of Q-tables, the Q-tables need to update at each time slot t. Q_A and Q_B can be given as follows:

$$Q^A_{t+1}(s,a) = Q^A_t(s,a) + \alpha\left(R + \gamma Q^B(s',a') - Q^A_t(s,a)\right) \quad (11)$$

$$Q^B_{t+1}(s,a) = Q^B_t(s,a) + \alpha\left(R + \gamma Q^A(s',a') - Q^B_t(s,a)\right) \quad (12)$$

where γ is the discount factor, α is the learning rate, and s is the next state after taking action A at state S. During training, only one function will be selected for update. In order to update the value of Q table in (11) or (12), the energy storage needs to select an action to be performed at time t. The action selection strategy is ϵ-greed. The algorithm not only takes advantage of the best moves, but also explores other moves that might be better. Specifically, the action will be randomly selected with a probability of ϵ, and the optimal action will be selected with a probability of $(1-\epsilon)$. When the training is completed, the optimal action in the Q table will be selected each time to maximize the cumulative profit. As follows:

$$a^* = argmax_a Q(s,a) \quad (13)$$

The exact procedures of Double-Q learning algorithm in this article are shown in Algorithm 1. Fig. 3 gives the training process of energy storage arbitrage based on Double Q-learning algorithm. One of the functions is randomly selected to update. Where α is the learning rate and γ is the discount factor. Specifically, α determines the extent to which new information covers old information, and γ determines the importance of future rewards. The system randomly uses one of these two functions to determine the greedy update strategy, and the other function is used to calculate the value based on this selected strategy. Compared with the standard Q-learning algorithm, this algorithm is more stable and efficient because its update mechanism can actually reduce the overestimation between actual and approximate values. This is important to the correctness of choosing the right operation for energy storage.

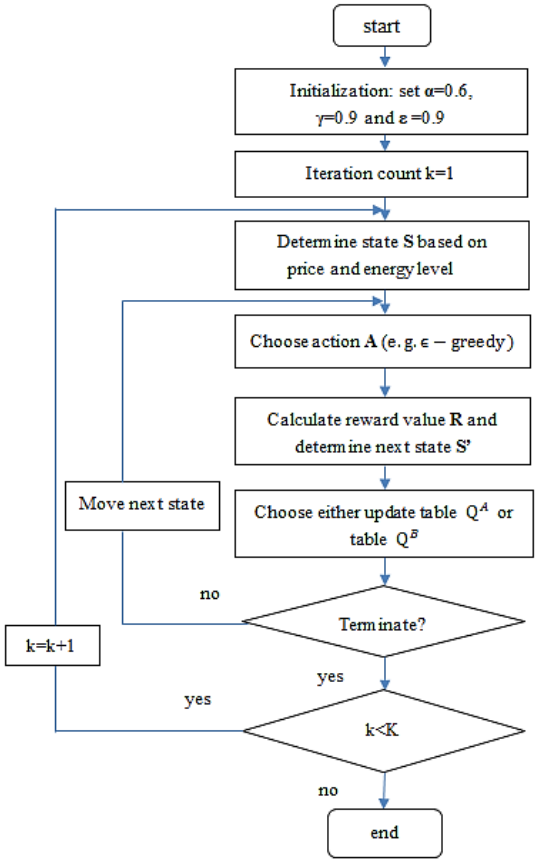


FIGURE 3. Flowchart for training process.

Algorithm 1 Double-Q Learning

Initialization
Repeat
  Observe state S;
  Choose A (using ϵ-greedy method) based on Q^A and Q^B observe R, S′.
Choose (in turn or randomly) either update table Q^A or table Q^B.
if update table Q^A
  Choose action A' = $\text{argmax}_{a'} Q^B(s', a)$ from Q^B
  Update table Q^A as given in (11).
else if update table Q^B
  Choose action a' = $\text{argmax}_{a'} Q^A(s', a)$ from Q^A
  Update table Q^B as given in (12).
end if
end if
  s ←s'
until end of operation.
end

TABLE 1. System parameters.

