Hindawi
International Transactions on Electrical Energy Systems
Volume 2022, Article ID 6854620, 14 pages
https://doi.org/10.1155/2022/6854620

Research Article

Charging Station Management Strategy for Returns Maximization via Improved TD3 Deep Reinforcement Learning

Hengjie Li,[1,2] **Jianghao Zhu**,[1] **Yun Zhou**,[1,2] **Qi Feng**,[1] **and Donghan Feng**[1,2]

[1]*School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China*
[2]*Key Laboratory of Control of Power Transmission and Conversion (Ministry of Education), Shanghai Jiao Tong University, Shanghai 200240, China*

Correspondence should be addressed to Yun Zhou; yun.zhou@sjtu.edu.cn

Received 16 July 2022; Revised 4 November 2022; Accepted 21 November 2022; Published 15 December 2022

Academic Editor: Santoshkumar Hampannavar

Maximizing the return on electric vehicle charging station (EVCS) operation helps to expand the EVCS, thus expanding the EV (electric vehicle) stock and better addressing climate change. However, in the face of dynamic regulation scenarios with large data, multiple variables, and low time scales, the existing regulation strategies aiming at maximizing EVCS returns many times fail to meet the demand. To handle increasingly complex regulation scenarios, a deep reinforcement learning algorithm (DRL) based on the improved twin delayed deep deterministic policy gradient (TD3) is used to construct basic energy management strategies in this paper. To enable the strategy to be more suitable for the goal of real-time energy regulation strategy, we used Thompson sampling strategy to improve TD3's exploration noise sampling strategy, which greatly accelerated the initial convergence of TD3 during training. Also, we use marginalised importance sampling to calculate the Q-return function for TD3, which ensures that the constructed strategies are more likely to learn high-value experiences while having higher robustness. It is shown in numerical experiments that the charging station management strategy (CSMS) based on the modified TD3 obtains the fastest convergence speed and the highest robustness and achieves the largest operational returns compared to the CSMS constructed using deep deterministic policy gradient (DDPG), actor-critic using Kronecker-factored trust region (ACKTR), trust region policy optimization (TRPO), proximal policy optimization (PPO), soft actor-critic (SAC), and the original TD3.

1. Introduction

A well-designed energy management strategy for electric vehicle charging station (EVCS) can not only balance the peak charge and discharge of electric vehicles (EVs), reduce the idle rate of charging posts at charging stations, and ease traffic congestion during peak periods [1]; It can also expand the operating returns of charging stations, attract more companies to enter the operating market, form healthy competition, and accelerate the construction of smart and connected cities. Our research aims to develop a real-time regulation strategy to maximize the return of EV charging station operators in a certain area with photovoltaic (PV) and grid as power reserves and regulated by dynamic electricity price while satisfying the load demand of EVs in the area through the regulation of energy management strategy (EMS).

In the research area of energy management strategies for EV charging stations, most studies tend to build energy management strategies by adapting the planning and operation of EV charging stations. In Paper [2], an EMS is formed by combining approximate dynamic planning (ADP) and evolutionary algorithms (EA) to extend the battery life and reduce the communication requirements of the control system. In literature [3], the service quality constraint decision model for EVCS is first formulated using network evolution theory based on network evolution, followed by a game energy interaction based on the equilibrium model of the supply function (SFE), established as an optimization problem with equilibrium constraints, and

finally solved by a hybrid optimization algorithm based on differential evolution and interior point method. A novel multistage stochastic structure for parking lot aggregators (PLAs) is proposed in [4] to manage the flexibility of PEVs, providing functional flexibility to the power market and meeting the needs of intermittent power systems with high penetration rates. A two-stage stochastic planning is applied to model a hybrid charging station consisting of an electrolyzer, fuel cell, and hydrogen storage integrated with a photovoltaic system in paper [5], and a linear risk-constrained planning strategy is proposed to reduce the scheduling risk based on a framework that enables EV charging station operators to utilize the energy storage capacity of n-Grids for resource scheduling to maximize their returns from energy and storage product purchases is developed in literature [6].

However, previous research on energy management strategies for EVCSs has been limited to approaches based on fixed models. This type of fixed-model-based research approach requires the construction of a system model considering a wide enough range of influencing factors, including the meteorological conditions of the area where the charging station is located, the spatio-temporal model of the traffic network, the psychological behavior portrait of EV users, the charging and discharging loads of EVs, the generation and tariff forecasts of renewable energy generation, and the capacity and lifetime of the energy storage system. The detailed modeling of these environmental influences requires a considerable amount of work, while the granularity of the model is strongly correlated with the effectiveness of the modulation strategy, and most training data collection and research would violate user privacy. Furthermore, due to the great differences in the EV and environment user groups in different regions, a fixed model management strategy does not have good generalizability and often needs to rewrite a large number of environmental conditions and recalculate them when generalizing the application model. At the same time, too many variables that influence these models make the calculation time increase significantly with the growth of added variables, which reduces the speed of regulation of charging station management strategies and is detrimental to the economics of charging station operators.

Therefore, model-free energy management strategies based on reinforcement learning (RL) are becoming increasingly popular. RL-based energy management strategies do not rely on large amounts of data, and each modulated action relies only on the experience learned by the intelligence in previously constructed environments. In literature [7], using a RL-based scheduling strategy to regulate electric vehicles to address the uncertainty of electricity supply and demand, the RL-based regulation strategy can reduce the energy cost by 22.05%, 22.57%, and 19.33%, respectively, compared to the regulation strategies based on genetic algorithm, particle swarm optimization, and artificial fish swarm algorithm. In literature [8], a Markov decision process (MDP) formulation with a scalable state representation independent of the number of charging stations is proposed based on the RL framework, followed by fitting Q iterations using the batch RL algorithm to generate a charging strategy that reduces the charging cost of charging stations by 37% compared to previous methods. In literature [9], a feature-based linear function approximator is proposed to improve the state value function and further improve the efficiency and generalization ability of the RL algorithm, and the strategy significantly extends the profit and reduces the peak load of the grid in the numerical case. In literature [10], an online reinforcement learning model is developed based on a combination of adaptive heuristic criticism and the recursive least squares algorithm to improve the stability of charging stations and increase the total revenue of charging stations. The simulation results show that the effectiveness of the proposed solution in terms of utility power savings and charging station profitability is significantly improved. In literature [11], by building a Markov decision process model to characterize the time series of uncertainty and then using RL techniques based on deep deterministic policy gradients (DDPG) to analyze the impact of uncertainty on the charging strategy, the proposed charging strategy maximizes the return to the distribution system operator while satisfying all physical constraints. These works demonstrate the strong potential of RL as an emerging approach in the EMS field.

However, some inherent drawbacks of RL make it still difficult to apply to charging station energy management strategies, especially the large number of states that makes it difficult to simulate traditional RL in a way that gives each variable a corresponding action. Therefore, deep reinforcement learning (DRL) that leverages the powerful representational power of neural networks to fit Q-tables or direct fitting strategies to solve the problem of oversized state-action spaces or continuous state-action spaces is developed [12], DRL has become an effective and important approach for developing model-free and real-time management strategies for hybrid electric vehicles (HEVs) and Plug-in hybrid electric vehicles (PHEVs) [13]. A new multi-intelligence DRL approach is proposed in literature [14], which can compute scheduling solutions for multiple electric vehicles charging stations in a distributed manner while processing dynamic data that changes in real-time. In literature [15], Q-learning is combined with deep neural networks to propose a dual deep Q-network (DQN) model to controlling charging and discharging actions under the constraints of hourly available tariffs, and experimental results show a significant increase in profitability compared to conventional charging schemes for electric vehicles. In literature [16], to minimize the operating cost of business storage systems (BSS), the DDPG algorithm belonging to DRL is used to control multiple charging piles simultaneously, and a BSS model is proposed to determine the optimal real-time charging and discharging power of the charging piles, which reduces the operating cost of BSS. In paper [17], a modified long- and short-term memory (LSTM) neural network is used as a representation layer to extract temporal features of the tariff signal, followed by a DDPG algorithm to solve the MDP class problem, which significantly reduces the cost of the payment. In literature [18], Gaussian noise is added to the output of the actor

network to prevent the agent from adhering to a nonoptimal policy, and the sparse reward limitation is addressed by using two replay buffers to meet the user's demand for battery energy and reduce charging costs. In literature [19], a cloud-based multiobjective EMS is explored using a hybrid architecture with DDPG that improves the thermal safety of electrical equipment while minimizing system energy losses and aging costs.

