Electric Vehicles Management in Distribution Network: A Data-Efficient Bi-level Deep Reinforcement Learning Method

Hongrong Yang, *Student Member, IEEE*, Yinliang Xu, *Senior Member*, Hongbin Sun, *Fellow IEEE*, Qinglai Guo, *Senior Member, IEEE*, and Qiong Liu, *Student member, IEEE*

***Abstract*—In this paper, a deep reinforcement learning (DRL)-based electric vehicles (EVs) management strategy is proposed to achieve peak shaving and regulate the voltage violations in distribution networks. We present a new approach for modeling the EV user willingness considering the effect of multi-attribute attitudes and price incentives on user' behaviors. The real-time optimal regulation problem is modeled as a bi-level optimization and then formulated as a Markov decision process. To solve the problem, we first define an EV energy availability criterion for K-means clustering method to assess the EV scheduling potential for hierarchical dispatching, which facilitates meeting the user charging demand constraint. Then we present a bi-level soft actor critic algorithm to solve the optimal gaming problem between distribution system operator (DSO) and charging stations (CSs) aggregator, and develop a data-enhanced method using Gaussian mixture model with a two-stage training process to improve the data efficiency and model robustness, as well as avoiding the constraint violations in the start-up process. Test results on the modified IEEE 33-bus and 118-bus systems with the real-world data are presented to validate the effectiveness of the proposed approach.**

***Index Terms*—electric vehicle, willingness model, peak shaving, voltage control, deep reinforcement learning, data efficient.**

I. Introduction

According to the reports by International Energy Agency, the scale of the global electric vehicle (EV) fleet is expected to reach 54 million in 2025. Uncoordinated large scale EVs charging with their random behaviors may impose severe burdens on the distribution network such as peak load stacking and voltage violations [1]. Under a coordinated EVs charging scheme, EVs can provide substantial ancillary services to assist the economical and safe operation of the distribution network.

The active power control of vehicle-to-grid (V2G) technology has been well applied in power grid operations, especially in peak shaving [2]. In [3], two coordinated charge and discharge schemes of EVs were proposed to shape the load curve of residential communities. However, the study did not set an effective SOC constraint to meet the charging demand of EV users. Sun *et al.* [4] proposed a customized EV charging criterion and set an average SOC constraint in reward function to limit the discharging power. But the weight coefficient which reflects a tradeoff between SOC and parking time of EVs is hard to determine. In [5], an optimal dispatch scheme considering the influence of both economic and emission was proposed to achieve peak shaving. But the results were based on the assumption that at least 80% EV users were willing to provide the V2G service without considering the incentives. Gupta *et al.* [6] proposed a charging willingness model based on scheduling parameters and Xu *et al.* [7] established a psychological model with the aim to evaluate the influence of incentives on EV users' discharging decisions. These two studies analyzed the users' willingness of participation and proposed linear models to reflect actual reactions of users to the price incentives.

As the development of four-quadrant smart charger, EVs can also play an important role in the voltage regulation of distribution networks. The charger not only realizes V2G functionality, but also achieves the bidirectional reactive power control (RPC) [8] [9]. Many recent works have provided effective solutions based on the RPC technology to harness EV flexibility on the voltage control. In [10], a novel optimal hybrid control scheme was proposed to regulate the dynamic voltage response by handling the injection of the reactive power of EV chargers. In [11], the authors developed a hierarchical coordination strategy to optimize the grid operations and EV charging process with reactive power support. In [12], photovoltaics (PVs) inverters and EVs were coordinated to mitigate voltage fluctuations by a decentralized voltage control algorithm considering both active and reactive power compensation. Hoque *et al*. [13] proposed a sensitivity-based real-time voltage control framework by regulating the EV energy.

However, the above methods for peak shaving and voltage regulation are all fully model-based approaches that rely on the complete and accurate information about the EV, users and power grid, which can be difficult to obtain in practice, especially information under privacy limitations, such as users trip. Besides, the model-based method may not be able to effectively address the complex optimization problem due to the time-varying operating conditions and the uncertainty of EV charging behaviors. Furthermore, with increasing size of the distribution network and scale of EVs, the computational complexity of model-based methods will impose a significant burden [14] to the system operator.