Rated Power	1MW
Rated Capacity	1 WMh
Maximum charge rate	1 C
Maximum discharge rate	1 C

IV. NUMERICAL RESULTS

A. ARBITRAGE OF ELECTRICITY PRICE

In order to show the results more intuitively, we consider using a small energy storage system in the micro-grid system and participating in the trading of the electricity market. In order to verify the effectiveness of energy storage arbitrage strategy based on the Double-Q learning algorithm, it was first applied to arbitrage in the real-time electricity market and compared with Q-learning algorithm. Parameters of the energy storage system are shown in Table 1.

According to the rate of charge and discharge, the charge and discharge capacity of energy storage per hour is 1WM. Source of market price data for energy storage arbitrage is very important. Therefore, the electricity price data is the real-time hourly Location Marginal Price (LMP) of a price node provided by PJM from 11/1/2017 to 11/1/2018 [31]. This date is very suitable for energy storage arbitrage. As shown in Figure 4, it can be observed that the price fluctuation is very obvious. The market price will change with the system load, power generation costs, system congestion, and renewable energy (wind, solar) conditions. Location Marginal Price (LMP) is a pricing model for spot electricity. This pricing model design is very reasonable or subtle. It combines the market with system operations, relies on the physical model (power flow model) of the power grid. It follows Security Constrained Unit Combination (SCUC) and Security Constrained Economic Dispatch (SCED).

To verify that the Double-Q Learning algorithm has more stable performance than the traditional Q-Learning

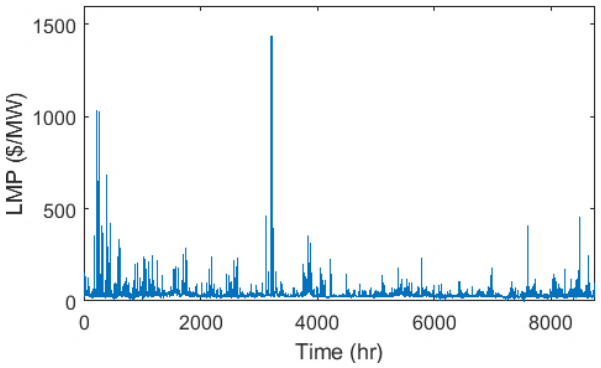


FIGURE 4. The real-time hourly LMP.

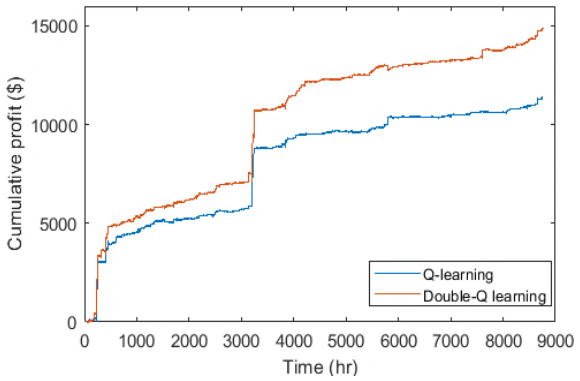


FIGURE 5. Cumulative profits under LMP.

algorithm, we divided the experimental data into four groups of 2190 hours each. As shown in Fig. 6(a-d). After experimenting with two algorithms on each set of data, compare the experimental results. The accumulative arbitrage results for each group are shown in the Fig. 7(a-d). The arbitrage profit for each set of data using Double-Q learning algorithm is higher than traditional Q-learning algorithm. Therefore, it can be concluded that our proposed scheme is better than the ordinary scheme and is more suitable for energy storage arbitrage.

These gains stem from the fact that the proposed algorithm aims to find the optimal arbitrage strategy to maximize the profits, while the Q-learning algorithm may result in sub-optimal policies which lead to a worse result. Therefore, we apply the two algorithms to the arbitrage experiment of one year of data and observe the overall performance of the algorithm. The arbitrage results are shown in Fig. 5. From the results below, it can be concluded that the Double-Q learning algorithm arbitrage profit is about 43% higher than the Q learning algorithm. From these experiment results, we can know the Double-Q learning algorithm is more stable than the Q learning algorithm in energy storage arbitrage, and the average profit is higher.

