Deep Reinforcement Learning Based Charging Scheduling for Household Electric Vehicles in Active Distribution Network

Taoyi Qi, Chengjin Ye, Yuming Zhao, Lingyang Li, and Yi Ding

Abstract—**With the booming of electric vehicles (EVs) across the world, their increasing charging demands pose challenges to urban distribution networks. Particularly, due to the further implementation of time-of-use prices, the charging behaviors of household EVs are concentrated on low-cost periods, thus generating new load peaks and affecting the secure operation of the medium- and low-voltage grids. This problem is particularly acute in many old communities with relatively poor electricity infrastructure. In this paper, a novel two-stage charging scheduling scheme based on deep reinforcement learning is proposed to improve the power quality and achieve optimal charging scheduling of household EVs simultaneously in active distribution network (ADN) during valley period. In the first stage, the optimal charging profiles of charging stations are determined by solving the optimal power flow with the objective of eliminating peak-valley load differences. In the second stage, an intelligent agent based on proximal policy optimization algorithm is developed to dispatch the household EVs sequentially within the low-cost period considering their discrete nature of arrival. Through powerful approximation of neural network, the challenge of imperfect knowledge is tackled effectively during the charging scheduling process. Finally, numerical results demonstrate that the proposed scheme exhibits great improvement in relieving peak-valley differences as well as improving voltage quality in the ADN.**

Index Terms—**Household electric vehicles, deep reinforcement learning, proximal policy optimization, charging scheduling, active distribution network, time-of-use prices.**

I. Introduction

In recent years, electric vehicles (EVs) are widely used to reduce air pollution and emissions of greenhouse gases with the appeal for sustainable development goals [1]. According to the estimation of International Energy Agency, EV sales in China have reached 3.3 million in 2021 [2], whose soring charging demands pose new challenges to the reliable operation of the distribution network, i.e., dramatical peak-valley load differences [3], power congestions [4], and undervoltage [5]. The active distribution network (ADN) enables a mass of flexible loads to deliver various regulation services to solve the above problems, which puts forward efficient and economic solutions for power systems [6]. Conventional flexible loads with promising regulation capacity, such as air conditioners [7], have been fully investigated into demand response for supporting system balancing. However, limited response capacity and duration time of residential loads restrict their further implementations on long-time-scale dispatch considering users’ comforts.

EVs are regarded as ideal alternatives to provide various regulation services in the ADN [8]. The significant regulation potential of EVs has attracted lots of research interests, and their prominent contributions to peak shaving, accommodation of renewable energy resources (RESs), and voltage regulations have been validated in [9], [10], and [11], respectively. Charging scheduling is fundamental for the ADN to utilize the flexibility of EVs due to the elasticity of departure time. A coordinated charging strategy for EVs in the distribution network is presented in [12] to manage power congestion. An intelligent charging scheduling algorithm is proposed in [13] to choose the suitable charging station (CS) and charging period with the goal of minimizing charging costs.

Private charging is currently the dominant charging mode for household EVs in many countries [2]. However, lower charging power and utilization rate as well as wide-area distributions make it uneconomic to retrofit household charging piles to achieve flexible regulation used in commercial direct current (DC) piles, such as the on/off charging control strategy and continuous charging power adjustment mentioned in [14] and [15]. As a result, time-of-use (TOU) prices are implemented for household EVs to transfer charging demands from peak period to valley period in some areas, e.g., Zhejiang Province of China. Household EVs possess competitive charging flexibility without sharing charging piles with others [16]. Considering the fixed electricity price during the valley period, it is tricky for TOU prices to dispatch EVs adequately with regard to the realistic operating characteristics of the ADN, except for transferring charging loads from

Manuscript received: July 27, 2022; revised: December 28, 2022; accepted: January 20, 2023. Date of CrossCheck: January 20, 2023. Date of online publication: February 24, 2023.

This work was supported by the National Key R&D Program of China (No. 2021ZD0112700) and the Key Science and Technology Project of China Southern Power Grid Corporation (No. 090000k52210134).

T. Qi, C. Ye (corresponding author), L. Li, and Y. Ding are with the College of Electrical Engineering, Zhejiang University, Hangzhou, China (e-mail: eeqty@zju.edu.cn; yechenjing@zju.edu.cn; lilingyang@zju.edu.cn; yiding@zju.edu.cn).

Y. Zhao is with Shenzhen Power Supply Bureau Co., Ltd., Shenzhen, China (e-mail: zhaoym97@sina.com).

DOI: 10.35833/MPCE.2022.000456

peak period to valley period. The fixed valley price means that the charging costs are minimum as long as the whole charging process is finished within the valley period. Consequently, owners instinctively decide to start charging at the beginning of valley period for convenience, which leads to new congestions in the ADN [17]. The intensive charging demands are likely to result in equipment failures and severely threaten the secure operation of the ADN, accounting for the large-scale integration of EVs. Therefore, developing a promising charging scheduling method for household EVs is of great importance to tackle these challenges.

Apart from the above regulation obstacles under TOU prices, previous approaches to solve the charging scheduling problem are not suitable for household EVs with private charging piles, accounting for their sequential arrivals and uncertain charging demands. For instance, the offline and online scheduling algorithms are proposed in [18] for EVs to save the charging cost, which is formulated as a mixed integer programming (MIP) problem assuming full knowledge of the charging demands. The bi-level optimal dispatching model proposed in [19] is also solved by converting it into an MIP problem. These studies assume that all charging demands are collected before optimization so as to convert them to solvable MIP problems. However, it is very difficult or even impossible to acquire all charging information in advance. In reality, the EV arrives sequentially and the charging demands can only be obtained precisely after arrival.

Under these circumstances, the charging process of the household EV can only be regarded as an uninterruptable process and the charging demands cannot be obtained in advance. The charging scheduling problem of household EVs can be formulated as a Markov decision process (MDP) [20], which aims to dispatch EVs sequentially with finite information, and achieve the global optimum for all EVs in the end. Therefore, how to determine the specific charging start time of the EV when arriving is one of the key priorities. On the basis of the charging reservation function, household EVs can be adopted appropriately in the charging scheduling of the ADN without extra equipment investments.

With the rapid development of deep learning (DL) and reinforcement learning (RL), deep reinforcement learning (DRL), which combines both advantages of DL and RL, is proposed to overcome the dimensional curse and solve the MDP problem with continuous action spaces [21]. Based on the powerful function approximation of neural networks and big data technology, DRL is emerged as an interesting alternative to address the sequential charging scheduling problem without full knowledge of charging demands [22]. First of all, the decision for the current EV only depends on the real-time environment states, i. e., arrival time, charging power, charging duration, and departure time, and it is notable for a DRL agent to address such a problem due to the sequential feature. Moreover, through interacting repeatedly with the dynamic environment, the agent can learn from the experience and investigate an excellent control policy in the absence of models, which is more appliable in uncertain environments.

In the field of charging scheduling problems, DRL has been implemented in various optimizations. Reference [23] proposes a novel DRL method based on the prioritized deep learning deterministic policy gradient method, so as to solve the bi-level optimization of the examined EV pricing problem. For the EV CS, an energy management based on DRL is proposed in [24] to tickle varying input data and reduce the cumulated operation costs. However, the above literatures mostly focus on minimizing the operation costs of CSs by dispatching EVs, while the contributions to the improvement of power quality in the ADN are not fully accounted for. Under TOU prices, EVs can be further dispatched to relieve the congestion and shorten the peak-valley differences without extra charging costs.

To address the above problems and take full use of substantial household EVs during valley period under TOU prices, this paper proposes a two-stage charging scheduling scheme for household EVs in the ADN. In the first stage, to relieve the power congestions and shorten the peak-valley differences, the optimal power flow (OPF) of the ADN is solved to determine the optimal charging profiles of CSs during the valley period. In the second stage, DRL based on proximal policy optimization (PPO) algorithm is employed to dispatch the household EVs sequentially within the low-cost period according to the optimal charging profiles. PPO algorithm was proposed by OpenAI in 2017 [25], which combines the advantages of trust region policy optimization and advantage actors-critic, to prevent the performance collapse caused by a large update of the policy. Besides, most decisions are finished by the distributed agents in the proposed scheme with lower communication requirements and computational burden, which makes it appliable easily in ADNs with numerous EVs.

The main contributions of this paper are as follows.