Although most DDPG-based EVCSs have performed well in research in related fields, some drawbacks inherent in the DDPG algorithm can still cause some impacts in practical applications. DDPG evolved from the DQN series. Although the dual critic architecture of double deep Q-network (DDQN) is used to calculate the Q-value, there is still a case of overestimating the Q-value in practice, which makes the algorithm converge too early or fails to find the optimal policy. When the Q-network is continuously updated, the actor may act according to the previous state to reach the expected maximum Q value, but after the Q-network is updated, the expected Q value is found to be wrong, but the wrong Q value has already guided the actor to choose the wrong action at the next moment. The twin delayed deep deterministic policy gradient (TD3) [20] algorithm based on DDPG uses three techniques, clipped double-q learning, delayed policy updates, and target policy smoothing, to improve and optimize the above problem, so the more well-performing TD3 algorithm is widely used in related fields. A differentiated pricing mechanism for multiservice PEV charging infrastructure (EVCI) is developed in literature [21] using the TD3 algorithm, which adaptively adjusts the service pricing of multiservice EVCI to maximize charging facility utilization while ensuring higher service quality satisfaction. In literature [22] a scheduling policy is constructed using TD3 and then the trained policy is deployed online to execute multiple actions simultaneously for coordinating the scheduling of mobile energy storage systems (MESS) and the integrated services of microgrid resource scheduling. In literature [23], an algorithm based on TD3 trajectory design for time minimization tasks (TD3-TDCTM) is proposed, which enables the shortest path for the UAV-IoT interaction process. In literature [24], a smart EMS for HEV is formulated using TD3 and a local controller based on heuristic rules (LC) is embedded in the DRL loop to eliminate l unreasonable torque distribution considering the characteristics of the components of the powertrain, after which a hybrid empirical replay (HER) method based on the hybrid empirical buffer (MEB) is proposed. Compared to other DRL-based EMS systems, the improved TD3 EMS system achieves the best fuel optimality, the fastest convergence speed, and the highest robustness under different operating conditions. In literature [25], a TD3-EMS for hybrid electrically driven rail vehicles is proposed that achieves a favorable balance between battery charging and discharging while minimizing hydrogen consumption and fuel cell aging costs and slows down fuel cell degradation.

Figure 1 shows the architecture of the actual operating environment of the DRL management strategy based on the modified TD3 composition, which consists of a CSMS strategy generation center and multiple distributed EVCSs with PV systems and ESSs. In this regulation model, the distributed EVCSs collect local data and perform preprocessing, after which the data are fed into the strategy generation center in real-time. The strategy center uses the improved TD3-based algorithm to construct a management strategy based on the principle of maximizing returns while satisfying the load demand of each charging station in the region, and the management strategy outputs the amount of electricity purchased by each EV charging station on the grid, the amount of electricity used by each PV system actions of EVCS and regulation actions of energy storage of EVCS, and all distributed EVCS in the area of these actions.

This study embodies an improved TD3-based CSMS. Compared with existing studies, this present paper encompasses four perspectives that may possibly contribute to relevant research:

The system integrates several functions, including shear double-q learning of critics, delayed policy updates, and smooth regularization of target policies, resulting in a CSMS based on TD3. To the authors' knowledge, this is one of the pioneering works on charging station management policy formulation for EVCS based on the TD3 algorithm.

To speed up the initial convergence of the management strategy at training, the initial exploration noise is determined using Thompson sampling-based sampling for the selection of the initial exploration strategy action, which results in a substantial increase in the initial convergence speed of the strategy.

A new way of calculating Q values was used to improve TD3, using marginalised importance sampling (MIS) to calculate Q values, making the selection of Q values more reasonable, avoiding the actor network from picking to the next highest point during the action, and increasing the stability increase of CSMS.

The performance of the improved TD3 is comprehensively compared with the original TD3, DDPG, trust region policy optimization (TRPO), proximal policy optimization (PPO), actor-critic using Kronecker-factored trust region (ACKTR), and soft actor-critic (SAC), and then a DRl-based CSMS is constructed in the same environment and comparatively analyzed to clarify the advantages of the CSMS constructed based on the improved TD3 in terms of convergence speed, robustness, and returns.

The remainder of the paper is structured as follows: in Section 2, the structure of a distributed EVCS containing a PV system, an energy storage system, a data acquisition system is briefly described, and a detailed cost constraint model is calculated for the construction of the CSMS. Section 3 describes the system design of the CSMS based on the improved TD3 and the improved training convergence speed by means of a new calculation method with Thompson sampling and determined Q-values. In Section 4, numerical experiments are conducted on the CSMS based on the improved TD3 and a comparison with six other DRL-based CSMSs is presented. Section 5 summarises the conclusions obtained in the study.

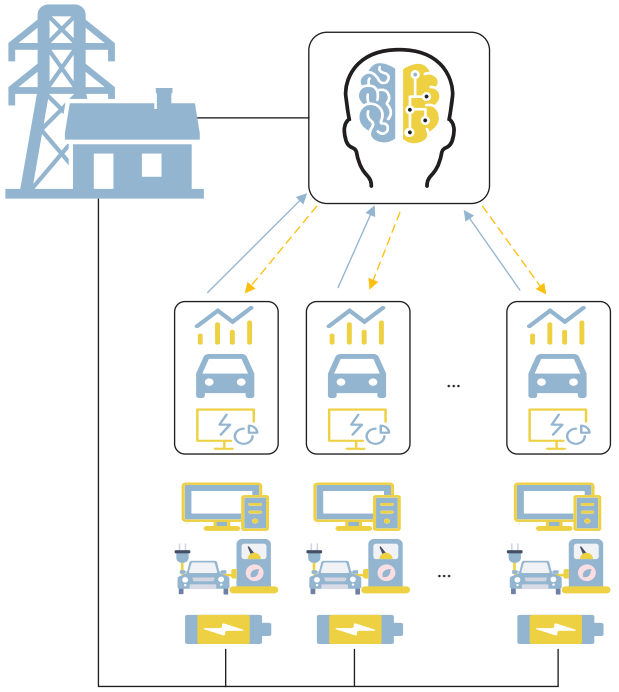

Figure 1: Framework for a charging station management strategy for regional operators with multiple charging stations.

2. Modeling of EVCS

To increase the return to regional charging operators and to increase the local consumption rate of PV, we propose a regional EV charging station energy management framework called twin delayed real-time management strategies (TDRMS), in which a modified TD3 algorithm is used to construct charging station management strategies to control the power purchase actions and energy management actions of multiple EVCSs in the region. In the TDRMS, the operation flow consists of the following two phases. In the first stage, each distributed independent charging station collects the condition information of its location and sends it to the dispatch center. The condition information includes weather conditions, grid power purchase price, PV generation forecast, charging postidle rate, charging station storage energy, EV charging power, and EV charging load forecast. In the second stage, the dispatching center takes all the information collected from each distributed charging station, merges the real-time power sales price of the grid as the environmental conditions for DRL strategy training, generates the management strategy for each distributed charging station using the improved TD3 algorithm, and distributes the strategy to each distributed charging station.

The DRL-based CSMS is theoretically model-free and has minimal dependence on data. Therefore, for the DRL problem, an explicit and accurate construction of the environment state where the agent is trained is extremely necessary, and an accurate environment model can make the generated policies closer to realistic application scenarios. In this section, a distributed smart EV charging station controlled by a real-time grid tariff and containing a PV system is modeled in terms of both energy management and operating cost.

2.1. Modeling of Energy Flows. In the framework established in this study, the operating environment model of EVCS in the region is shown in Figure 1, where the energy flow process mainly consists of purchasing power from the grid, the PV system of CS responding to the load demand of EV owners, and the CS storing energy and responding to the load demand of EV owners at a specific time. The energy flow model in the whole region is divided into two categories, the stored energy model and the transmitted energy model, and their models are constructed separately and input as state variables in DRL.

In the regulation strategy of this study, the group of EVs that are undergoing control is divided into five categories, such as battery electrical vehicle (BEV), HEV, PHEV, extended-range electric vehicles (EREV), and fuel cell electric vehicle (FCEV), as shown in Table 1.

In the operational framework established in this study, the primary power for each EVCS comes from the grid under dynamic pricing, and the supplementary power comes from the PV system equipped with each EVCS. Each distributed charging station has an independent energy management system, which is responsible for recording the real-time energy demand in the area, later forecasting the charging load of EV users in the area, and recording and regulating the charging station's energy storage system after receiving the regulation strategy. The method for accurate prediction of the charging load demand of EV users and power generation of the PV system has been done in our previous work [26], so here the power generation of the PV system and charging load demand of EV users are assumed as the dynamic variables for accurate prediction.