These shortcomings of model-based methods motivate the application of deep reinforcement learning (DRL) method which learns the approximate control policy by interacting with the practical system or high-fidelity simulator. In terms of voltage control and peak shaving, there have been a few studies on solving the complex decision problem using DRL methods [15]-[19]. Wang *et al*. [15] formulated the grid voltage regulation problem as an Markov Game solved by a multi-agent DRL

approach. Nguyen *et al.* [16] proposed a three-stage inverter-based peak shaving and reactive power control framework solved by the online safe DRL algorithm. Li *et al.* [17] proposed a safe DRL method to utilize EVs to support the stable operation of system and reduce peak loads. Cao *et al.* [18] proposed a customized actor–critic learning framework to minimize EV charging cost and curtail the peak charging loads. In [19], a dueling deep Q network DRL method interruptible load was presented to reduce both peak load demand and operating costs in the premise of the voltage safety.

For the above-mentioned works, EV charging regulation schemes for ancillary services can be summarized into two categories: EV charging coordination and price incentives, however, the existing schemes consider only one of them, which makes the proposed schemes impractical due to the strong correlation between these two terms. Therefore, this paper proposes a two-stage deep reinforcement learning (DRL)-based EVs coordinated dispatching for ancillary service provision considering price incentives and EV users' willingness of participation. The main contributions of this paper are summarized as follows:

1) We propose a DRL-based coordinated charging scheme considering price incentives to solve the real-time peak shaving and voltage control issue. The problem is modeled as a bi-level optimization and then formulated as an Markov decision process.

2) We propose a reasonable approach for modeling the EV user's charging and discharging willingness considering the influence of multi-attribute attitudes and price incentives on EV users' discharging decisions.

3) We design a data-efficient bi-level soft actor critic algorithm to solve the proposed bi-level optimization problem. The algorithm is trained in two-stage training process using a Gaussian mixture model-based data-enhanced method to improve the data efficiency and the robustness of the trained policy against to the uncertainty of user's charging behavior, as well as avoiding the severe violations in the start-up process of online training.

4) Unlike the works in [3] - [5], which directly use SOC and parking time as charging constraints, we propose an EV energy availability criteria using K-means clustering method to assess the EV energy state, which contributes to ensure EV charging demand upon departure.

The rest of this paper is organized as follows: Section II formulates the mathematical models for the distribution network, EV, users' willingness evaluation, and the bi-level optimization objective function, respectively. Section III presents the real-time peak shaving and voltage regulation scheme with the data-efficient bi-level SAC method. Section IV conducts the case study on the modified IEEE 33 and 118 distribution networks, and Section V concludes this paper.

II. Mathematic Models

A. Distribution Network Model

The distribution network is defined as follows: Let $\boldsymbol{B} = \{0,1,2,\dots,b\}$ denote the set of buses and $\boldsymbol{L} \subset \boldsymbol{B} \times \boldsymbol{B}$ denote the set of lines. For each $(m,n) \in \boldsymbol{L}$, m is the unique parent bus of bus n. Then, we can obtain the radial distribution network model as:

$$\begin{cases} \sum_{mn\in\boldsymbol{L}} (P_{mn,t} - r_{mn} l_{mn,t}) - \sum_{nk\in\boldsymbol{L}} P_{nk,t} = P_{n,t}^{CS} + P_{n,t}^{O} \\ \sum_{mn\in\boldsymbol{L}} (Q_{mn,t} - r_{mn} x_{mn,t}) - \sum_{nk\in\boldsymbol{L}} Q_{nk,t} = Q_{n,t}^{CS} + Q_{n,t}^{O} \end{cases} \tag{1}$$