1) A two-stage charging scheduling scheme for household EVs is proposed to improve the power quality of the ADN and achieve the optimal charging scheduling of EVs simultaneously during the valley period, which consists of the OPF of the ADN and charging dispatch of EVs. On this basis, the contributions of household EVs to power congestion management and peak-valley difference elimination are further exploited.

2) The realistic characteristics of household EVs are taken into consideration, including the limited controllability and uncertain charging demands. The charging process of EV is regarded as an uninterruptable procedure with constant power, and the charging scheduling process is modelled as a sequential MDP problem, thereby the owner can make a charging reservation to achieve charging scheduling without extra equipment investments.

3) The intelligent DRL agent based on the PPO algorithm is developed to schedule the charging process of EVs. Through the remarkable approximation function of the neural network, the agent can accumulate rich experience when interacting with various environments repeatedly to break the limitations on imperfect information. Hence, numerous household EVs are dispatched effectively to formulate the optimal charging profile even when lacking full knowledge of charging demands in advance.

The remainder of this paper is organized as follows. Section II establishes the two-stage charging scheduling scheme of household EVs. Section III introduces the MDP model of EV charging scheduling and the intelligent DRL agent based on the PPO algorithm. Case studies are conducted in Section IV using the real-world data of residential and EV loads, which proves the effectiveness of the proposed scheme. Section V concludes the remarks of this paper.

II. Two-stage Charging Scheduling Scheme of Household EVs

In this section, an overview of the charging scheduling scheme is introduced first to illustrate the coordination between the problems in the two stages. Then, the first-stage problem which considers the mutual impacts of different nodes is put forward to determine the optimal operation of the ADN with household EVs. On the basis of the optimal charging profiles provided by the first-stage problem, the detailed sequential decision problem in the second stage is formulated to describe the charging scheduling process.

A. Overview of Charging Scheduling Scheme

As important flexible loads of the ADN, household EVs are not fully exploited for further regulation potential under TOU prices. Generally, most charging durations of household EVs are much shorter than their sojourn time [26]. Substantial charging demands of EVs concentrate on the prophase of the valley period lacking effective guidance, which results in extra power congestions and wastes the regulation potential of EVs to a large extent.

At the same time, the ADN is suffering from power quality issues including dramatic peak-valley differences, power congestions, and voltage limit violations. Consequently, the managers of the ADN, i. e., distribution network operator (DSO) and energy supplier, are motivated to further dispatch household EVs to improve the power quality under TOU prices without extra equipment investments and charging costs, even earning profits through delivering ancillary services for power systems. Apart from the DSOs, estates or community administrators are also encouraged to implement such a charging scheduling, so as to satisfy increasing charging demands accounting for the limited carrying capacities of ADNs.

The schematic diagram of two-stage charging scheduling scheme of household EVs is shown in Fig. 1, assuming that the ADN at the residential side is operated by the DSO and consists of several residential loads and EV charging loads. Considering the relatively centralized installations of private charging piles, e. g., underground parking spaces, nearby charging piles are aggregated as a CS and managed by the aggregator. Assume that only EVs can provide flexible regulation service while other residential loads are regarded as fixed loads. To relieve the power congestion caused by intensive charging demands and transfer them to appropriate time periods, a two-stage problem is formulated.

In the first stage, determining the optimal charging profiles of CSs is the key point. Because of the various operating characteristics of different ADNs, it is of great importance for the DSO to choose favorable optimization objectives at first. In this paper, charging scheduling of household EVs is employed to flatten the tie-line power to provide ancillary services to power systems. Considering the mutual impacts between different nodes, the optimal charging profiles of CSs are not appropriate to be determined simply according to their electricity sectors. Therefore, the OPF algorithm is used to solve the problem with regard to the secure and stable operation. The optimal charging power is calculated with the goal of shortening the peak-valley differences, based on historical and forecasted load data during the valley period.

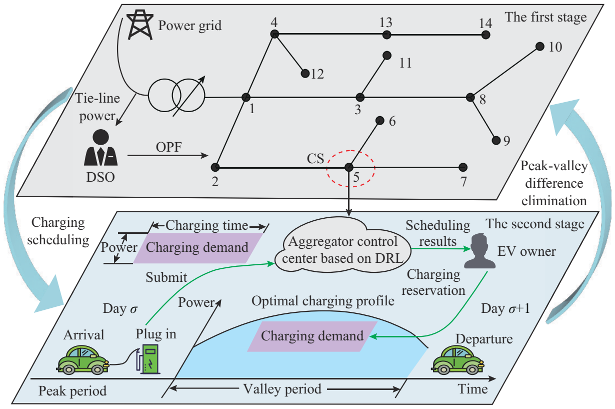

Fig. 1. Schematic diagram of two-stage charging scheduling scheme of household EVs.

In the second stage, to overcome the obstacle of limited charging information, the aggregator control center based on DRL is used to make decisions with imperfect knowledge and dispatch the charging processes of EVs in terms of the determined charging profile. Before the valley period, when the k^{th} EV arrives home, the user needs to plug in the EV and submit the charging demands to the control center, including charging power, charging duration time, and departure time. Then, according to the optimal charging profile and previously scheduled EV power, the aggregator will determine the most suitable charging time period for the k^{th} EV immediately. At last, the owner makes a charging reservation to realize the charging scheduling without extra charging costs, and even gets some incentives for the contributions to the operation of the ADN.

B. First-stage Problem: OPF Model of ADN

The first-stage problem aims to involve EVs participating in shortening peak-valley differences and managing congestions of the ADN. Considering the fact that an ADN typically features radial topology, as shown in Fig. 2, the complex power flow at each node can be described by the classic DistFlow model [27].

$$P_{i+1}=P_i-r_i(P_i^2+Q_i^2)/V_i^2-p_{i+1} \tag{1}$$

$$Q_{i,i+1}=Q_{i-1,i}-x_i(P_i^2+Q_i^2)/V_i^2-q_{i+1} \tag{2}$$

$$V_{i+1}^2=V_i^2-2(r_iP_i+x_iQ_i)+(r_i^2+x_i^2)(P_i^2+Q_i^2)/V_i^2 \tag{3}$$

$$\begin{cases} p_i=p_i^D-p_i^g \\ q_i=q_i^D-q_i^g \end{cases} \tag{4}$$

where P_i and Q_i are the active and reactive power flows from node i to node $i+1$, respectively; p_i and q_i are the active and reactive power demands at node i, respectively, which are determined by the load demands (with superscript D) and generator outputs (with superscript g); V_i is the voltage at node i; and r_i and x_i are the resistance and reactance of the branch from node i to node $i+1$, respectively.

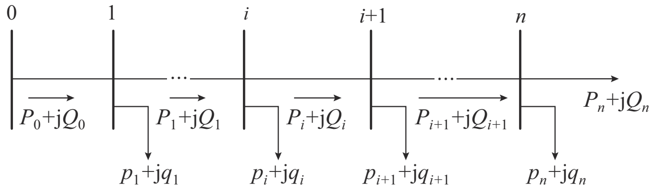

Fig. 2. Diagram of ADN with radial topology.

The DistFlow equations above are nonlinear and difficult to solve. Ignoring network losses, the DistFlow equations can be converted to linearized power flow equations as (5), which have been widely used in distribution network analysis [28].

$$\begin{cases} P_{i+1} = P_i - p_{i+1} \\ Q_{i+1} = Q_i - q_{i+1} \\ V_{i+1} = V_i - (r_i P_i + x_i Q_i)/V_1 \\ p_i = p_i^D - p_i^g \\ q_i = q_i^D - q_i^g \end{cases} \tag{5}$$

The OPF model proposed in this subsection aims to flatten the tie-line power, as well as maintain the node voltage within the acceptable range. The objective tie-line power should be determined in advance and power profiles of all nodes can be calculated using the OPF model with the goal of minimizing differences between real tie-line power and objective power. Considering the limited penetrations of EVs in the ADN at present, it is difficult to eliminate the peak-valley differences completely without abundant regulation capacity. Hence, the objective tie-line power which is related to residential consumption and the charging electricity can be computed as:

$$P_{obj}(t) = P_u(t) + \frac{E_{EV}}{E_{dev}}(P_{u,\max} - P_u(t)) \tag{6}$$

$$E_{dev} = \int_{t_s}^{t_e} (P_{u,\max} - P_u(t))\mathrm{d}t \tag{7}$$

$$E_{EV} = \int_{t_s}^{t_e} P_{EV}(t)\mathrm{d}t \tag{8}$$

where $P_{obj}(t)$ is the objective tie-line power of the ADN at time t; $P_{EV}(t)$ and $P_u(t)$ are the power of EVs and residential loads at time t, respectively; $P_{u,\max}$ is the maximum power of residential loads during the valley period; E_{EV} is the total electricity consumption of EVs; E_{dev} is the electricity deviation between the residential loads and that of the maximum power; and t_s and t_e are the start time and end time of the valley period, respectively.