Here, the sum of the charging load predictions for EVCSs is $E_t^{\mathrm{EV}} = [E_{1,t}^{\mathrm{EV}}, E_{2,t}^{\mathrm{EV}}, \ldots E_{n,t}^{\mathrm{EV}}]$, $E_{i,t}^{\mathrm{EV}}$ denotes the electric vehicle charging load predicted by distributed charging station i at moment t. The total forecasted electricity production of PV systems with EVCSs is $E_t^{\mathrm{PV}} = [E_{1,t}^{\mathrm{PV}}, E_{2,t}^{\mathrm{PV}}, \ldots E_{n,t}^{\mathrm{PV}}]$, $E_{i,t}^{\mathrm{PV}}$ denotes the PV generation of the charging station at the moment t predicted by distributed charging station i. The total fixed consumption load of EVCSs is $E_t^{\mathrm{Fixed}} = [E_{1,t}^{\mathrm{Fixed}}, E_{2,t}^{\mathrm{Fixed}}, \ldots E_{n,t}^{\mathrm{Fixed}}]$, $E_{i,t}^{\mathrm{Fixed}}$ denotes the fixed load electricity consumption of charging station i at moment t. The sum of all charging station unrealized charging load at moment t for EVCSs is $E_t^{\mathrm{EV,uf}} = [E_{1,t}^{\mathrm{EV,uf}}, E_{2,t}^{\mathrm{EV,uf}}, \ldots E_{n,t}^{\mathrm{EV,uf}}]$, $E_{i,t}^{\mathrm{EV,uf}}$ denotes the amount of unfulfilled charging load at the charging station of distributed charging station i at moment t. $E_{i,t}^{\mathrm{EV,total}}$ denotes the sum of the predicted energy demand of distributed charging station i at moment t and the previously unfulfilled energy demand of electric vehicles. Meanwhile, it is also necessary to represent the power stored in each distributed EVCS. The sum of the stored power of EVCSs is $\mathrm{SOE}_t = [\mathrm{SOE}_{1,t}, \mathrm{SOE}_{i,t}, \ldots, \mathrm{SOE}_{i,t}]$, $\mathrm{SOE}_{i,t}$ denotes the amount of power stored at distributed charging station i at moment t. There are two parts of losses in the operation of the energy storage system, namely charging losses and discharging

Table 1: Classification of electric vehicles.

Classification	Description	Included/not included in the model
Typical EV I	BEV	Included
Typical EV II	HEV	Included
Typical EV III	PHEV	Included
Typical EV IV	EREV	Included

losses, which are mainly influenced by the cycle efficiency η. The losses of the energy storage system also need to be accounted for in the state variables, which are calculated as follows:

$$E_{ec,t} = \sum_{i=1}^{N} \eta \times E_{i,t}^{G},$$
$$E_{\mathrm{edc},t} = \sum_{i=1}^{N} \frac{E_{i,t}^{EV,\mathrm{total}}}{\eta}. \tag{1}$$

2.2. Modeling of Costing. To maximize the return for regional charging station operators, it is equally important to model all costs involved in the operation process when building an agent's training environment. P_t^{grid} denotes the dynamic electricity price of the grid at time t, P_t^{PV} denotes the price per unit of electricity saved by the PV system at moment t under the dynamic tariff. The access to photovoltaic systems and the regulation of energy storage in charging stations reduces the carbon emissions to the environment and saves the treatment costs of traditional power generation methods, thus generating environmental benefits P_t^{En}, which are calculated as follows:

$$P_t^{\mathrm{En}} = P_{\mathrm{poll}} \cdot \sum_{i=1}^{N} E_{i,t}^{\mathrm{PV}}, \tag{2}$$

where P_{poll} is the cost of treatment required per unit of electricity generated by conventional power generation methods for a fixed value of pollutants. The life of the energy storage system decreases as the number of charges and discharges increases, so the net annual value (NAV) over the life cycle of energy storage and the cost of loss of energy storage per unit need to be calculated. At the same time, the life decay of the energy storage system brings additional costs of system overhaul, system maintenance, and energy storage system replacement during the operating cycle. The energy storage cost model is analyzed using the full life cycle, and its calculation formula is shown as follows:

$$C_{\mathrm{RE}} = C_{\mathrm{SE}} \sum_{k_r=1}^{K_{\mathrm{R}}} \frac{(1-\alpha)^{k_r r}}{(1+i_c)^{k_r l}},$$
$$K_{\mathrm{R}} = \frac{N}{L} - 1,$$
$$C_{\mathrm{FCSNPV}} = C_{\mathrm{IC}} + C_{\mathrm{RE}} + \sum_{j=1}^{n} \frac{C_{\mathrm{IS},j}}{(1+i)^{j}} + \frac{C_{\mathrm{end}}}{(1+i)^{n}}, \tag{3}$$
$$C_{\mathrm{FCSNAV}} = C_{\mathrm{FCSNPV}} \frac{i(1+i)^{N}}{(1+i)^{N}-1},$$

C_{RE} is the replacement cost of energy storage batteries; α is the percentage of annual decrease in battery cost; K_{R} is the number of battery replacements; C_{SE} is the unit cost of the battery. N is the battery energy storage plant operating cycle; i_c is the discount on the cost of purchasing the battery; L is the energy storage battery replacement cycle; C_{IC} a is the fixed investment cost of the energy storage plant, including the personnel cost, equipment input cost, charging station construction cost, etc., during the operation life cycle; $C_{\mathrm{IS},j}$ is the overhaul cost in year j; i is the discount rate. C_{end} is the disposal cost of the energy storage system at the end of life. C_{FCSNPV} is the whole-life net present value (NPV) of EVCS, which is the sum of the present value of the net cash flows occurring in each year of the entire calculation period discounted by a predetermined or a set discount rate to the start of the investment program, respectively and can reflect the profitability of EVCS; C_{FCSNAV} is the net annual value of the whole-life cycle of EVCS, which is the equivalent annual value converted from the equivalent net cash flow during the calculation period of the project with a certain base rate of return and can determine whether the EVCS under this scenario has investment value [27]. It is also necessary to calculate the operating return of EVCS power sales, which is shown in the following formula:

$$C_{\mathrm{RETURN}} = E_t^{\mathrm{EV,total}} P_{i,t}^{\mathrm{sell}} + P_t^{En} - \left(E_t^{\mathrm{EV,total}} + E_t^{\mathrm{Fixed}} + \mathrm{SOE}_t - E_{i,t}^{\mathrm{PV}}\right) P_t^{\mathrm{grid}}. \tag{4}$$

The meaning of the variables is the same as described in the previous section, and the CSMS based on the modified TD3 is trained with the goal of maximizing C_{RETURN} by judging the value of C_{FCSNAV} for feasibility analysis and finding the CSMS that maximizes C_{RETURN}.

3. Methodology on Framework Design of Td3-Based CSMS

3.1. Preliminary Formulation of TD-3 Based CSMS

3.1.1. Brief Review of DDPG. DDPG is a modified form of deterministic policy gradient algorithm, which combines DPG and DQN in DRL [28]. The loss function of DQN is shown as follows:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s') \sim U(D)}\left[\left(r + \gamma \max_{a'} Q(s', a'; \theta_i^-) - Q(s, a; \theta_i)\right)^2\right]. \tag{5}$$

When finding the action with the maximum value, it is almost impossible to find the action with the maximum value if the action space is very large, or even if the action in the space of consecutive actions needs to be solved [29]. Even if the continuous space is discretized to find the approximate solution, only a very low solution efficiency is obtained [30]. But in DPG policy gradient does not have the above-mentioned maximization operation, avoiding this problem, and the formula is shown as follows:

$$\begin{aligned}\nabla_{\theta^\mu} J &\approx \mathbb{E}_{s_t \sim \rho^\beta}\left[\nabla_{\theta^\mu} Q\left(s, a|\theta^Q\right)\Big|_{s=s_t, a=\mu\left(s_t|\theta^{\mu^k}\right)}\right] \\ &= \mathbb{E}_{s_t \sim \rho^\beta}\left[\nabla_a Q\left(s, a|\theta^Q\right)\Big|_{s=s_t, a=\mu\left(s_t\right)} \nabla_{\theta^\mu} \mu\left(s|\theta^\mu\right)\Big|_{s=s_t}\right].\end{aligned} \tag{6}$$

In DQN, finding the next action to take is probabilistically distributed, so the action with the highest value must be found when updating the Q-value [31]. In contrast to DQN, DPG uses a deterministic policy, thus avoiding the operation of finding value-maximizing actions [32]. Consequently, DDPG combines the advantages of DQN and DPG and uses the two tricks used by the neural network in DQN to fit the Q values in DPG, turning the Q values in DPG into a neural network as well. In addition to combining the advantages of DQN and DPG, DDPG also adds three additional tricks for optimizing the learning process.

First, DDPG adopts a soft-mode network parameter update method, in which the parameters in the target net are updated at each step, but only a small portion of them are updated proportionally, thus greatly improving the stability of the learning process, expressed in the following equation:

$$\begin{aligned}\theta Q' &\leftarrow \tau\theta Q + (1-\tau)\theta Q', \\ \theta\mu' &\leftarrow \tau\theta\mu + (1-\tau)\theta\mu'.\end{aligned} \tag{7}$$

Due to the hyperparameter $\tau \ll 1$, the target network changes slowly and smoothly, which improves the stability of training. Here, the Q-value is learned by the bellman equation in the DQN with the following equation:

$$Q'(s, a) = E\left[r(s, a) + \gamma Q'\left(s', a'\right)\right]. \tag{8}$$

DDPG introduces batch normalization (BN) to solve the problem that different inputs have different magnitudes of units and data ranges for their features. Bn belongs to global adjustment, which linearly transforms the inputs of each layer and can restrict the parameters to run on different batches to a certain extent to prevent the gradient direction from competing into the vicious competition. Bn's will introduce mini information of other samples within the batch, resulting in predicting an independent sample with other sample information equivalent to the canonical term, making the loss surface smoother and easier to find the optimal solution. It is equivalent to one independent sample prediction that can look at multiple samples, and the learned features are more generalized.