B. JOINT ARBITRAGE OF ELECTRICITY PRICE AND CARBON PRICE

As global greenhouse gas emission pressures increase, the carbon trading market continues to improve and expand. It is increasingly feasible to add carbon prices to energy storage arbitrage. Because our algorithm can reduce the losses caused by overestimation, we use this algorithm to design a

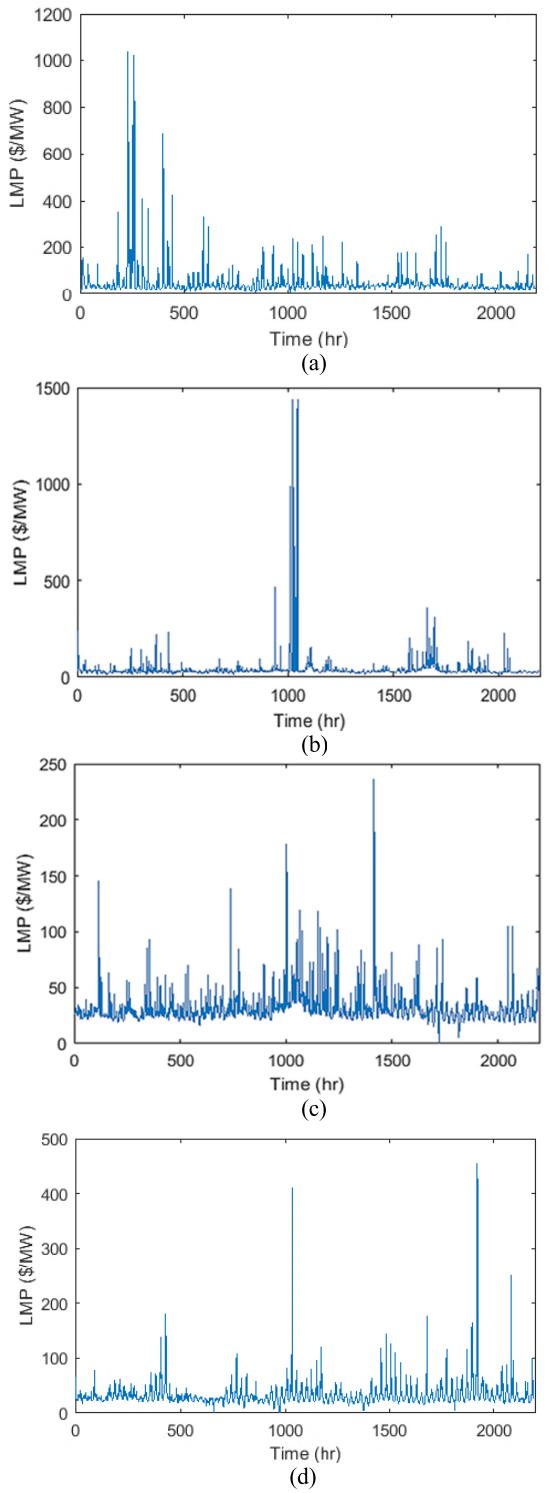


FIGURE 6. Electricity price for each group.