Therefore, the objective function of the OPF model can be represented by:

$$\min|\Delta P| = \int_{t_s}^{t_e} |P_{sub}(t) - P_{obj}(t)|\mathrm{d}t \tag{9}$$

where $P_{sub}(t)$ is the real tie-line power at time t. The OPF problem aims to optimize the power profiles of all nodes to minimize the differences between $P_{sub}(t)$ and $P_{obj}(t)$.

Assume F and R represent the set of EV nodes and the set of residential nodes, respectively. Considering the continuous characteristic of the charging process, it is difficult to regulate the charging power of CS dramatically in a short period, thereby the ramp rate of the CS needs to be limited within λ.

$$|p_i(t+1) - p_i(t)| \leq \lambda p_i(t) \quad \forall i \in F \tag{10}$$

In addition, the constraints of the ADN mainly include the nodal voltage and feeder ampacity as shown in (11) and (12), respectively.

$$V_{i,\min} \leq V_i(t) \leq V_{i,\max} \quad \forall i \in F \cup R, t \in [t_s, t_e] \tag{11}$$

$$P_{i,\min} \leq P_i(t) \leq P_{i,\max} \quad \forall i \in F \cup R, t \in [t_s, t_e] \tag{12}$$

where $V_{i,\min}$ and $V_{i,\max}$ are the minimum and maximum nodal voltages at node i, respectively; and $P_{i,\min}$ and $P_{i,\max}$ are the minimum and maximum ampacities of the branch from node i to node $i+1$, respectively.

C. Second-stage Problem: Scheduling Model of Household EVs

After calculating the OPF of the ADN, the optimal charging profiles of CSs are determined. Then, the agent based on DRL will dispatch EVs to approach the optimal charging power.

The charging process of EVs can be divided into three parts, which are trickle charging, constant current charging, and constant voltage charging, where the constant current charging process accounts for 80% duration and has relatively constant power [29]. On the other hand, considering the actual situations where household charging piles are installed, it is difficult for charging piles to achieve continuous power regulation due to the lack of communication conditions. Therefore, the charging process of a household EV is regarded as a continuous process with constant power [30], and the charging demands of the k^{th} EV CD_k can be represented using a tuple as:

$$CD_k = (t_{arr,k}, P_{c,k}, t_{c,k}, t_{dep,k}) \tag{13}$$

where $t_{arr,k}$ is the arrival time of the k^{th} EV; $P_{c,k}$ and $t_{c,k}$ are the constant charging power and charging time duration, respectively; and $t_{dep,k}$ is the departure time, which means that the charging process needs to be finished before the departure of EVs to satisfy the owner's traveling energy requirements.

EVs arrive sequentially and the specific charging demands can only be obtained precisely when an EV is plugged in. The aggregator control center aims to transfer the charging demands to formulate a redistribution scheme of charging demands based on the objective charging power. Through the charging scheduling of EVs, not only power congestions at the prophase of valley period can be alleviated, but also the ancillary service for shortening the peak-valley differences can be delivered to power systems.

EVs can be divided into adjustable groups and non-adjustable groups. The non-adjustable EV, whose charging time duration is longer than its sojourn time, will not be regulated. The start charging time of non-adjustable EVs needs to be set as their arrival time to satisfy charging demands and there is no need to involve them in the proposed charging scheduling. Therefore, the following charging scheduling focuses on adjustable EVs. When dispatching EVs to formulate the optimal charging profile, the charging demands can be described using a rectangle as demonstrated in Fig. 3, whose length and height indicate the charging time duration and charging power, respectively. The valley period is from t_s of day σ to t_e of day $\sigma+1$. Once the k^{th} EV arrives, the charging demand CD_k is submitted to the control center. Then the control center determines the charging start time of the k^{th} EV according to its charging demands and the optimal charging profile.

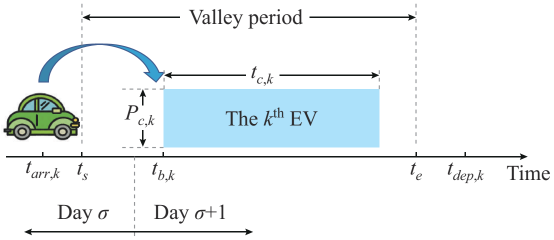

Fig. 3. Diagram of charging scheduling process.

Denoting $t_{b,k}$ as the optimized charging start time of the k^{th} EV, the real-time charging power of the k^{th} EV during the valley period can be represented as:

$$P_{EV,k}(t)=\begin{cases}0 & t_s \le t < t_{b,k}\\ P_{c,k} & t_{b,k} \le t < t_{b,k}+t_{c,k}\\ 0 & t_{b,k}+t_{c,k} \le t < t_e\end{cases} \tag{14}$$

$$t_{c,k}=\frac{C_{EV,k}(SOC_{e,k}-SOC_{s,k})}{\eta P_{c,k}} \tag{15}$$

where $C_{EV,k}$ is the rated battery capacity of the k^{th} EV; $SOC_{e,k}$ and $SOC_{s,k}$ are the expected SOC and the starting SOC of the k^{th} EV, respectively; and η is the charging efficiency and regarded as a fixed value.

These adjustable EVs have already decided to charge during the valley period with lower electricity price though they have arrived earlier, and most of them instinctively choose to start charging at the beginning of valley period t_s due to the lack of effective guidance. Therefore, the range of actionable space is set from t_s to t_e, which aims to determine their optimal charging periods. Besides, to satisfy the charging demands and save charging costs of EVs, $t_{b,k}$ is also constrained as:

$$t_s \le t_{b,k} \le \min(t_e, t_{dep,k})-t_{c,k} \tag{16}$$

After dispatching the k^{th} EV, the control center will update the scheduled charging power of the first k EVs as follow:

$$P_{sum,k}^{EV}(t)=\sum_{j=1}^{k} P_{EV,j}(t) \quad t\in[t_s,t_e] \tag{17}$$

where $P_{sum,k}^{EV}(t)$ is the scheduled power of the first k EVs at time t.

III. MDP Model of EV Charging Scheduling and Intelligent DRL Agent Based on PPO Algorithm

In this section, the MDP model of EV charging scheduling is developed at first. Then, the intelligent DRL agent based on the PPO algorithm is introduced, followed by the training workflow of the PPO algorithm.

A. MDP of EV Charging Scheduling Process

The charging scheduling of household EVs can be modelled as an MDP due to the discrete arrivals of EVs and the randomness of charging demands, which can be appropriately solved using the DRL algorithm. An MDP can be represented as a tuple $(\boldsymbol{S},\boldsymbol{A},\boldsymbol{R},\boldsymbol{T})$ [31], where $\boldsymbol{S}$ is the state space; $\boldsymbol{A}$ is the action space; $\boldsymbol{R}$ is the reward function; and $\boldsymbol{T}$ is the state transition function, which is determined by (14) and (17). The specific illustrations of the MDP are follows.

1) The state space observed by the agent is represented as:

$$\boldsymbol{S}=(CD_k, P_{dev,k-1}(t)) \tag{18}$$

$$P_{dev,k-1}(t)=P_{opt,i}(t)-P_{sum,k-1}^{EV}(t) \tag{19}$$

where CD_k is the charging demand of the k^{th} EV, including the arrival time $t_{arr,k}$, the charging power $P_{c,k}$, the charging time duration $t_{c,k}$, and the departure time $t_{dep,k}$; and $P_{dev,k-1}(t)$ is the deviation between the optimal charging power of node i $P_{opt,i}(t)$ and the scheduled charging power $P_{sum,k-1}^{EV}(t)$ of the first $k-1$ EVs. The state space $\boldsymbol{S}$ contains all knowledge of the current environment which can be obtained by the agent when scheduling the k^{th} EV at time t. The properties of MDP have decided that the future state only depends on the present state and the action taken by the agent. To be specific, the agent can only determine the charging start time of the current k^{th} EV and the scheduled charging power only depends on the scheduling result of the k^{th} EV.