Thus, the standard DDPG can be seen as an evolution of DQN for handling control tasks with continuous action spaces, especially for real-time control of a single target. However, DDPG also has some inherent limitations that need to be improved. As a deterministic policy, the trajectory generated with a deterministic policy is always fixed for a given state s and policy parameters, so the agent cannot explore other trajectories or consider other states, and the agent cannot find the optimal policy under many missions.

Therefore, an off-policy learning approach was adopted, using an actor-critic-based framework, where the actor action strategy uses a stochastic strategy to ensure sufficient exploration, and the critic evaluation strategy is deterministic, using a function approximation method to estimate the value function, while noise is introduced in it to add randomness.

3.1.2. EMS Based on the Improved TD3. All RL algorithms based on Q-value learning have the problem of overestimation for Q-values when maximizing Q-values including noise [33]. As a result, DDPG may become more unstable [34] in the face of complex control tasks, heavily dependent on the hyperparameters searched for the task at hand, and the generated model does not generalize well to the same types of problems [35]. The risk of overestimating the Q value in the critical network in the DDPG algorithm can lead to the accumulation of estimation errors as the training process proceeds to make the intelligence fall into a local optimum or suffer catastrophic oblivion [36]. The emergence of TD3 alleviates the problem of overestimation bias, while TD3 shows a significant improvement in learning speed and performance in complex continuous control domains compared to DDPG [37].

In the actor-critic framework of RL, parameter updates of the strategy function are related to the valuation function's estimate of the action value and always lead to an overestimation bias. This theoretical overestimation of the Q-value function that would occur also appears in the 2015 state-of-the-art (SOTA) algorithm DDPG [28]. In contrast to standard Q-value learning, TD3 learns the strategy in double Q-learning [38], choosing the lower value of the two valuation networks as a way to mitigate the overestimation bias for the value target. The high variance of the valuation makes the gradient of the strategy update noisy, which reduces the training speed of the model and affects the final training quality. To minimize the error at each iteration of the intelligence, TD3 proposes a strategy called delayed policy updates. By updating the policy network less frequently than the valuation network, TD3 reduces the estimation error before performing policy updates, ensuring that the TD-error becomes sufficiently small. Eventually, TD3 has faster training speed, better training results, and is easier to implement in any simulation environment compared to DDPG.

Therefore, in this article, the TD3 algorithm is selected as the basis for building a distributed CSMS, and the network architecture of the improved TD3 is shown in Figure 2.

First, TD3 uses two Q-value networks to calculate the state value at the next moment, as shown in the following equation:

$$\begin{cases} y_1 = r + \gamma Q_{\theta_2'}\left(s', \theta_{\mu 1}(s)\right), \\ y_2 = r + \gamma Q_{\theta_1'}\left(s', \theta_{\mu 2}(s)\right). \end{cases} \tag{9}$$

Although there is the possibility of both overestimation and underestimation of Q values, the true value is usually overestimated after adding noise for maximizing the estimate, and even in some regions of the state space, the overestimation is further exaggerated. Therefore, choosing the smaller of the two estimates as the target Q value can offset the overestimation of the Q value and substitute it into the bellman equation to calculate the TD error and loss function, as follows:

$$\begin{aligned} y &= r + \gamma \min_{i=1,2} Q_{\theta_i}\left(s', \theta_{\mu 1}(s)\right), \\ L_{ki} &= \frac{1}{M}\sum_{j=1}^{M}\left(y_j - Q_i\left(s_j, a_j\right)\right)^2. \end{aligned} \tag{10}$$

Although this rule of selecting lower Q values for updating may lead to underestimation bias for the standard Q-learning method, the underestimated actions are not propagated explicitly through the update of the strategy. Thus, the error is substantially reduced compared to the previous update rule that would lead to a constant accumulation of overestimation bias.

Secondly, a target network is established as a depth function approximator. Deep neural networks require multiple gradient updates to converge and fit the target, while updating using the target network provides a stable target, thus allowing the network to fit a larger range of training data. The update frequency of the strategy network is set lower than that of the valuation network to ensure that the delayed update of the strategy is performed only after the value error has been minimized before the strategy is updated. An adequate delayed update strategy limits repeated updates by critics and uses a lower variance when performing Q-value estimation, so the quality of strategy updates becomes higher, substantially increasing stability when training actors:

$$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'. \tag{11}$$

Finally, the use of a regularization strategy for smoothing the selection of the target strategy can make the Q values selected for the pre-estimation more stable. In a practical application, a noise needs to be added to the range of the original Q-value calculation, and the range values around the estimated Q-value are used to merge the original Q-value for the pre-estimation calculation:

$$\widetilde{a} \sim \theta_\mu(s) + \widehat{\theta}_{(l)}, \widehat{\theta}_{(l)} \sim \mathrm{clip}\left(W\left(x_l\right)\mathcal{N}\left(0, \widetilde{\sigma}\right), -c, c\right). \tag{12}$$

3.2. Improvement of Actor Noise. Also as a deterministic policy gradient algorithm, TD3 adds many unique hyperparameters compared to DDPG, which makes the training process more controllable and also increases the training speed and stability of the generated model. Therefore, the selection strategy of hyperparameters becomes more important in influencing the change of model performance.

The exploration noise is a hyperparameter unique to the TD3 algorithm and is used to adjust the variance of the noise attached to TD3's exploration actions during the exploration process. TD3 explores the action space in such a way that it is easy to explore the boundary actions in the space. In most RL task scenarios, the optimal strategy exists in the boundary actions, and TD3 often shows good training speed in such application scenarios. In the exploration, TD3 first adjusts the output tensor in the strategy network to (−1, +1) after the activation function and afterwards adds a noise parameter to the action that can be adjusted directly by exploring the noise variance after the clip, finally performs another clip operation on the action to adjust it to the interval (−1, +1) for performance. However, the selection of noise values has a greater impact on the actions between clip to (−1, +1) after increasing noise. Too small noise variance makes it difficult to explore suitable boundary actions during exploration and leads to inefficient exploration actions, increasing the training time of the model, reducing the speed of modulation, and making it more difficult to explore suitable actions during exploration. Excessive noise variance causes the exploration activities to be extremely biased towards the boundary action, which may reduce the diversity of strategy selection and the generalization performance of the model in application scenarios of different missions.

In this article, Thompson sampling is used to optimize the process of selecting the exploration noise variance. Thompson sampling is a natural stochastic Bayesian algorithm that is easy to implement and generalize and is not