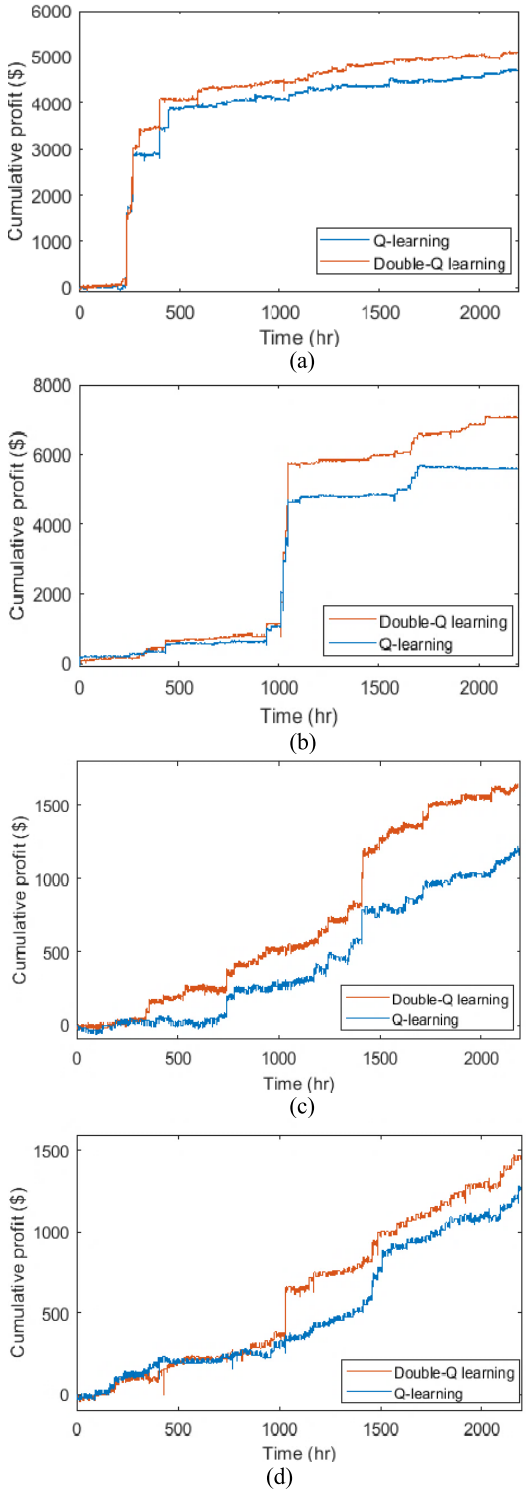


FIGURE 7. Results of the arbitrage experiment.

joint arbitrage strategy for energy storage. Then compare the experimental results of single and joint arbitrage.

In joint arbitrage, add carbon price information to the state space in the first. There are still three types of energy storage actions. The energy storage selects the appropriate action in the two price distributions to maximize the cumulative arbitrage profit. The problem of joint energy storage arbitrage maximization is a continuous decision under the constraints of price and system conditions. The goal is to maximize arbitrage in the two prices. In addition to the increase in state space information, the setting of the reward and punishment function is especially important. It is the key to judging which

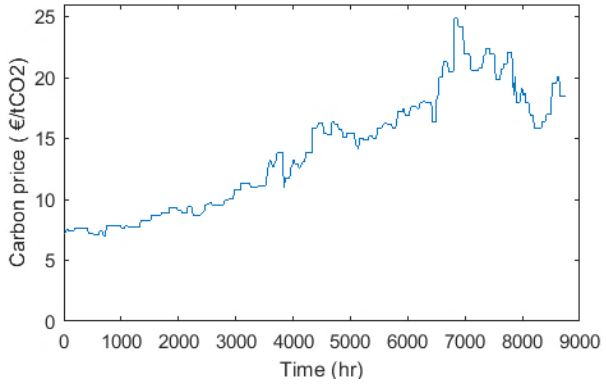


FIGURE 8. The real-time hourly carbon price.

TABLE 2. Cumulative profits of arbitrage.

Algorithm	Arbitrage source	Arbitrage profit($)
Q-learning	LMP	11351.2
	LMP and carbon price	26104.7
Double-Q learning	LMP	14278.3
	LMP and carbon price	31119.6

action can maximize the arbitrage from the two prices (14):

$$R = \begin{cases} \{(\overline{P} - P_t) + \rho\left(\bar{Q} - Q_t\right)\}\check{C}_{max} & if\ Charge \\ 0 & if\ Idle \\ \{(P_t - \bar{P}) + \rho\left(Q_t - \bar{Q}\right)\}\check{D}_{max} & if\ Discharge \end{cases} \tag{14}$$

where ρ is the discount rate of electricity and carbon price, and $\bar{Q}$ is the average price of carbon price.