2) The action space is represented as $\boldsymbol{A}=(t_{b,k})$ because the actions are taken sequentially, which determines the specific charging profile of the k^{th} EV combining its charging demands. In other words, the agent needs to make a decision when an EV arrives instead of scheduling all EVs together in the end. Due to the fixed electricity price, the feasible charging start time should be limited within the valley period to prevent extra charging costs. Considering the discrete feature for charging reservations, the action space is set as a discrete space with 1-min interval.

3) Every action taken by the agent will obtain a reward, which describes the performance of this action and contributes to improving the agent to achieve the maximum cumulative rewards. The reward function is defined as:

$$r_k=\rho(Dev_{k-1}-Dev_k) \tag{20}$$

$$Dev_k=\int_{t_s}^{t_e}\left|P_{opt,i}(t)-P_{sum,k}^{EV}(t)\right|\mathrm{d}t \tag{21}$$

where r_k is the reward gained by the agent after taking the action a_k; Dev_k is the deviation between the optimal charging power and the scheduled charging power of k EVs; $P_{opt,i}(t)$ is the optimal power of node i at time t; and ρ is the coefficient of reward, which is used to normalize the reward between different nodes with various EVs.

Moreover, the reward function can also reveal how much the charging demand is not satisfied or the charging costs have increased. Figure 4 illustrates the specific penalty when charging demands are satisfied or not. If the charging process of the k^{th} EV is beyond the boundary of valley period, the power deviation $Dev_k^{(b)}$ will be smaller than $Dev_k^{(a)}$ after dispatching the k^{th} EV, and the agent will obtain a smaller reward r_k.

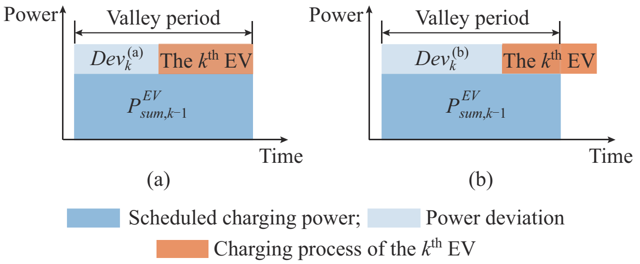

Fig. 4. Specific penalty when charging demands are satisfied or not satisfied. (a) Demands are satisfied. (b) Demands are not satisfied.

Theoretically, the total reward with optimal charging scheduling can be represented as:

$$\max E\left(\sum r_k\right)=\int_{t_s}^{t_e} P_{opt,i}(t)\mathrm{d}t \tag{22}$$

$$\rho=\frac{R_{norm}}{\max E\left(\sum r_k\right)} \tag{23}$$

where R_{norm} is a constant value for normalizing the reward.

From the perspective of the whole scheduling process, the agent will schedule all EVs to approach the optimal charging profiles so as to maximize the total reward. Nevertheless, considering the indivisibility of EV charging processes, it is tricky to realize the global optimum through the optimal decision of every single step. To be specific, the present decision has durable effects on the later charging scheduling processes, which are difficult to be involved into the optimization problem and solved using conventional methods. Based on the outstanding approximation ability of the neural network, DRL can take the subsequent effects into consideration. For example, the DRL agent may take an action that cannot gain the maximum reward at present, but it contributes to obtaining more rewards in the future and achieving the maximum total reward.

B. PPO Algorithm

Policy gradient is an essential method for training the DRL agent to maximize the cumulative reward, which works by computing an estimator of the policy gradient and plugging it into a stochastic gradient ascent algorithm [25]. The most commonly used gradient estimator $\hat{g}$ can be represented as:

$$\hat{g}=\hat{E}_t(\nabla_\theta \lg \pi_\theta(a_t|s_t)\hat{A}_t) \tag{24}$$

where π_θ is the stochastic policy function with parameter θ; $\hat{A}_t$ is the estimator of the advantage function at time t; a_t and s_t are the action and state, respectively; and $\hat{E}_t$ is the empirical average with finite samples.

As a result, the loss function is defined as:

$$L_{PG}(\theta)=\hat{E}_t(\lg \pi_\theta(a_t|s_t)\hat{A}_t) \tag{25}$$

However, traditional policy gradient methods have low utilization efficiency of sampling data and have to spend too much time on sampling new data once the policy is updated. Besides, it is difficult to determine appropriate steps for updating policy so as to prevent resulting in large differences between the new policy and the old policy.

Therefore, the PPO algorithm was proposed in 2017 to address the above shortcomings. The detailed training workflow of the DRL agent with PPO algorithm is demonstrated in Fig. 5. PPO algorithm consists of three networks, including two actor networks with the new policy π_θ and the old policy $\pi_{\theta'}$ (parameterized by θ and θ', respectively) and a critic network V_ϕ (parameterized by ϕ).

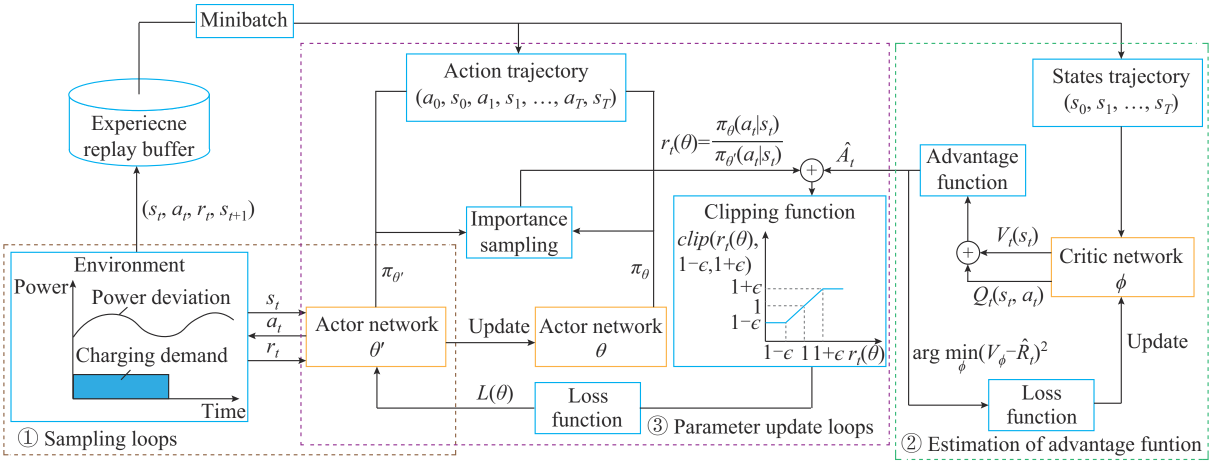

Fig. 5. Training workflow of DRL agent with PPO algorithm.

To increase the sample efficiency, $\pi_{\theta'}$ is used to interact with environments and sample N trajectory sets with T timesteps, while π_θ is the actual network that needs to be trained according to the demonstrations of $\pi_{\theta'}$. Utilizing importance sampling technology, the same trajectory sets can be used multiple times although there are differences between $\pi_{\theta'}$ and π_θ. The probability ratio of new policy and old policy can be expressed as:

$$r_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta'}(a_t|s_t)} \tag{26}$$

Another point of PPO algorithm is that the new policy should avoid significant evolution from the old policy after every update, so as to maintain the accuracy of importance sampling and avoid accident performance collapse. Hence, a clipped surrogate function is used to remove the incentive for moving $r_t(\theta)$ outside of the interval $[1-\epsilon, 1+\epsilon]$, so the loss function of PPO algorithm can be represented as [25]:

$$L(\theta)=\hat{E}_t(\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)) \tag{27}$$

where ϵ is the clipping parameter, which aims to clip the probability ratio. For instance, the objective will increase if the advantage function $\hat{A}_t$ is positive, but the increase is maintained within $1+\epsilon$ by a limit set by the clipping function.