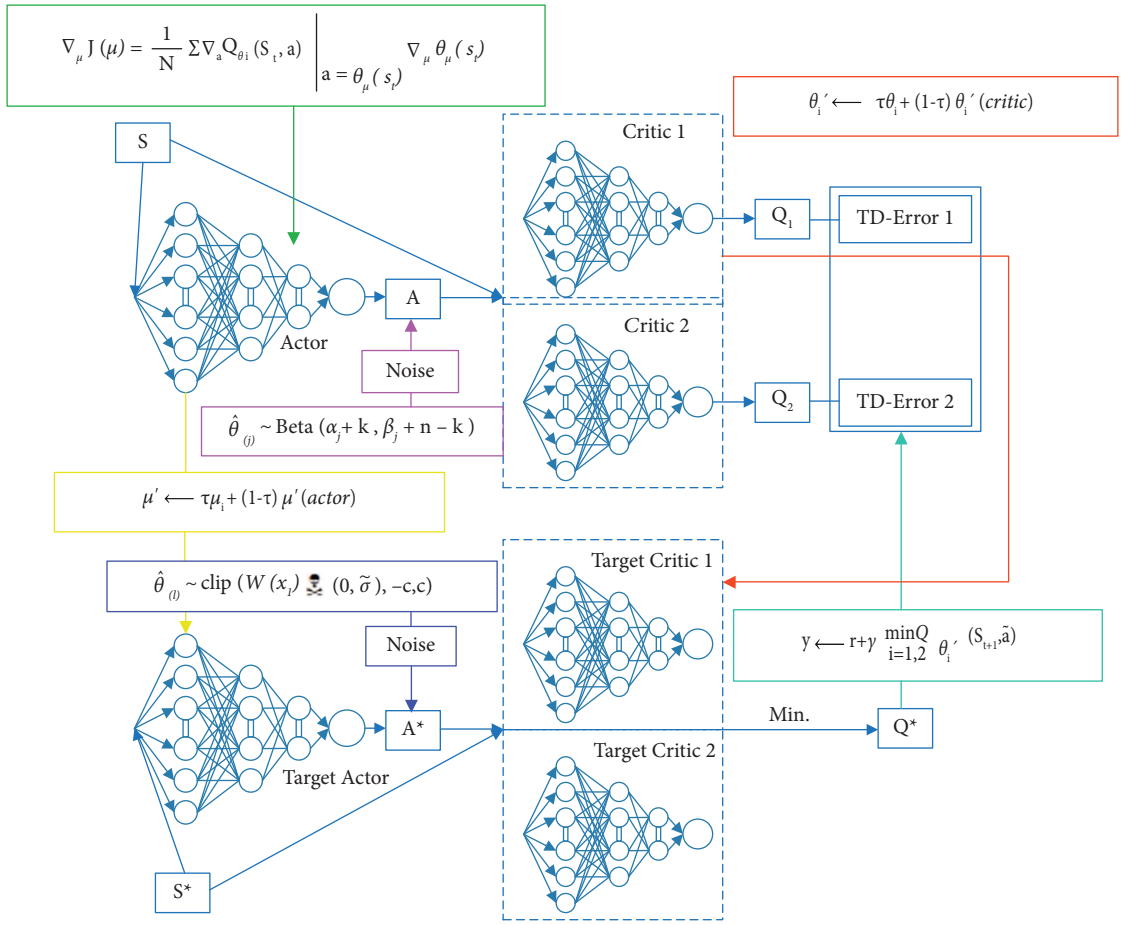

Figure 2: The architecture of the improved TD3.

prone to fall into early wrong decisions during training, making it more competitive with other sampling methods that are currently more advanced but has fewer relevant applications and research [39]. Thompson sampling is sampled based on the beta probability distribution, which is a rare family of continuous probability distributions among common distributions that take values on a finite interval [40], and it contains two positive parameters, called shape parameters, generally denoted by α and β. The probability density function of the beta distribution has the following form [41]:

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}. \tag{13}$$

Controlled by two parameters, α, and β, the beta distribution can well describe various events that occur in the interval (0, 1) and the probability of success of the event. In this article, we use Thompson sampling for exploring the noise variance applicable parameters are set on a 0.1 scale, and the two parameters α and β are the number of times selected and the number of times not selected, respectively. The beta distribution is generated for all available parameters based on the initial input values of α and β. The parameter with the best effect is selected as the current option, and then the parameters α and β are adjusted to a smaller fraction to further generate the beta distribution for the parameter, and the process is repeated iteratively. The final exploration noise variance that is most applicable to the current task is derived after Thompson sampling and used to clip on the exploration action to speed up the initial convergence of TD3.

3.3. Selection of More Stable Q Values. The phenomenon of Q-value overfitting is prevalent in determining policy gradient algorithms. A regularization strategy is introduced in TD3 for smoothing the target strategy by mimicking similar actions should have similar value (SARSA), and the training process is modified to explicitly reflect this connection:

$$y = r + \mathbb{E}_{\epsilon}\left[Q_{\theta'}\left(s', \pi_{\phi'}(s') + \epsilon\right)\right]. \tag{14}$$

TD3 makes the valuation smoother by estimating the action values by bootstrapping. In practice, by adding a small variance of noise to the target strategy and updating the action expectation in small batches on average, the valuation of the next action is made more accurate by combining all the optional valuations around the estimated initial valuation when estimating the value:

$$y = r + \gamma Q_{\theta'}\left(s', \pi_{\phi'}(s') + \epsilon\right), \epsilon \sim \mathrm{clip}(\mathcal{N}(0,\sigma), -c, c). \tag{15}$$

When updated multiple times, this computational process approximates using a small fraction of the values near the valuation point to estimate the return function at the next moment, which makes the estimated value more accurate and avoids suboptimal returns in the fetching space. But it also makes the strategy noise variance significantly affects the valuation range of the valuation point, too large strategy noise variance makes the valuation point range too large, including too many return function values, making the estimated value distorted. Too small a strategy noise variance makes the range of valuation points too small, too few return function values will make the calculation of the estimated return function inaccurate, but still easy to fall into the local optimal strategy. Therefore, to avoid the impact of too small or too large strategy noise variance, the Q-value expectation function is calculated using marginal importance sampling.

MIS is an importance sampling method used for nonpolicy assessments [42] in which the payoff function Q is assessed by reweighting the rewards for resampling subsequent actions from the data $\mathscr{D} = \{(s, a, r, s')\} \sim p(s'|s, a)d^{\mathscr{D}}(s, a)$. Here, $d^{\mathscr{D}}$ is an arbitrary distribution that does not require some specific behavioural strategy to generate. The improved payoff function Q by MIS is shown as follows:

$$Q(\pi) = \mathbb{E}_{(s,a)\sim d^{\mathscr{D}}, r(s,a)}\left[\frac{d^{\pi}(s,a)}{d^{\mathscr{D}}(s,a)} r(s,a)\right]. \tag{16}$$

The goal of the MIS method is to use the data in $\mathscr{D}$ to resample the data according to the weight $w(s, a) \approx (d^{\pi}(s, a)/d^{\mathscr{D}}(s, a))$. MIS differs from traditional importance-based sampling methods in that the ratios apply to individual transitions rather than complete trajectories, which can reduce the variance of long or infinite level problems. The redefinition of the training payoff function for TD3 using MIS allows the training strategy to estimate the action payoff at the next moment more accurately and to take smoother values in the Q-value space, enhancing the robustness of the output moderation strategy.

3.4. System Design of Improved TD3-Based CSMS. The CSMS based on the boosted TD3 constructed in this paper has three features (1) SOTA algorithm TD3 is used to construct the DRL-based CSMS. (2) Thompson sampling is used to optimize the exploration noise sampling strategy of DRL, which makes the training initial convergence of the strategy faster. (3) The calculation of the Q-return function is optimized using MIS, which makes the robustness of the strategy improved. The system flow chart of the improved TD3-based CSMS is shown in Figure 3, and its pseudocode is attached in Table 2.

4. Results and Discussions

To evaluate the global optimality of the CSMS based on the improved TD3, in the following sections, we select TD3, DDPG, PPO, TRPO, ACKTR, and SAC to construct the CSMS with the improved TD3 and compare them. First, the convergence speed of the CSMS based on the improved TD3 and other DRL-based CSMSs at the early stage of training is compared. Secondly, the stability of the CSMS after improving the Q-value return function using MIS is evaluated. Finally, the impact of the improved TD3-based CSMS and other DRL-based CSMSs on operator returns is compared.

4.1. Simulation Environment Setup. The structure of the operation area environment faced by the regulation policies constructed based on TD3, DDPG, PPO, TRPO, ACKTR, and SAC with the DRL of the modified TD3 is shown in Figure 1. In this operation regulation area, multiple charging stations are managed by the policies generated by one operation dispatch center. All DRL-based CSMSs need to be trained with the goal of maximizing operational returns by considering load demand, grid tariff, and PV generation. The main constraint of all EVCS comes from the grid-controlled real-time tariff, and the real-time tariff information comes from ComEd [43] at 19th February 2022, and the real-time tariff curve is shown in Figure 4.

The load demand at each individual EV charging station in the region varies slightly from day to day, and the training of the DRL-based CSMS is regulated by the next instantaneous predicted load value, so accurate ultrashort-term EV charging load demand prediction is critical. The ultrashort-term prediction of EV charging load in a single EV charging station has been done in previous work [26], and Figure 5 shows the actual EV charging load values compared with the predicted values, which are entered as variables in the training process of the DRL-based CSMS strategy.

Each individual EV charging station in the region has a photovoltaic power generation system that is connected to the charging station's system, and this power is supplied directly to EV users or fed into energy storage as needed. Photovoltaic power generation is highly influenced by the climate, so the power generated by PV has a large fluctuation. The 24 hour PV power generation in real time is shown in Figure 6.

4.2. Initial Convergence Rate of Improved TD3-Based CSMS. The initial convergence speed is an important metric to evaluate the performance of a regulation strategy. In practical regulation scenarios, a strategy that can reach convergence faster can save a lot of training costs and also support faster regulation scales. As described in Section 3.1.2, TD3 has three main improvements, among which, the exploration noise variance is a unique hyperparameter of TD3. Different noise variances control the range of actions performed by the actor by generating different noises, which allows the intelligence to explore more ranges of action values and approach the optimal boundary actions faster, making the regulation strategy reach convergence faster. We use Thompson sampling to select the most suitable exploration noise variance, then load the generated exploration noise on the actor for spatial action exploration, and compare the initial convergence speed of the improved TD3 with TD3, DDPG, PPO, TRPO, ACKTR, and SAC, and the results are shown in Figure 7.