The European Energy Exchange (EEX) in Leipzig, Germany, is one of Europe's largest carbon spot trading platforms. It has a high position in the electricity market, natural gas market and carbon emission rights market. As the EU carbon market stabilization mechanism effectively reduced the supply of carbon allowances in the market, the price of the EU carbon trading market continued to rise from 2017 to 2018 and reaching a record high. Therefore, carbon price is the European Union Allowance (EUA) price from 11/1/2017 to 11/1/2018 by EEX [32]. As shown in Fig. 8. Taking coal-fired power generation as an example, saving 1 KWh of electricity is equivalent to reducing emissions by 0.997 kg of carbon dioxide, so ρ takes 0.997. The discount for the euro and the dollar takes the most recent transaction price: $1 € = 1.1205\$(2019/05/16)$.

The parameters of the energy storage system are the same as above. We use Double-Q learning algorithm and Q-learning algorithm for joint arbitrage. The date of electricity price and carbon price are based on the data mentioned above for the whole year. The joint arbitrage of cumulative profits curve is shown in Fig. 9. The profit of electricity price arbitrage and joint arbitrage is shown in Table 2.

Through the arbitrage results, it can be found that the profit of joint arbitrage increase by more than 110% compared to the arbitrage of electricity price only. At the same time, arbitrage strategies based on Double Q learning have higher

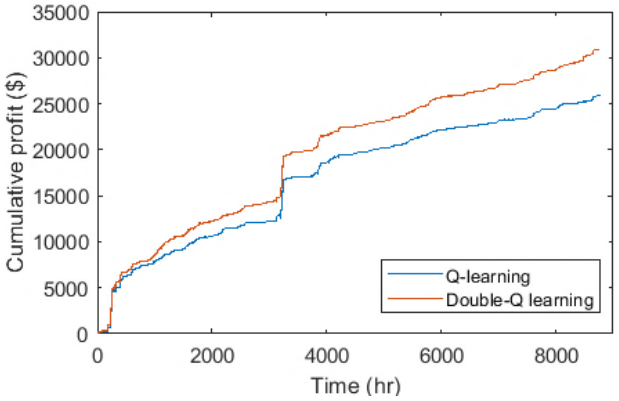


FIGURE 9. Cumulative profits of joint arbitrage.

profits than Q-learning. There are two main reasons. On the one hand, the performance of Double-Q learning algorithm is better than Q-learning. On the other hand, the setting of the reward function in joint arbitrage consists of two parts. Compared with the electricity price arbitrage only, the fluctuation of the value of the reward is increased, which makes the energy storage better to learn the action. It is equivalent to increasing the fluctuation of the electricity price, so that the energy storage can better distinguish the quality of the action. Although the gains from the carbon market are not high, this is beneficial to the overall gains. And if the carbon market price is more volatile, profits will increase further. This will provide new ideas for the increase of energy storage arbitrage profits, inspiring for subsequent research.

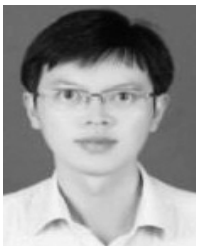
V. CONCLUSION

In this paper, we propose a reinforcement learning-based energy storage arbitrage strategy. One property of this scheme is to employ Double-Q learning algorithm. This algorithm has two estimators for actual value approximation so that it can reduce the uncertainty in the selection of optimal arbitrage strategy. From the results of Electricity price arbitrage, we can observe that the performance of our algorithm is more stable then Q-learning algorithm and can get higher profits. It shows that the overestimation has limited effect on our scheme during policy election. Applying our method to the joint arbitrage of electricity price and carbon price, we found arbitrage profits will increase by more than 110% compared to the arbitrage of electricity price only. This will bring potentially huge market benefits, prompting us to apply energy storage arbitrage to the carbon market. On the one hand, it can promote the development of the carbon market. On the other hand, it can increase the income of energy storage. And when the carbon price is more volatile, it will be more beneficial to energy storage arbitrage. Problems such as deterioration of the battery will be considered in future research.