Consequently, the network parameters θ of the new policy are updated using:

$$\theta=\arg\max_\theta \underset{s_t,a_t\sim\pi_{\theta'}}{E}(L(s_t,a_t,\theta',\theta)) \tag{28}$$

Apart from the actor network, a critic network is used to estimate the state value function and the advantage function. The advantage function describes how much an action is better than other actions on average, which is defined as:

$$\hat{A}_t=Q_t(s_t,a_t)-V_t(s_t) \tag{29}$$

$$Q_t(s_t,a_t)=E\left(\sum_{k=0}^{\infty}\gamma^k r_{t+k}\middle| s=s_t, a=a_t\right) \tag{30}$$

$$V_t(s_t)=E\left(\sum_{k=0}^{\infty}\gamma^k r_{t+k}\middle| s=s_t\right) \tag{31}$$

where γ is the discounting factor, which aims to balance the importance between immediate and future rewards; and V_t and Q_t are the value function and the action-value function, respectively. Therefore, $V_t(s_t)$ is the expected value on average at state s_t, which contains all optional actions. $Q_t(s_t,a_t)$ is the expected value at state s_t when taking action a_t.

The critic network V_ϕ is updated using regression to minimize a mean-squared-error objective [22]:

$$\phi=\arg\min_\phi (V_\phi(s_t)-\hat{R}_t)^2 \tag{32}$$

$$\hat{R}_t=\sum_{t'=t}^{T} r_t(s_{t'}, a_{t'}, s_{t'+1}) \tag{33}$$

where $\hat{R}_t$ is the reward-to-go, which is the sum of rewards after a point in the trajectory.

The DRL agent with PPO algorithm is trying to schedule the EV charging process according to the optimal charging profile, with the goal of maximizing the total expected rewards. The training workflow of PPO algorithm is summarized in Algorithm 1. The corresponding parameters are shown in Table I, where *lr* is the learning rate; and *MB* is the minibatch size.

The discounting factor γ and the clipping parameter ϵ are important hyperparameters that influence the performance agent observably. The importance of current action depends on the discounting factor γ, and a larger γ means that an agent is more long-sight so as to take full consideration of future uncertainties to achieve the maximum cumulative rewards. Thus, γ is set to be 0.99 [14].

Algorithm 1: training workflow of PPO algorithm

```
Initialize policy network $\pi_{\theta'}$ and value function network $V_\phi$
for $i=0; i<N; i++$ do
    Run policy $\pi_{\theta'}$ to interact with the environment for $T$ timesteps and obtains the trajectory samples $(s_t, a_t, r_t, s_{t+1})$
    Calculate the reward-to-go $\hat{R}_t$
    Use $V_\phi$ to estimate the advantage function $\hat{A}_t$
    Compute the loss function $L(\theta)$ with regard to $\theta$ with $K$ epochs of gradient decent
    $\pi_{\theta'} \leftarrow \pi_\theta$; $V_{\phi'} \leftarrow V_\phi$
end for
```

TABLE I
PARAMETERS OF PPO ALGORITHM

Parameter	Value	Parameter	Value
λ	0.15	N	2048
ϵ	0.2	K	10
γ	0.99	R_{norm}	1000
lr	3×10^{-4}	*MB*	64

Both the convergence speed and performance stability depend on ϵ [22], hence ϵ is set to be 0.2 to balance the training speed and total reward of the agent [25], [32]. The multilayer perceptrons of the policy network are composed of two hidden layers and the neurons of each layer are 64. The number of training episodes is set to be 250.

IV. CASE STUDY

To evaluate the performance of the proposed two-stage DRL-based charging scheduling scheme, case studies are conducted in this section.

A. Parameter Setting of ADN

An ADN for simulation is established based on the IEEE 14-node test feeder, as shown in Fig. 6, where 10 nodes are set as residential loads without regulation flexibility and 3 nodes are set as CSs. The ADN is operated by the DSO, with the goal of relieving the power congestion and shortening the peak-valley differences of tie-line power.

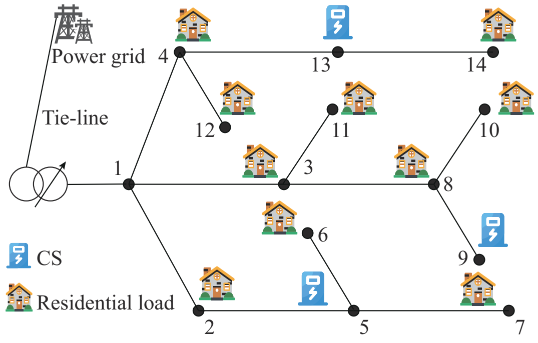

Fig. 6. ADN based on IEEE 14-node test feeder.

In accordance with the realistic situations in Zhejiang, China, the valley period of TOU is set from 22:00 to 08:00 of the next day. Meanwhile, the residential load data during the valley period are obtained from a housing estate in Hangzhou, Zhejiang, as shown in Fig. 7. Most residential loads have similar features, and the summits of the electricity consumption appear at 22:00 around. Then, the power demands continue to decline and reach the nadir at 03:00 of the next day. Finally, the electricity demands recover gradually as residents wake up. Therefore, there are significant peak-valley differences in residential distribution networks.

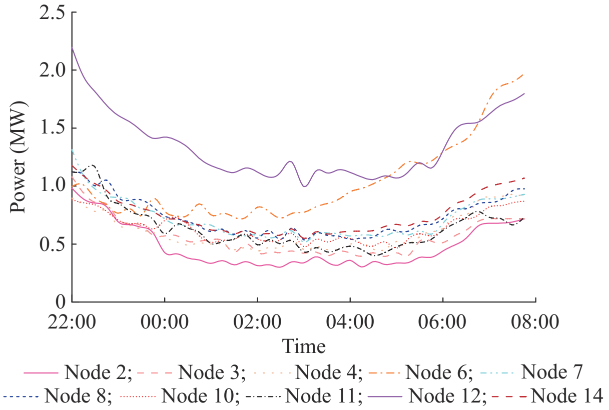

Fig. 7. Residential load data profiles during valley period.

B. OPF Results in First-stage Problem

The original and optimal charging profiles of EVs at different CSs are shown in Fig. 8.

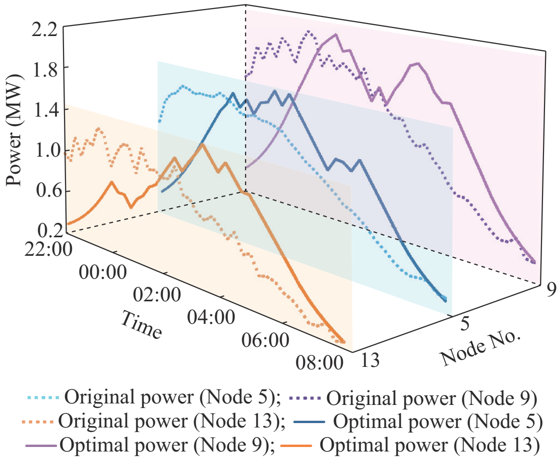

Fig. 8. Original and optimal charging profiles of EVs at different CSs.

The numbers of household EVs are set to be 179, 225, and 146 at node 5, node 9, and node 13, respectively. It is assumed that the charging power obeys the uniform distribution and the starting SOC obeys the normal distribution [16], whose parameters can be found in Table II. Moreover, the charging efficiency and battery capacity are set to be 90% and 50 kWh, respectively. Detailed charging data of household EVs are obtained from 550 individual meters which are equipped for household EVs. To simplify the charging scheduling, the arrival time distributes uniformly from 16:00 to 22:00, and the departure time is set consistently at 08:00 of the next day.

TABLE II
PARAMETERS OF EVS AT DIFFERENT NODES

Parameter	Description	Value
$P_{c,k}$ (kW)	Charging power	$U(5,25)$
η (%)	Charging efficiency	90
$C_{EV,k}$ (kWh)	Battery capacity	50
$SOC_{s,k}$ (%)	Starting SOC	$N(50,20)$
$SOC_{e,k}$ (%)	Expected SOC	100
$t_{arr,k}$ (hour)	Arrival time	$U(16{:}00, 22{:}00)$
$t_{dep,k}$ (hour)	Departure time	08:00

Note: normal distribution with the mean value of μ and the standard deviation of σ is abbreviated to $N(\mu,\sigma^2)$; uniform distribution with the minimum and maximum values of a and b, respectively, is abbreviated to $U(a,b)$.