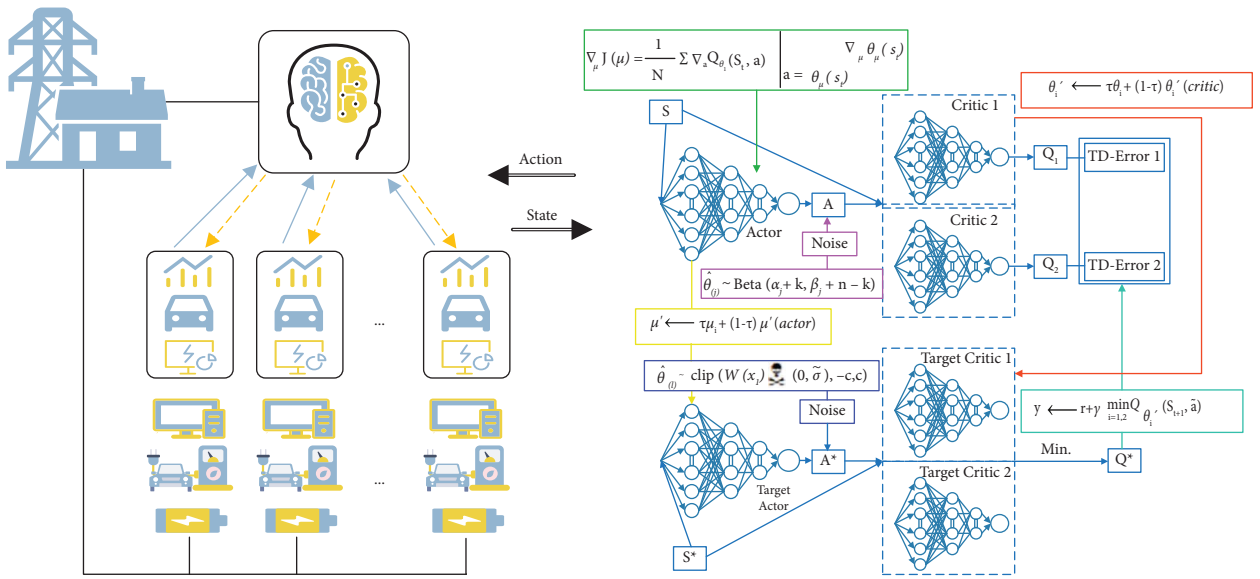

Figure 3: Management framework for CSMS based on the improved TD3.

Table 2: Pseudocode of proposed Improved TD3 algorithm.

Pseudocode of improved TD3 algorithm for sales price and energy scheduling	
	Initialization critic network with Q_{θ_1}, Q_{θ_2}, and actor network θ_μ with random parameters θ_1, θ_2, μ
1	Initialization target networks $\theta_1' \leftarrow \theta_1$, $\theta_2' \leftarrow \theta_2$, $\mu' \leftarrow \mu$
	Initialize replay buffer $\mathscr{B}$ and beta parameter (α_j, β_j)
2	for $t_1 = 1{:}T$ (time step of exploring cycle) do
3	Select a random noise $\widehat{\theta}_{(j)} \sim \text{Beta}(\alpha_j + k, \beta_j + n - k)$ for action exploration
4	get initial states: $E_t^{\text{EV}}, E_t^{\text{PV}}, E_t^{\text{EV,uf}}, SOE_t, P_t^{\text{grid}}$
5	for $t_2 = 1{:}T$ (time step of training cycle) do
6	Select action with exploration noise $a \sim \theta_\mu(s) + \widehat{\theta}_{(j)}$
7	Execute action a observe rewards r_t and new states s_{t+1}
8	Store transition (s_t, a, r_t, s_{t+1}) in $\mathscr{B}$
9	Sample mini-batch of N transitions from $\mathscr{B}$
10	$\tilde{a} \sim \theta_\mu(s) + \widehat{\theta}_{(l)}$, $\widehat{\theta}_{(l)} \sim \text{clip}(W(x_l)\mathscr{N}(0, \tilde{\sigma}), -c, c)$
11	$y \leftarrow r + \gamma \min_{i=1,2} Q_{\theta_i'}(s_{t+1}, \tilde{a})$
12	Update critics $\theta_i \leftarrow \arg\min_{\theta_i} N^{-1} \sum (y - Q_{\theta_i}(s_t, a))^2$
13	If t mod d then
14	Update the actor policy by the deterministic policy gradient:
15	$\nabla_\mu J(\mu) = (1/N) \sum \nabla_a Q_{\theta_1}(s_t, a)|_{a=\theta_\mu(s_t)} \nabla_\mu \theta_\mu(s_t)$
16	Update target networks: $\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$(critic)
17	$\mu' \leftarrow \tau\mu_i + (1-\tau)\mu'$ (actor)
18	end if
19	end for

It can be observed from the initial convergence speed graph that DDPG, ACKTR, TRPO, and PPO clearly pull apart from TD3 and SAC by a large margin when the same training objective is to maximize operational returns, with the CSMS composed of the modified TD3 having the fastest convergence speed among all CSMS because it finds the most suitable exploration noise variance earlier and has a larger and fuller exploration range for the action space in the early training.

4.3. Robustness of the Improved TD3-Based CSMS. For DRL strategies, not only the convergence speed of the strategy is a very important metric, but also the mean episode length of the strategy during the iterative process. The smaller the mean episode length, the smaller the combination of episodes required for the strategy to converge near the convergence limit, and the faster the agent can find the best strategy, thus increasing the payoff. In the scenario of practical management strategy application, the smaller the mean episode length, the higher the robustness of the strategy, and the larger the payoff that the strategy can achieve.

From Figure 8, we can observe that the mean episode length of each DRL strategy decreases significantly during the initial training process. The CSMS constructed by DDPG, SAC, and PPO has a large fluctuation of the mean episode length during the iterative process, and cannot

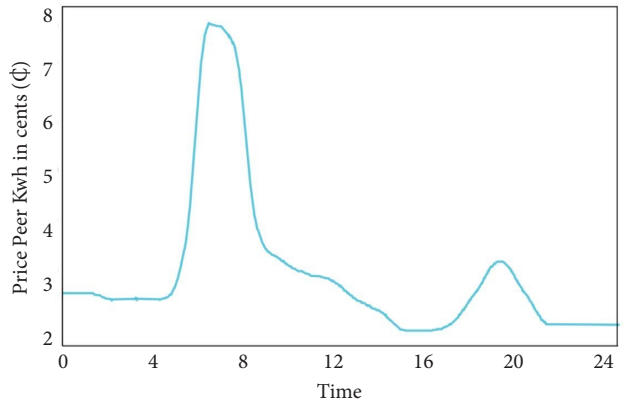

Figure 4: Hourly electricity prices of San ComEd.

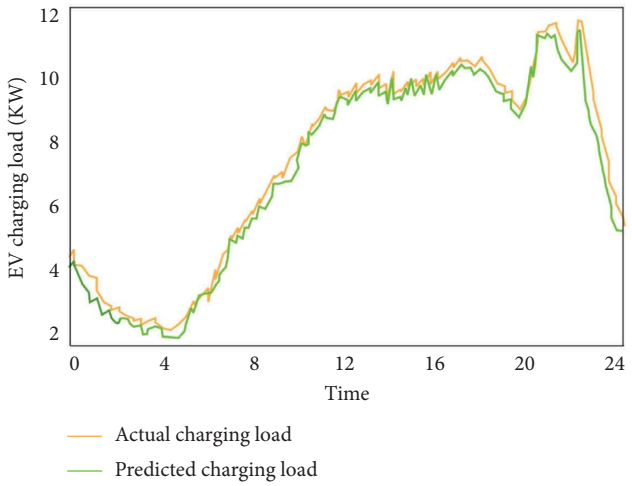

Figure 5: Actual vs. predicted values of hourly EV charging load demand.

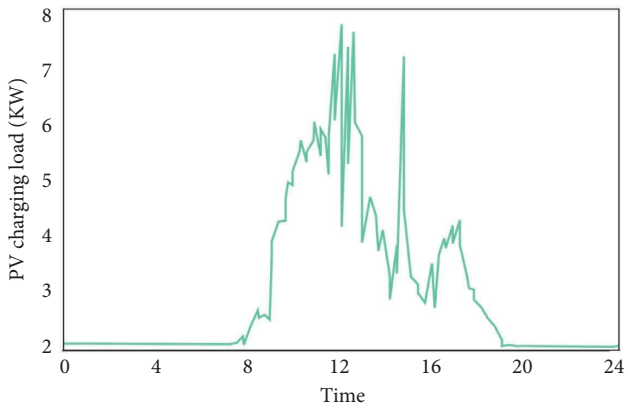

Figure 6: The predicted value of hourly PV generation.

guarantee good stability. The CSMS constructed by ACKTR has an obvious convergence trend, and the mean episode length decreases continuously with the iterative process, but the global convergence speed of this strategy is slow. The global convergence speed and the final convergence

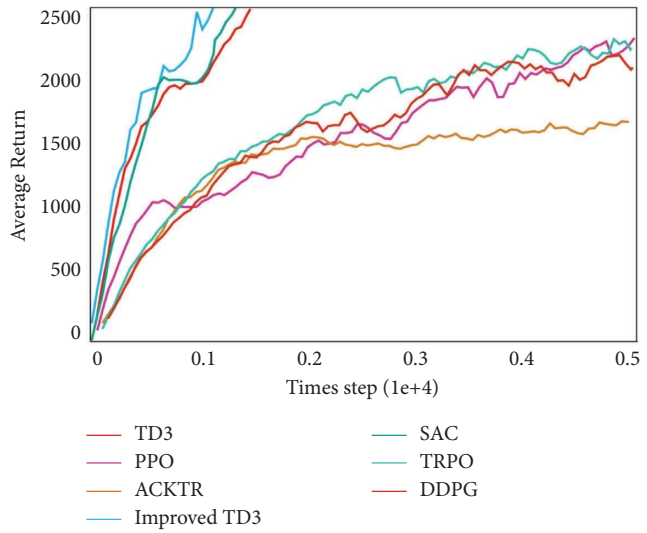

Figure 7: Comparison of initial convergence speed using seven different DRL methods.