$$\begin{cases} P_{n,t}^{CSs} = \sum_{i=1}^{N_{n,t}^{S}} P_{n,i,t}^{EV,C} + \sum_{i=1}^{N_{n,t}^{RPC}} P_{n,i,t}^{EV,C} - \sum_{i=1}^{N_{n,t}^{V2G}} P_{n,i,t}^{EV,D} \\ Q_{n,t}^{CSs} = \sum_{i=1}^{N_{n,t}^{S}} Q_{n,i,t}^{EV,C} - \sum_{i=1}^{N_{n,t}^{RPC}} Q_{n,i,t}^{EV,S} - \sum_{i=1}^{N_{n,t}^{V2G}} Q_{n,i,t}^{EV,S} \end{cases} \tag{2}$$

$$v_{n,t} = v_{m,t} + 2(P_{mn,t} r_{mn} + Q_{mn,t} x_{mn}) - l_{mn,t}(r_{mn}^2 + x_{mn}^2) \tag{3}$$

$$\left\| \begin{matrix} 2P_{mn,t} \\ 2Q_{mn,t} \\ l_{mn,t} - v_{n,t} \end{matrix} \right\|_2 \le l_{mn,t} + v_{m,t} \tag{4}$$

$$V_{\min}^2 \le v_{n,t} \le V_{\max}^2 \tag{5}$$

$$0 \le l_{mn,t} \le I_{\max}^2 \tag{6}$$

$$P_{n,t}^{CSs} \le P_{n,\max}^{DSO} \tag{7}$$

where $P_{mn,t}/Q_{mn,t}$ and r_{mn}/x_{mn} are the active/reactive power and impedance/inductive resistance of line (m, n). $P_{n,t}^{CSs}/Q_{n,t}^{CSs}$, $P_{n,t}^{O}/Q_{n,t}^{O}$ are the active/reactive power of charging stations (CSs), non-EV loads, respectively. $P_{n,i,t}^{EV,C}/P_{n,i,t}^{EV,D}$ is the charging/discharge active power of EV i at bus n during time stage t. $Q_{n,i,t}^{EV,C}$ is the reactive power consumed in charging process while $Q_{n,i,t}^{EV,S}$ is the reactive power support from EVs to power grid. $N_{n,t}^{S}, N_{n,t}^{V2G}$, $N_{n,t}^{RPC}$ is the number of EVs in standard charging /V2G / RPC mode. (1) formulates the linearized power balance. (2) denotes the components of CSs active and reactive power. (3) describes the relationship between node voltage and branch current. (4) is the convex relaxation of power flow equality constraints based on second order cone programming. (5)-(6) are the constraints on node voltage amplitude and the line current, where V_{min}/V_{max} is the minimum/maximum voltage magnitude limits and I_{min} is the maximum line current capacity. (7) is the power constraint for peak shaving, where $P_{n,max}^{DSO}$ is the maximum peak power of bus n, which is set by the distribution system operator (DSO).

B. EV Model

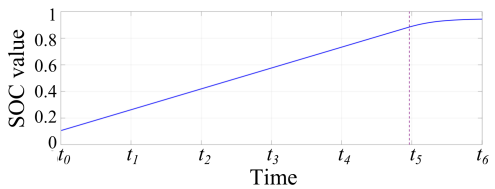

Fig. 1 Trajectory of the EV's SOC in charging process.

Practically, EV charging process is relatively complex, which generally goes through four stages: 1) current limiting stage (i.e., the battery pre-charging stage, the voltage and current rise slowly), 2) constant current stage, 3) constant power stage and 4) constant voltage stage. An EV generally starts charging with a residual capacity of the battery, so daily charging only goes through the last two stages, and the constant voltage charging stage is very short.