TOU prices make great contributions to transferring charging demands from peak period to valley period. However, the charging processes cannot be dispatched effectively due to TOU prices are unable to describe the various demands of the ADN precisely during different time periods. Therefore, EV owners instinctively decide to start charging at the beginning of the valley period. As shown in Fig. 8, most charging processes start at 22:00, but the charging durations are much shorter than the valley period. The charging demands overlap with the residential peak, resulting in new power congestions at the beginning of the valley period, which threatens the secure and stable operation of the ADN.

To alleviate the power congestions and schedule the charging demands according to distribution network operations, the DSO needs to determine the optimal charging profiles of EV CSs by solving the OPF. Utilizing the DistFlow model introduced in Section II, the OPF of the ADN is calculated with the goal of flattening tie-line power, and the optimal charging profiles are shown in Fig. 8.

It can be observed that the main charging demands are transferred to 01: 00-05: 00, during which other electricity consumptions are the lowest. Moreover, the regulation targets are not allocated simply according to the total electricity demands of CSs; nodal voltages and impacts from other nodes are also taken into account to realize the multidimensional optimum. Therefore, the CSs are coordinated and the optimal charging profiles at different nodes are various, as shown in Fig. 8. For example, the charging summit of node 9 appears at 01:00 while that of node 13 appears at 03:00.

C. Charging Scheduling Results in Second-stage Problem

On the basis of optimal charging profiles calculated in the first stage, the DRL agent needs to schedule the charging processes of EVs sequentially to approach the optimal profiles. The charging scheduling results of household EVs at different nodes are shown in Fig. 9, where the deviation represents its absolute value. During the scheduling process, the agent makes decisions based on probability, which is calculated through massive pieces of training. All feasible actions are possible to be taken by the agent, although the probability of making a bad decision is very low. Therefore, it is inevitable for the agent to take bad actions that will cause deviations in a series of decision processes. It can be observed that the real power profiles are very close to the optimal power profiles except for some points, which proves the effectiveness of the DRL agent with PPO algorithm on charging scheduling.

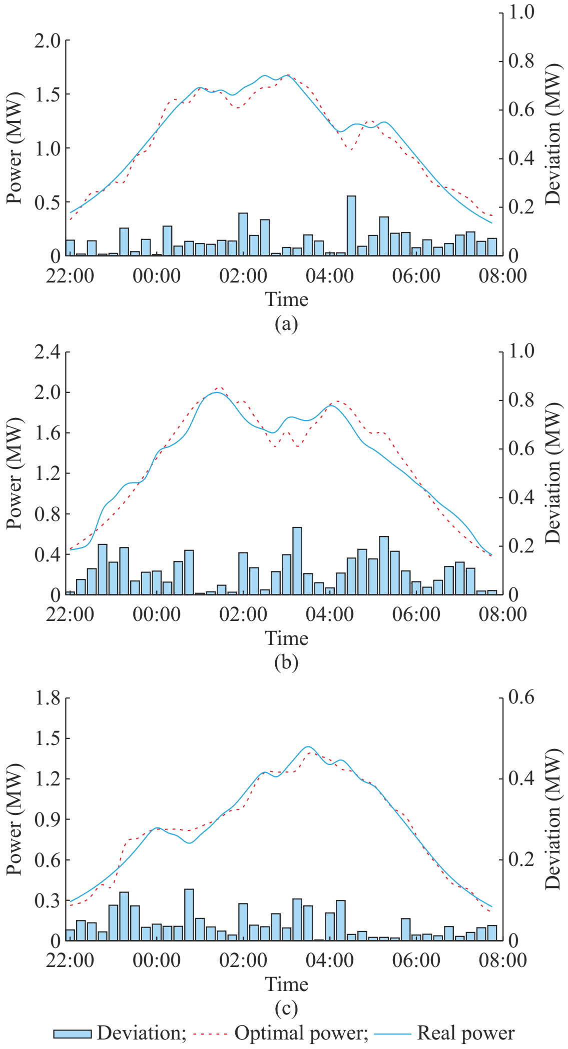

Fig. 9. Charging scheduling results of household EVs at different nodes. (a) Node 5. (b) Node 9. (c) Node 13.

As shown in Table III, the average deviations of node 5, node 9, and node 13 are 0.101 MW, 0.067 MW, and 0.044 MW, respectively, which are restricted at a relatively low level. Besides, it should be noted that significant deviations are likely to appear at the turning points of the optimal charging profile, e.g., the deviation of node 5 at 04:30 reaches 0.246 MW.

TABLE III
PERFORMANCE OF CHARGING SCHEDULING

Node No.	Average deviation (MW)	The maximum deviation (MW)	Total reward
5	0.101	0.246	939.5
9	0.067	0.277	923.1
13	0.044	0.127	946.6

During the whole charging scheduling processes, the DRL agent makes efforts to maximize the reward and obtains a total reward of 939.5, 923.1, and 946.6 for node 5, node 9, and node 13 in the end, respectively. Similar to the indexes of average deviation and the maximum deviation, the total reward indicates that the agent performs better with a smoother objective charging profile.

Moreover, the average SOC and median SOC at specific hours are further analyzed, as shown in Fig. 10. The median SOC reflects the charging completion result of every EVs. It can be observed that more than 50% EVs have finished their charging before 03:00 in the original charging scheduling, even though only half of valley period has passed. The results also indicate there is a significant regulation potential to be exploited for household EVs. The average SOC represents the overall charging progress of all EVs. It can be observed that the charging speed of the original charging is much faster than that of the proposed charging scheduling in the first half of the valley period, when the electricity demand is decreasing towards the nadir. Hence, the original charging scheduling cannot match the regulation demand of the ADN. On the contrary, the proposed charging scheduling takes full use of the shiftability of charging demands to reshape the charging curve with the goal of eliminating peak-valley differences in the ADN.

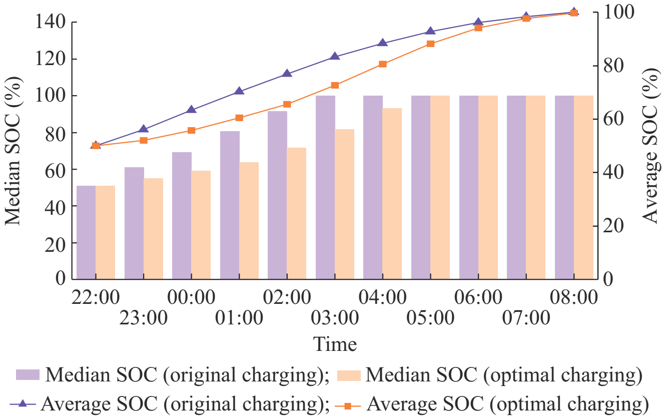

Fig. 10. Average SOC and median SOC during valley period.

D. Performance of PPO Algorithm

To verify the advantages of PPO algorithm, advantage actor critic (A2C) and deep Q-network (DQN) algorithms are implemented as the benchmarks. All training timesteps are set to be 512200 to analyze the total reward and the convergence speed. The cumulative reward is regarded as an index to evaluate the performance of the agents trained by different algorithms. Figure 11 illustrates the reward evolution curves of the PPO, A2C, and DQN algorithms during the training process.

The number of start points of the reward curve is regarded as the performance of the random policy, which is around 600. The PPO algorithm achieves the highest reward about 937. The PPO algorithm reaches a relatively stable state after 50 episodes (102400 timesteps). In the following 250 episodes, the PPO algorithm keeps exploring the optimal strategy and stabilizes its policy networks. Finally, the agent comprehensively reaches convergence with lower reward variances.

The A2C algorithm has a sharp increase at the beginning of the training process, which appears much faster than that of the PPO algorithm. The results prove that the clipped function of PPO algorithm has worked and limited the drastic change of policy network, so as to effectively avoid performance collapse and local optimum. As illustrated in Fig. 11, the A2C algorithm experiences an oscillation period after aggressive policy update and performance improvement, then it converges to the total reward around 908.

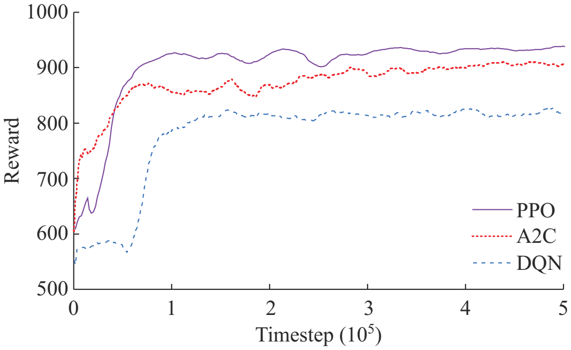

Fig. 11. Reward evolution curves of PPO, A2C, and DQN algorithms.