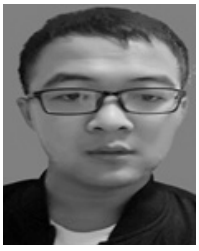
REFERENCES

[1] Z. Peng, L. Haitao, Z. Minglu, H. Sipeng, and L. Shuang, "Research overview of optimization of multi-source and energy storage in microgrid," in *Proc. 12th IET Int. Conf. AC DC Power Transmiss. (ACDC)*, Beijing, China, May 2016, pp. 1–5.

[2] G. C. Gissey, D. Subkhankulova, P. E. Dodds, and M. Barrett, "Value of energy storage aggregation to the electricity system," *Energy Policy*, vol. 128, pp. 685–696, May 2019.
[3] J. Eyer and G. Corey, "Energy storage for the electricity grid: Benefits and market potential assessment guide," *Sandia Nat. Lab.*, vol. 45, pp. 5–6, Jul. 2010.
[4] C. K. Woo, I. Horowitz, J. Moore, and A. Pacheco, "The impact of wind generation on the electricity spot-market price level and variance: The Texas experience," *Energy Policy*, vol. 39, no. 7, pp. 3939–3944, Jul. 2011.
[5] W. Rahul, A. Jay, and M. Rick, "Economics of electric energy storage for energy arbitrage and regulation in New York," *Energy Policy*, vol. 35, no. 4, pp. 2558–2568, Apr. 2007.
[6] R. Sioshansi, P. Denholm, T. Jenkin, and J. Weiss, "Estimating the value of electricity storage in PJM: Arbitrage and some welfare effects," *Energy Econ.*, vol. 31, no. 2, pp. 269–277, Mar. 2009.
[7] D. Connolly, H. Lund, P. Finn, B. V. Mathiesen, and M. Leahy, "Practical operation strategies for pumped hydroelectric energy storage (PHES) utilising electricity price arbitrage," *Energy Policy*, vol. 39, no. 7, pp. 4189–4196, Jul. 2011.
[8] K. Abdulla, J. de Hoog, V. Muenzel, F. Suits, K. Steer, A. Wirth, and S. Halgamuge, "Optimal operation of energy storage systems considering forecasts and battery degradation," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 2086–2096, May 2018.
[9] D. Krishnamurthy, C. Uckun, Z. Zhou, P. R. Thimmapuram, and A. Botterud, "Energy storage arbitrage under day-ahead and real-time price uncertainty," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 84–93, Jan. 2018.
[10] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *Int. J. Forecasting*, vol. 30, no. 4, pp. 1030–1081, Oct. 2014.
[11] D. R. Jiang and W. B. Powell, "Optimal hour-ahead bidding in the real-time electricity market with battery storage using approximate dynamic programming," *INFORMS J. Comput.*, vol. 27, no. 3, pp. 525–543, Aug. 2015.
[12] J. Qin, Y. Chow, J. Yang, and R. Rajagopal, "Online modified greedy algorithm for storage control under uncertainty," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 1729–1743, May 2016.
[13] H. Wang and B. Zhang, "Energy storage arbitrage in real-time markets via reinforcement learning," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Portland, OR, USA, Aug. 2018, pp. 1–5.
[14] H. V. Hasselt, "Double Q-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2613–2621.
[15] H. Huang, M. Lin, and Q. Zhang, "Double-Q learning-based DVFS for multi-core real-time systems," in *Proc. IEEE Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Jun. 2017, pp. 522–529.
[16] X. Liu, M. Chen, and C. Yin, "Optimized trajectory design in UAV based cellular networks for 3D users: A double Q-learning approach," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Chengdu, China, Feb. 2019, pp. 13–18.
[17] G. Ludovic and M. Kaveh, "Energy storage race: Has the monopoly of pumped-storage in Europe come to an end?" *Energy Policy*, vol. 126, pp. 22–29, Mar. 2019.
[18] M. Mehling and E. Tvinnereim, "Carbon pricing and the 1.5°C target: Near-term decarbonisation and the importance of an instrument mix," *Carbon Climate Law Rev.*, vol. 12, no. 1, pp. 50–61, Feb. 2018.