As shown in Fig. 1, EV charging process can mainly be divided into two stages, we approximate the charging power curve with a linear function when SOC is less than 0.9, then the EV charging and discharging characteristics can be generally modelled as:

$$P_{n,i,t}^{EV,C} = \begin{cases} P_{n,i,t}^{EV,RC}, & \text{SOC}<0.9 \\ \tau_1 t + \tau_2, & SOC \geq 0.9 \end{cases} \tag{8}$$

$$E_{n,i,t+1}^{EV} = E_{n,i,t}^{EV} + \Delta t\left(P_{n,i,t}^{EV,C}\eta_{n,i,t}^{EV,C} - P_{n,i,t}^{EV,D} / \eta_{n,i,t}^{EV,D}\right) \tag{9}$$

$$SOC_{n,i,t+1}^{EV} = SOC_{n,i,t}^{EV} + (E_{n,i,t+1}^{EV} - E_{n,i,t}^{EV}) / B_{n,i}^{EV} \tag{10}$$

$$0 \leq P_{n,i,t}^{EV,C} \leq b_{n,i,t}^{EV,C} \cdot P_{n,i,\max}^{EV,C} \tag{11}$$

$$0 \leq P_{n,i,t}^{EV,D} \leq b_{n,i,t}^{EV,D} \cdot P_{n,i,\max}^{EV,D} \tag{12}$$

$$0 \leq b_{n,i,t}^{EV,D} + b_{n,i,t}^{EV,C} \leq 1 \tag{13}$$

$$SOC_{n,i,\min}^{EV} \leq SOC_{n,i,t}^{EV} \leq SOC_{n,i,\max}^{EV} \tag{14}$$

$$0 \leq \left|Q_{n,i,t}^{EV}\right| \leq \alpha_{n,i,t}^{EV}\sqrt{S_{n,i,\max}^{EV\ 2} - P_{n,i,t}^{EV\ 2}} \tag{15}$$

$$SOC_{n,i,t_l}^{EV} \geq SOC_{n,i,t_l}^{EV,EX} \tag{16}$$

where $P_{n,i,t}^{EV,RC}$ is the EV rated charging power, τ_1 and τ_2 is the factors of the approximate charging power function when the SOC is higher than 0.9. $\eta_{n,i,t}^{EV,C}/\eta_{n,i,t}^{EV,D}$ is the charging/discharging efficiency. $E_{n,i,t}^{EV}$ is the EV energy and $B_{n,i}^{EV}$ is the EV battery capacity. $P_{n,i,max}^{EV,C}/P_{n,i,max}^{EV,D}$ is the maximum charging /discharging power with the binary charging/discharging status $b_{n,i,t}^{EV,C}/b_{n,i,t}^{EV,D}$ and $SOC_{n,i,min}^{EV}/SOC_{n,i,max}^{EV}$ is the minimum/maximum SOC of EV battery. $P_{n,i,t}^{EV}, Q_{n,i,t}^{EV}$ and $S_{n,i,max}^{EV}$ are the charging active, reactive and maximum apparent power, respectively. $\alpha_{n,i,t}^{EV}$ is the charger power factor. SOC_{n,i,t_l}^{EV} denotes the EV SOC at departure and $SOC_{n,i,t_l}^{EV,EX}$ is the expected EV energy state. (8)-(15) describe the EV charging/discharging process and establish the EV energy storage model with $t_{n,i}^{a}/t_{n,i}^{l}$ as arrival/leaving time stage. (16) indicates that SOC at departure should meet the EV user's energy demand.

C. Price-Charging Willingness Model

Here, the price-reasonableness curve is adopted to describe the reasonableness of charging price. Each user's charging price-reasonableness curve is an S-shaped curve which can be formulated as follows:

$$R_{n,t} = c_{n,t} / (1 + e^{a_{n,t} - b_{n,t}/w_{n,t}^{T}}),\ a_{n,t}, b_{n,t}, c_{n,t} > 0 \tag{17}$$

$$w_{n,t}^{T} = w_{n,t}^{C} + w_{n,t}^{S} \tag{18}$$

$$w_{n,t}^{\min} \leq w_{n,t}^{S} \leq w_{n,t}^{\max} \tag{19}$$

where $R_{n,t}$ is the reasonableness of charging price which ranges from 0 to 1. The total charging price w_t^T is the sum of charging price w_t^C and service fee w_t^S. The parameters $a_{n,t}$, $b_{n,t}$ and $c_{n,t}$ decide the shape of the curve, which is mainly affected by the historical charging price. Such S-shaped value function is general and basic to measure the consumer's attitude to pay for goods [20] and has been widely applied in management and economics science [21]. Study results have demonstrated that S-shaped curve is highly consistent with the actual consumer's behavior. [22] [23].