The DQN algorithm converges to the total reward around 825 and the PPO algorithm outperforms it by more than 13%, which proves the advantages of actor-critic networks. Besides, DQN spends much time collecting abundant samples and filling up its replay buffer, so there is no improvement at the beginning of the training process.

It takes a total time of 782.75 s, 586.83 s, and 542.21 s for PPO, A2C, and DQN algorithms in the entire training process, respectively. Besides, the decision time of the proposed DRL agent with PPO algorithm is also tested and the results indicate that the average decision time per EV is about 2.5 ms. These tests have been carried out using Python 3.7 on an Intel$^{(R)}$ i7 12700kf, 32 GB RAM desktop.

In terms of test results, the PPO algorithm outperforms the A2C algorithm, DQN algorithm, and random policy, although the PPO algorithm has the lowest training speed with the same timesteps. To be specific, the PPO algorithm can obtain the total reward of 937 when scheduling EV charging processes, which is 29, 112, and 337 more than that of the A2C algorithm, DQN algorithm, and random policy, respectively.

Then, the loss function performance of PPO algorithm is presented, as shown in Fig. 12. The value loss and loss represent the performance of PPO algorithm on training sets and test sets, respectively. It can be observed that the value loss and loss share similar trends during the training process, which demonstrate the remarkable adaptability on various data sets.

Hence, the PPO algorithm is suited for addressing the charging scheduling problem and can be adopted to handle the uncertainty of environment.

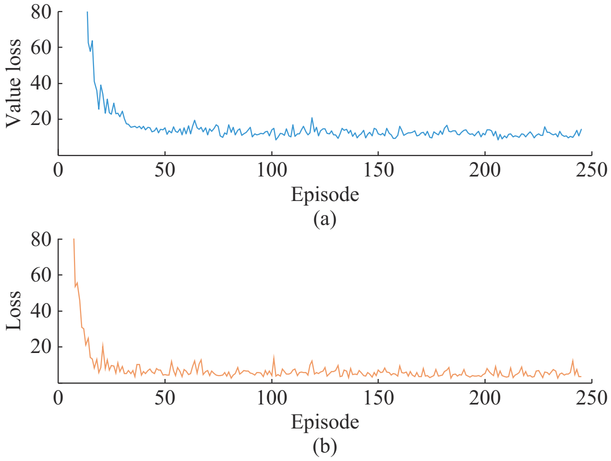

Fig. 12. Loss function performance of PPO algorithm during training process. (a) Value loss. (b) Loss.

E. Improvement for ADN

The original and optimized tie-line power profiles are demonstrated in Fig. 13. At first, the power congestions caused by intensive charging demands at the beginning of the valley period are eliminated effectively. Moreover, these charging demands are allocated to smooth the tie-line power. Thus, the regulation potential of household EVs is further exploited by ADN without extra costs, which benefits both the power system and EV owners. It can be observed that the peak-valley differences of ADN are dramatically eliminated from 6.61 MW to 2.76 MW, and the curtailment of peak load will save remarkable investments in electric power facilities. With further integrations of household EVs in the ADN, the proposed scheme can also be used to formulate a smooth tie-line power during the peak period under TOU prices.

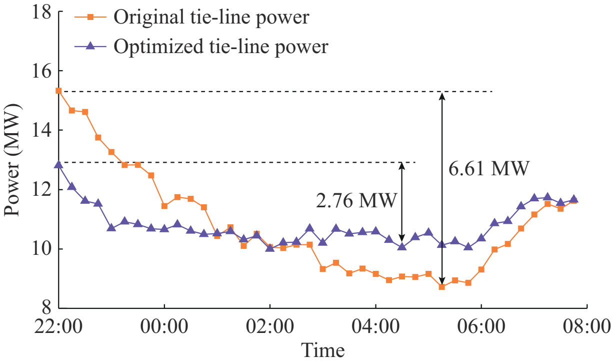

Fig. 13. Comparison between original and optimized tie-line power profiles.

Different from the transmission network, the distribution network possesses much higher resistance, and the active power has more significant effects on voltages. As a result, apart from great contributions to the elimination of peak-valley differences, the voltage quality of the ADN is also improved through scheduling household EVs during the valley period. As shown in Fig. 14, due to the overlap of the peak of residential electricity consumption and the intensive charging demands, some nodal voltages are extremely low at the beginning of the valley period, especially at node 9, whose nadir has reached 0.969 p.u.. In the near future, with more penetration of household EVs into the ADN, the voltage limit violation problem will be more serious if no strategies are taken.

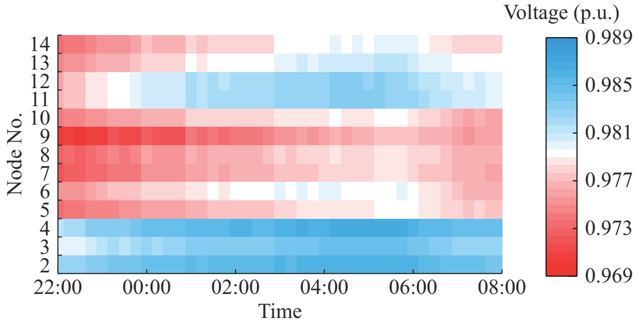

Fig. 14. Original nodal voltages.

Utilizing the proposed two-stage charging scheduling scheme, the voltage violation problem is addressed effectively, as shown in Fig. 15.

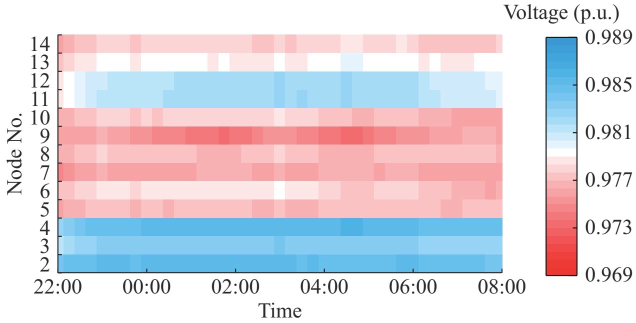

Fig. 15. Nodal voltages after charging scheduling.

Besides, the oscillations of nodal voltages are also limited with smoother tie-line power, which is beneficial for reducing the operating times of voltage regulation equipment such as on-load tap changers. For example, the voltage variations of node 4 decrease from 0.0051 p.u. to 0.0028 p.u. during the whole valley period. Meanwhile, the contributions to voltage regulation are not restricted to the nodes of CSs. As shown in Table IV, all nodal voltages have different degrees of improvement.

Because OPF is involved in the first-stage optimization problem, all nodal voltages are taken into consideration when determining the optimal charging profiles of CSs. Specifically, the voltage nadir of node 8 has 0.39% improvement. For these old communities with relatively poor electricity infrastructure, the proposed scheme can also satisfy residential power consumption and charging demands simultaneously with limited carrying capacity.

Therefore, under the existing TOU price circumstances, the proposed two-stage charging scheduling scheme can take full use of the regulation potential of household EVs during valley periods to improve the power quality of the ADN without extra equipment investments and charging costs, including peak-valley difference elimination, congestion management, and nodal voltage regulation.

TABLE IV
NODAL VOLTAGE IMPROVEMENT IN ADN

Node No.	Original voltage nadir (p.u.)	Optimized voltage nadir (p.u.)	Voltage improvement (%)
2	0.982	0.984	0.17
3	0.979	0.981	0.19
4	0.981	0.983	0.14
5	0.973	0.975	0.26
6	0.975	0.976	0.11
7	0.971	0.974	0.28
8	0.972	0.976	0.39
9	0.969	0.973	0.35
10	0.974	0.976	0.21
11	0.976	0.978	0.23
12	0.976	0.978	0.23
13	0.974	0.977	0.29
14	0.973	0.975	0.28

V. CONCLUSION

In the context of taking full use of the regulation potential of household EVs under TOU prices, this paper proposes a two-stage charging scheduling scheme to dispatch household EVs. The first-stage problem aims to involve the charging scheduling of household EVs in operation and optimization of the ADN, and the optimal charging power profiles of CSs are determined by calculating the OPF so as to relieve the power congestions and shorten the peak-valley differences. Furthermore, a PPO-based DRL agent is developed to dispatch the charging processes of EVs in terms of the optimal charging power. Case studies with realistic data are conducted to illustrate the multidimensional performance of the proposed scheme. It is demonstrated that the PPO-based DRL agent can be adopted in different CSs with various objective charging profiles and EV amounts. Besides, the charging scheduling of EVs contributes to significant improvement in power quality, including decreasing the peak-valley differences and stabilizing the nodal voltages.