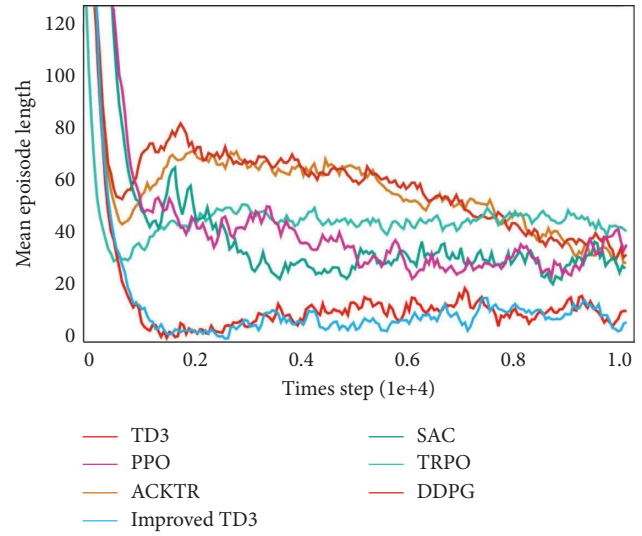

Figure 8: Comparison of robustness using seven different DRL methods.

limit of TD3 and the improved TD3 both show obvious advantages, whereas the CSMS composed based on the improved TD3 has a smaller mean episode length and higher robustness.

4.4. Operational Returns of the Improved TD3-Based CSMS. CSMS is a typical continuous type regulation task with multiple input states, and the addition of multiple input states inevitably greatly affects the convergence performance of DRL-based CSMS. In the actual operation of CSMS, the most important goal is to maximize the operational return, so it is extremely important for all DRL-based CSMS to observe its average return per training as the number of iterations increases until the final convergence.

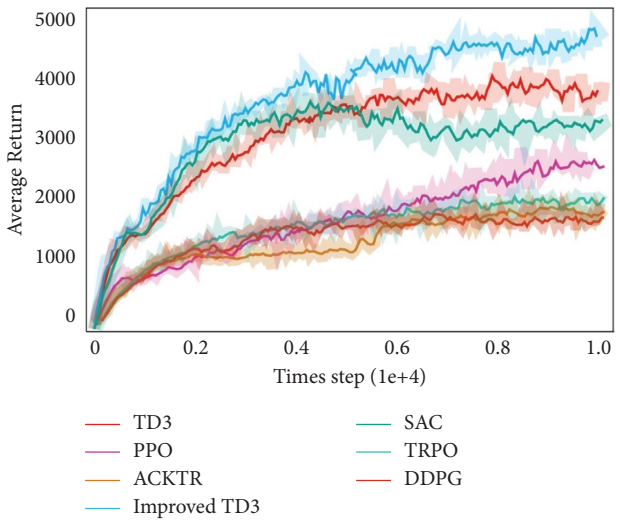

Figure 9: Comparison of average returns using seven different DRL methods.

It can be observed from Figure 9 that although the average returns of all DRL-based CSMS continue to increase with the number of iterations during the training process, the initial convergence speed and the average returns of the final convergence of PPO, TRPO, ACKTR, and DDPG are significantly lower than those of SAC, TD3 and the improved TD3. The CSMS based on the improved TD3 outperforms the CSMS based on other DRL algorithms in terms of both initial convergence speed and returns to operators. The CSMS based on the improved TD3 also outperforms the CSMS based on other DRL algorithms in terms of stability.

5. Conclusions

In this study, a CSMS based on an improved deterministic policy gradient DRL algorithm TD3 composition for managing multiple charging stations in an operation region is proposed to make the maximum return for regional charging station operators. First, a detailed model is built for the regional operation scenario constructed by multiple charging stations in terms of both energy flow and operation cost, and the CSMS is constructed with the objective function of maximizing the operation return after considering the EV load demand, PV generation, grid tariff, and operation cost, etc. To accelerate the initial convergence speed of the management strategy during training, Thompson sampling is used to improve the selection strategy of the exploration noise variance of the exploration noise attached by the agent to the exploration action. After that, considering the need for stability of the management strategy in the actual operating environment, the calculation of the Q-value return function is improved using MIS to make the management strategy more stable in the convergence phase. Compared with other CSMS based on DRL algorithms, the improved TD3 shows a faster initial convergence speed and higher global stability during training. Meanwhile, the CSMS based on the improved TD3 has the maximum payoff.

For the return problem of management strategies for charging station operators, the strategy constructed in this paper considers a small area for application. Therefore, in order to meet the management strategy requirements of charging station operators on a larger scale, the structure of the large scale management strategy can be further refined in the subsequent research to distribute the CSMS over a large scale area to meet the management requirements on a larger spatial scale.

Symbols and Abbreviations

E_t^{EV}: The sum of the charging load predictions for all EVCSs
$E_{i,t}^{\mathrm{EV}}$: The predicted value of energy demands from EV users of EVCS i at time t
E_t^{PV}: The total forecasted electricity production of PV systems with EVCSs
$E_{i,t}^{\mathrm{PV}}$: The predicted value of the energy of PV system connected to EVCS i at time t
E_t^{Fixed}: The total fixed consumption load for all EVCSs to maintain operation
$E_{i,t}^{\mathrm{Fixed}}$: The fixed load electricity consumption of charging station i at moment t
$E_t^{\mathrm{EV,uf}}$: The charging load unfulfilled by the charging station at time t
$E_{i,t}^{\mathrm{EV,uf}}$: The charging load unfulfilled by charging station i at moment t
$E_t^{\mathrm{EV,total}}$: The predicted and unsatisfied charging loads at all charging stations at moment t
SOE_t: The sum of the stored power of EVCSs at moment t
$\mathrm{SOE}_{i,t}$: The amount of power stored at distributed charging station i at moment t
η: Efficiency during charging and discharging of EVCSs
$E_{\mathrm{ec},t}$: The charge of the energy storage system at moment t after considering the cycle efficiency
$E_{\mathrm{edc},t}$: The discharge of the energy storage system at moment t after considering the cycle efficiency
$\mathrm{SOC}_{i,t}$: Power stored in the charging station i at time t
P_t^{grid}: Dynamic electricity price of the grid at time t
P_t^{PV}: The price per unit of electricity saved by the PV system at moment t
P_t^{En}: The savings in environmental management costs through the use of clean energy
P_{poll}: Treatment cost per unit of electricity generated to be treated
C_{RE}: The replacement cost of energy storage batteries
α: The percentage of annual decrease in battery cost
K_{R}: The number of battery replacements
i_c: The discount on the cost of purchasing the battery
C_{SE}: The unit cost of the battery
N: The battery energy storage plant operating cycle
L: The energy storage battery replacement cycle
C_{IC}: The fixed investment cost of the energy storage plant
$C_{\mathrm{IS},\,j}$: The overhaul cost in year j
dr: The discount rate
C_{end}: The disposal cost of the energy storage system at the end of life

C_{FCSNPV}: The whole-life net present value of EVCS
C_{FCSNAV}: The whole-life net annual value of EVCS
$P_{i,t}^{\text{sell}}$: Selling price factor for EVCS i at time t
C_{RETURN}: The operating return of EVCS power sales
s': State space variables
$\theta_{\mu 1}(s)$: Storage of action variables
y: The value of the next moment
M: The number of batches split
$Q_{\theta i}$: Set of critic agents for EVCS
$\tilde{a}$: The noise used for smoothing
ϵ: The noise added to the target strategy
Q_{θ}: Initial value of the critical network
$d^{\mathscr{D}}$: Distribution range of parameters
θ_{μ}: Set of actions agents for EVCS
$\widehat{\theta}_{(j)}$: Random noise in his distribution
$\mathscr{B}$: The replay buffer of the training network
α_j, β_j: Shape parameters of the Beta distribution
$\widehat{\theta}_{(j)}$: Random noise of action exploration
$R_{n,t}^{\text{total}}$: The total reward for EVCS agent n at the time t.