Because EV users generally would not travel across regions when the battery is low, it is defaulted that each EV user would go to the local CSs to charge the EV. By collecting evaluation information of users in their local CSs, the charging preference $F_{n,i,t}$ is established according to the Fishbein Model, which is formulated as follows:

$$F_{n,i,t} = \sum_{j=1}^{k} B_{n,i,t}^{j} D_{n,i,t}^{j} \tag{20}$$

$$0 \leq B_{n,i,t}^{j} \leq 1 \tag{21}$$

$$\sum_{j=1}^{k} D_{n,i,t}^{j} = 1 \tag{22}$$

where $F_{n,i,t}$ is the charging preference of user on their local CSs. $B_{n,i,t}^{j}$ indicates the user satisfaction in attribute j of CSs and $D_{n,i,t}^{j}$ indicates the user preference degree for attribute j. k is the number of attributes of CSs. The attributes include geographical location, service quality, average charging waiting time, *etc*.

$R_{n,i,t}^{H}$ is the reasonableness threshold of charging behavior, which is defined as:

$$R_{n,i,t}^{H} = S_{n,i,t}(1 - F_{n,i,t}) \tag{23}$$

$$S_{n,i,t} = \begin{cases} 1 - (SOC_{n,i,t}^{C} - 1)^{\xi_{n,i,t}} & \xi_{n,i,t} = 2k \\ 1 + (SOC_{n,i,t}^{C} - 1)^{\xi_{n,i,t}} & \xi_{n,i,t} = 2k+1 \end{cases},\ k \in \boldsymbol{Z}^{*} \tag{24}$$

where $S_{n,i,t}$ is the stress of losing electricity that measures the effect of SOC variation on user's charging behavior. $SOC_{n,i,t}^{C}$ is the threshold of battery energy level for the user's charging decision. $\xi_{n,i,t}$ is the coefficient that reflects the user's resistance to the low battery anxiety.

One can get the threshold price of charging behavior $w_{n,i,t}^{H}$ according to (17) and (23), then the expected charging participation rate $A_{n,h,t}^{C}$ of the discrete total charging price point w_h^T can be obtained as:

$$A_{n,h,t}^{C} = \frac{N_{n,h,t}^{C}}{N_{n,s}} \tag{25}$$

$$N_{n,h,t}^{C} = \sum_{i}^{N_{n,s}} d_{n,i,t}^{h} \tag{26}$$

$$d_{n,i,t}^{h} = \begin{cases} 1 & w_h^T \geq w_{n,i,t}^{H} \\ 0 & w_h^T < w_{n,i,t}^{H} \end{cases} \tag{27}$$

where $N_{n,h,t}^{C}$ is the number of surveyed users who wish to charge their EVs at price w_h^T while $N_{n,s}$ is the total number of users surveyed. $d_{n,i,t}^{h}$ is the charging decision Boolean variable.

Then we use the interpolation method to get the expected continuous price-participation curve and apply the neural network to fit the nonlinear function relationship between total charging price and user charging participation. That means the actual charging willingness may not be possible to formulated as a mathematics function in reality.

D. Profits-Discharging Willingness Model

1) *EV battery cost model*

The EV battery cost model is a nonlinear model with respect to the depth of discharge (DOD), discharge power and other factors due to the nonlinear degradation process of lithium-ion battery. According to the simplified battery cost model in [24], EV battery degradation cost $C_{n,i,t}^{EV,D}$ during the V2G process is formulated as:

$$C_{n,i,t}^{EV,D} = w_{n,i,t}^{L} \Delta e_{n,i,t}^{D} \tag{28}$$

$$w_{n,i,t}^{L} = In_{n,i}^{EV,B} / B_{n,i}^{EV} l_{n,i,t}^{EV,B}(dod) \eta_{n,i,t}^{EV,D} \tag{29}$$