Moreover, the proposed scheme can be adopted properly in substantial distributed communities with the combination of edge computing technology. On this basis, numerous flexible loads, e.g., thermostatic loads, energy storage, RES, can be involved into the proposed scheme to be managed efficiently, so as to activate their flexibility and enhance the regulation capacity of ADNs in the near future.

REFERENCES

[1] T. Chen, X.-P. Zhang, J. Wang, *et al.*, "A review on electric vehicle charging infrastructure development in the UK," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 2, pp. 193-205, Mar. 2020.

[2] IEA. (2022, May). Global EV outlook 2022. [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2022

[3] H. Liu, P. Zeng, J. Guo *et al.*, "An optimization strategy of controlled electric vehicle charging considering demand side response and regional wind and photovoltaic," *Journal of Modern Power Systems and Clean Energy*, vol. 3, no. 2, pp. 232-239, Jun. 2015.

[4] Fco. J. Zarco-Soto, J. L. Martínez-Ramos, and P. J. Zarco-Periñán, "A novel formulation to compute sensitivities to solve congestions and voltage problems in active distribution networks," *IEEE Access*, vol.

9, pp. 60713-60723, Apr. 2021.

[5] B. Wei, Z. Qiu, and G. Deconinck, "A mean-field voltage control approach for active distribution networks with uncertainties," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1455-1466, Mar. 2021.

[6] Y. Luo, Q. Nie, D. Yang, *et al.*, "Robust optimal operation of active distribution network based on minimum confidence interval of distributed energy beta distribution," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 2, pp. 423-430, Mar. 2021.

[7] K. Xie, H. Hui, Y. Ding *et al.*, "Modeling and control of central air conditionings for providing regulation services for power systems," *Applied Energy*, vol. 315, p. 119035, Jun. 2022.

[8] H. Wei, J. Liang, C. Li *et al.*, "Real-time locally optimal schedule for electric vehicle load via diversity-maximization NSGA-II," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 4, pp. 940-950, Jul. 2021.

[9] E. Hadian, H. Akbari, M. Farzinfar *et al.*, "Optimal allocation of electric vehicle charging stations with adopted smart charging/discharging schedule," *IEEE Access*, vol. 8, pp. 196908-196919, Oct. 2020.

[10] H.-M. Chung, S. Maharjan, Y. Zhang *et al.*, "Intelligent charging management of electric vehicles considering dynamic user behavior and renewable energy: a stochastic game approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 12, pp. 7760-7771, Jul. 2021.

[11] J. Hu, C. Ye, Y. Ding *et al.*, "A distributed MPC to exploit reactive power V2G for real-time voltage regulation in distribution networks," *IEEE Transactions on Smart Grid*, vol. 13, no. 1, pp. 576-588, Sept. 2022.

[12] S. Deb, A. K. Goswami, P. Harsh *et al.*, "Charging coordination of plug-in electric vehicle for congestion management in distribution system integrated with renewable energy sources," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, pp. 5452-5462, Sept. 2020.

[13] S. Das, P. Acharjee, and A. Bhattacharya, "Charging scheduling of electric vehicle incorporating grid-to-vehicle and vehicle-to-grid technology considering in smart grid," *IEEE Transactions on Industry Applications*, vol. 57, no. 2, pp. 1688-1702, Mar. 2021.

[14] L. Yan, X. Chen, J. Zhou *et al.*, "Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5124-5134, Jul. 2021.

[15] F. L. D. Silva, C. E. H. Nishida, D. M. Roijers *et al.*, "Coordination of electric vehicle charging through multiagent reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2347-2356, May 2020.

[16] L. Gan, X. Chen, K. Yu *et al.*, "A probabilistic evaluation method of household EVs dispatching potential considering users' multiple travel needs," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, pp. 5858-5867, Sept. 2020.

[17] E. L. Karfopoulos and N. D. Hatziargyriou, "A multi-agent system for controlled charging of a large population of electric vehicles," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1196-1204, May 2013.

[18] A.-M. Koufakis, E. S. Rigas, N. Bassiliades *et al.*, "Offline and online electric vehicle charging scheduling with V2V energy transfer," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 2128-2138, May 2020.

[19] Y. Li, M. Han, Z. Yang *et al.*, "Coordinating flexible demand response and renewable uncertainties for scheduling of community integrated energy systems with an electric vehicle charging station: a bi-level approach," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 4, pp. 2321-2331, Oct. 2021.

[20] 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, May 2022.

[21] D. Silver, A. Huang, C. J. Maddison *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484-489, Jan. 2016.

[22] B. Huang and J. Wang, "Deep-reinforcement-learning-based capacity scheduling for PV-battery storage system," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2272-2283, May 2021.

[23] D. Qiu, Y. Ye, D. Papadaskalopoulos *et al.*, "A deep reinforcement learning method for pricing electric vehicles with discrete charging levels," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, pp. 5901-5912, Sept. 2020.

[24] 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, May 2020.

[25] J. Schulman, F. Wolski, P. Dhariwal *et al.* (2017, Jul.). Proximal policy optimization algorithms. [Online]. Available: http://arXiv: 1707.06347

[26] S. Yoon and E. Hwang, "Load guided signal-based two-stage charging coordination of plug-in electric vehicles for smart buildings," *IEEE Access*, vol. 7, pp. 144548-144560, Oct. 2019.

[27] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Transactions on Power Delivery*, vol. 4, no. 2, pp. 1401-1407, Apr. 1989.

[28] S. Tan, J.-X. Xu, and S. K. Panda, "Optimization of distribution network incorporating distributed generators: an integrated approach," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2421-2432, Aug. 2013.

[29] C. Zhang, Y. Liu, F. Wu *et al.*, "Effective charging planning based on deep reinforcement learning for electric vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 542-554, Jan. 2021.

[30] S. Han, S. Han, and K. Sezaki, "Development of an optimal vehicle-to-grid aggregator for frequency regulation," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 65-72, Jun. 2010.

[31] Z. Zhao and C. K. M. Lee, "Dynamic pricing for EV charging stations: a deep reinforcement learning approach," *IEEE Transactions on Transportation Electrification*, vol. 8, no. 2, pp. 2456-2468, Jun. 2022.

[32] W. Zhu and A. Rosendo, "A functional clipping approach for policy optimization algorithms," *IEEE Access*, vol. 9, pp. 96056-96063, Jul. 2021.

Taoyi Qi received the B.S. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2020, where he is currently pursuing the M.S. degree in electrical engineering. His research interests mainly focus on demand response, flexible loads, and deep reinforcement learning.

Chengjin Ye received the B.E. and Ph.D. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2010 and 2015, respectively. From 2015 to 2017, he served as a Distribution System Engineer with the Economics Institute, State Grid Zhejiang Electric Power Company Ltd., Hangzhou, China. From 2017 to 2019, he was an Assistant Research Fellow with the College of Electrical Engineering, Zhejiang University. Since 2020, he has been a Tenure-track Professor there. His research interests mainly include resilience enhancement of power grids and integrated energy systems, as well as market mechanism and control strategy towards the integration of demand resources into power system operation.

Yuming Zhao received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, Tsinghua University, Beijing, China, in 2001 and 2006, respectively. He is currently a Senior Engineer (Professor level) with Shenzhen Power Supply Co., Ltd., Shenzhen, China. His main research interest includes the DC distribution power grid.

Lingyang Li is currently pursuing the Ph.D. degree in electrical engineering at Zhejiang University, Hangzhou, China. His research interests include the optimal operation of distribution networks and mobile energy storage system optimization scheduling.

Yi Ding received the bachelor's degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2002, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2007. He is currently a Professor at the College of Electrical Engineering, Zhejiang University, Hangzhou, China. His current research interests include power systems reliability analysis incorporating renewable energy resources, smart grid performance analysis, and engineering systems reliability modeling and optimization.