Data Availability

The data supporting this study's findings are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was sponsored in part by two National Natural Science Foundation of China (Grant nos. 52167014 and 52077139), in part by the Science and Technology Commission of Shanghai Municipality (Grant no. 21DZ2204800), and in part by the Science and Technology Project of State Grid Shanghai Municipal Electric Power Company of China (Grant no. 52094022000G), and in part by the Key Laboratory of Control of Power Transmission and Conversion (SJTU), Ministry of Education (Grant no. 2022AA04).

References

[1] S. Saharan, S. Bawa, and N. Kumar, "Dynamic pricing techniques for Intelligent Transportation System in smart cities: a systematic review," *Computer Communications*, vol. 150, pp. 603–625, 2020.

[2] Y. Wu, A. Ravey, D. Chrenko, and A. Miraoui, "Demand side energy management of EV charging stations by approximate dynamic programming," *Energy Conversion and Management*, vol. 196, pp. 878–890, 2019.

[3] Y. Li, Z. Ni, T. Zhao et al., "Supply function game based energy management between electric vehicle charging stations and electricity distribution system considering quality of service," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, pp. 5932–5943, 2020.

[4] MK. Daryabari, R. Keypour, and H. Golmohamadi, "Stochastic energy management of responsive plug-in electric vehicles characterizing parking lot aggregators," *Applied Energy*, vol. 279, Article ID 115751, 2020.

[5] A. Akbari-Dibavar, V. Sohrabi Tabar, S. Ghassem Zadeh, and R. Nourollahi, "Two-stage robust energy management of a hybrid charging station integrated with the photovoltaic system," *International Journal of Hydrogen Energy*, vol. 46, no. 24, pp. 12701–12714, 2021.

[6] M. Khoshjahan, M. Soleimani, and M. Kezunovic, "Optimal participation of PEV charging stations integrated with smart buildings in the wholesale energy and reserve markets," in *Proceedings of the 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference*, pp. 1–5, New Orleans, LA, USA, February 2020.

[7] M. Alqahtani and M. Hu, "Dynamic energy scheduling and routing of multiple electric vehicles using deep reinforcement learning," *Energy*, vol. 244, Article ID 122626, 2022.

[8] N. Sadeghianpourhamami, J. Deleu, and C. Develder, "Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 203–214, 2020.

[9] S. Wang, S. Bi, and YA. Zhang, "Reinforcement learning for real-time pricing and scheduling control in EV charging stations," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 849–859, 2021.

[10] V. Moghaddam, A. Yazdani, H. Wang, D. Parlevliet, and F. Shahnia, "An online reinforcement learning approach for dynamic pricing of electric vehicle charging stations," *IEEE Access*, vol. 8, pp. 130305–130313, 2020.

[11] T. Ding, Z. Zeng, J. Bai, B. Qin, Y. Yang, and M. Shahidehpour, "Optimal electric vehicle charging strategy with Markov decision process and reinforcement learning technique," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, pp. 5811–5823, 2020.

[12] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[13] X. Hu, T. Liu, X. Qi, and M. Barth, "Reinforcement learning for hybrid and plug-in hybrid electric vehicle energy management: recent advances and prospects," *IEEE Industrial Electronics Magazine*, vol. 13, no. 3, pp. 16–25, 2019.

[14] M. Shin, D.-H. Choi, and J. Kim, "Cooperative management for PV/ESS-enabled electric vehicle charging stations: a multiagent deep reinforcement learning approach," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3493–3503, 2020.

[15] F. Kiaee, "Integration of electric vehicles in smart grid using deep reinforcement learning," in *Proceedings of the 2020 11th International Conference on Information and Knowledge Technology*, pp. 40–44, Osaka Japan, January 2020.

[16] Y. Gao, J. Yang, M. Yang, and Z. Li, "Deep reinforcement learning based optimal schedule for a battery swapping station considering uncertainties," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, pp. 5775–5784, 2020.

[17] S. Li, W. Hu, D. Cao et al., "Electric vehicle charging management based on deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 3, pp. 719–730, 2022.

[18] F. Zhang, Q. Yang, and D. An, "CDDPG: a deep-reinforcement-learning-based approach for electric vehicle charging control," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3075–3087, 2021.

[19] W. Li, H. Cui, T. Nemeth et al., "Cloud-based health-conscious energy management of hybrid battery systems in electric vehicles with deep reinforcement learning," *Applied Energy*, vol. 293, Article ID 116977, 2021.

[20] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of the International conference on machine learning*, pp. 1587–1596, Stockholm, Sweden, July 2018.

[21] A. Abdalrahman and W. Zhuang, "Dynamic pricing for differentiated PEV charging services using deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 1415–1427, 2022.

[22] S. Yao, J. Gu, H. Zhang, P. Wang, X. Liu, and T. Zhao, "Resilient load restoration in microgrids considering mobile energy storage fleets: a deep reinforcement learning approach," in *Proceedings of the 2020 IEEE Power & Energy Society General Meeting*, pp. 1–5, Denver, CO, USA, July 2020.

[23] Y. Wang, Z. Gao, J. Zhang et al., "Trajectory design for UAV-based internet-of-things data collection: a deep reinforcement learning approach," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3899–3912, 2022.

[24] J. Zhou, S. Xue, Y. Xue, Y. Liao, J. Liu, and W. Zhao, "A novel energy management strategy of hybrid electric vehicle via an improved TD3 deep reinforcement learning," *Energy*, vol. 224, Article ID 120118, 2021.

[25] K. Deng, Y. Liu, D. Hai et al., "Deep reinforcement learning based energy management strategy of fuel cell hybrid railway vehicles considering fuel cell aging," *Energy Conversion and Management*, vol. 251, Article ID 115030, 2022.

[26] H. Li, J. Zhu, X. Fu, C. Fang, D. Liang, and Y. Zhou, "Ultra-short-term load forecasting of electric vehicle charging stations based on ensemble learning," *Journal of Shanghai Jiaotong University*, vol. 56, no. 8, pp. 1004–1013, 2022.

[27] M. M. Elbaz and A. Eldesouky, "Determination of upper and lower limits of concession period for PPP projects using probabilistic NPV," *Journal of Engineering Research*, vol. 3, no. 9, pp. 31–39, 2019.

[28] T. P. Lillicrap, J. J. Hunt, A. Pritzel et al., "Continuous control with deep reinforcement learning," 2015.

[29] G. Matheron, N. Perrin, and O. Sigaud, "The problem with DDPG: understanding failures in deterministic environments with sparse rewards," 2019, https://arxiv.org/abs/1911.11679.

[30] C. Tallec, L. Blier, and Y. Ollivier, "Making deep q-learning methods robust to time discretization," *Proceedings of the 36th International Conference on Machine Learning, PMLR*, vol. 97, pp. 6096–6104, 2019.

[31] C. J. C. H. Watkins, *Learning Form Delayed Rewards*, University of Cambridge, Cambridge CB2 1TN, 1989.

[32] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," *Proceedings of the 31st International Conference on Machine Learning, PMLR*, vol. 32, pp. 3873951, Bejing, China, June 2014.

[33] H. Dong, Z. Ding, and S. Zhang, *Deep Reinforcement Learning*, Springer Singapore, Singapore, 2020.

[34] R. Islam, P. Henderson, M. Gomrokchi, and D. Precup, "Reproducibility of benchmarked deep reinforcement learning tasks for continuous control," 2017, https://arxiv.org/abs/1708.04133.

[35] R. Hoque, D. Seita, A. Balakrishna et al., "Visuospatial foresight for multi-step, multi-task fabric manipulation," 2020, https://arxiv.org/abs/2003.09044.

[36] T. Tiong, I. Saad, K. T. K. Teo, and H. bin Lago, "Deep reinforcement learning with robust deep deterministic policy gradient," in *Proceedings of the 2020 2nd International Conference on Electrical, Control and Instrumentation Engineering*, Kuala Lumpur, Malaysia, November 2020.

[37] S. Dankwa and W. Zheng, "Twin-delayed ddpg: a deep reinforcement learning technique to model a continuous movement of an intelligent robot agent," in *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*, vol. 66, pp. 1–5, Vancouver BC Canada, August 2019.

[38] H. Hasselt, "Double Q-learning," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[39] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," *Proceedings of the 25th Annual Conference on Learning Theory, PMLR*, vol. 23, no. 39, pp. 1–26, 2012.

[40] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan, "Statistical methods for eliciting probability distributions," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 680–701, 2005.

[41] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 2, Wiley, Hoboken, NJ, USA, 1995.

[42] T. Xie, Y. Ma, and Y. Wang, "Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[43] C. E. ComEd, 2022, https://hourlypricing.comed.com/live-prices/?date=20220218.