[19] X. Chen, T. Zhou, and X. Li, "Structure identification of CO_2 emission for power system and analysis of its low-carbon contribution," *Automat. Electr. Power Syst.*, vol. 36, no. 2, pp. 18–25, Jan. 2015.
[20] Z. Yang, M. Fan, S. Shao, and L. Yang, "Does carbon intensity constraint policy improve industrial green production performance in China? A quasi-DID analysis," *Energy Econ.*, vol. 68, pp. 271–282, Oct. 2017.
[21] Q. Ji, T. Xia, F. Liu, and J.-H. Xu, "The information spillover between carbon price and power sector returns: Evidence from the major European electricity companies," *J. Cleaner Prod.*, vol. 208, pp. 1178–1187, Jan. 2019.
[22] P. Da Silva, B. Moreno, and N. C. Figueiredo, "Firm-specific impacts of CO_2 prices on the stock market value of the Spanish power industry," *Energy Policy*, vol. 94, pp. 492–501, Jul. 2016.
[23] A. C. Christiansen, A. Arvanitakis, K. Tangen, and H. Hasselknippe, "Price determinants in the EU emissions trading scheme," *Climate Policy*, vol. 5, no. 1, pp. 15–30, Feb. 2005.
[24] E. Alberola, J. Chevallier, and B. Chèze, "Price drivers and structural breaks in European carbon prices 2005–2007," *Energy Policy*, vol. 36, no. 2, pp. 787–797, Feb. 2008.
[25] B. Hintermann, "Allowance price drivers in the first phase of the EU ETS," *J. Environ. Econ. Manage.*, vol. 59, no. 1, pp. 43–56, Jan. 2010.
[26] Q. Ji, D. Zhang, and J.-B. Geng, "Information linkage, dynamic spillovers in prices and volatility between the carbon and energy markets," *J. Cleaner Prod.*, vol. 198, pp. 972–978, Oct. 2018.
[27] Z. Guoping, W. Weijun, and M. Longbo, "An overview of microgrid planning and design method," in *Proc. IEEE 3rd Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Chongqing, China, Oct. 2018, pp. 326–329.
[28] T. T. Kim and H. V. Poor, "Scheduling power consumption with price uncertainty," *IEEE Trans. Smart Grid*, vol. 2, no. 3, pp. 519–527, Sep. 2011.
[29] S. Borenstein, "The long-run efficiency of real-time electricity pricing," *Energy J.*, vol. 26, no. 3, pp. 93–116, 2005.
[30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. 2011.
[31] *Hourly Real-Time LMP (Locational Marginal Price)*. Accessed: May 21, 2019. [Online]. Available: https://www.iso-ne.com
[32] *Carbon Price in EUA (European Union Allowance) Primary Market*. Accessed: May 21, 2019. [Online]. Available: http://www.eex.com/en/market-data/environmental-markets



YUNJUN YU (Member, IEEE) was born in Shangrao, China. He received the B.Sc. and M.Sc. degrees in control theory and control engineering from Nanchang University, China, in 2000 and 2007, respectively, and the Ph.D. degree from the Chinese Academy of Sciences, in 2013. He is currently an Associate Professor with the Department of Automation, Information Engineering School, Nanchang University. His research interests including photovoltaic forecasting, fault diagnosis, data-driven optimal control and it's applied in photovoltaic micro-grid systems, ADRC, and low-carbon electricity technology.



ZHENFEN CAI received the B.S. degree in automation from the Anyang Institute of Technology, Henan, China, in 2018. He is currently pursuing the M.S. degree in control engineering with Nanchang University, Jiangxi, China. His research interest mainly focuses on the application of reinforcement learning in micro-grids.

YUSHUI HUANG is currently a Professor with the Department of Automation, Nanchang University, engaged in the automation of power electronic equipment.