$$l_{n,i,t}^{EV,B}(dod) = 694(dod_{n,i,t}^{EV,B})^{-0.795} \tag{30}$$

$$dod_{n,i,t}^{EV,B} = 1 - E_{n,i,t}^{EV} / \mathrm{B}_{n,i,t}^{EV} \tag{31}$$

$$\Delta e_{n,i,t}^{D} = P_{n,i,t}^{EV,D} \Delta t \tag{32}$$

where $w_{n,i,t}^{L}$ is unit battery loss price and $\Delta e_{n,i,t}^{D}$ is the energy discharged during time stage t. $In_{n,i}^{EV,B}$ is the investment of EV battery. $dod_{n,i,t}^{EV,B}$ and $l_{n,i,t}^{EV,B}(dod)$ are the DOD and battery cycle number, respectively.

2) *EV user's profits of V2G process*

For an EV user, the total charging price remains the same since the EV arrival while the discharging price changes every time stage. Then the user's profits $\Pi_{n,i,t}^{EV,V2G}$ is calculated as:

$$\Pi_{n,i,t}^{EV,V2G} = \Pi_{n,i,t}^{EV,S} - C_{n,i,t}^{EV,CB} - C_{n,i,t}^{EV,D} \tag{33}$$

$$\Pi_{n,t}^{EV,V2G} = w_{n,i,t}^{D} \Delta e_{n,i,t}^{D} \tag{34}$$

$$C_{n,i,t}^{EV,CB} = w_{n,i,t_a}^{T} \Delta e_{n,i,t}^{D} \tag{35}$$

where $\Pi_{n,i,t}^{EV,S}$ is the profits for selling the energy in V2G service, and $\mathrm{C}_{n,i,t}^{EV,CB}$ is the cost for charging back the discharged energy $\Delta e_{n,i,t}^{D}$. w_{n,i,t_a}^{T} is the total charging price of CSs at bus n during the arrival time of EV i. $w_{n,i,t}^{D}$ is the discharging price of CSs.

3) *Profits of CSs and DSO*

The total profits $\Pi_{n,t}^{CSs}$ of CSs is formulated as follows:

$$\Pi_{n,t}^{CSs} = \Pi_{n,t}^{CSs,C} + \Pi_{n,t}^{CSs,V2G} - C_{n,t}^{CSs,E} - C_{n,t}^{CSs,V2G} - C_{n,t}^{CSs,P} \tag{36}$$

$$\Pi_{n,t}^{CSs,C} = \sum_{i}^{N_{n,t}^{RPC}+N_{n,t}^{S}} w_{n,i,t_a}^{T} \Delta e_{n,i,t}^{C} \tag{37}$$

$$\Delta e_{n,i,t}^{C} = P_{n,i,t}^{C} \Delta t \tag{38}$$

$$C_{n,t}^{CSs,E} = \sum_{i}^{N_{n,t}^{RPC}+N_{n,t}^{S}} w_{n,t}^{E} \Delta e_{n,i,t}^{C} \tag{39}$$

$$\Pi_{n,t}^{CSs,V2G} = \begin{cases} w_{n,t}^{I} P_{n,t}^{CSs,D} \Delta t, & P_{n,t}^{CSs,D} \le P_{n,t}^{CSs,B} \\ w_{n,t}^{I} P_{n,t}^{CSs,B} \Delta t, & 0.7 P_{n,t}^{CSs,B} \le P_{n,t}^{CSs,D} \le P_{n,t}^{CSs,B} \\ 0, & P_{n,t}^{CSs,D} \le 0.7 P_{n,t}^{CSs,B} \end{cases} \tag{40}$$

$$w_{n,t}^{E} < w_{n,t}^{I} \le w_{n,t}^{G,MAX} \tag{41}$$

$$C_{n,t}^{CSs,P} = \begin{cases} w_{n,t}^{P}(0.7 P_{n,t}^{CSs,B} - P_{n,t}^{CSs,D}) \Delta t, & P_{n,t}^{CSs,D} < 0.7 P_{n,t}^{CSs,B} \\ 0, & P_{n,t}^{CSs,D} \ge 0.7 P_{n,t}^{CSs,B} \end{cases} \tag{42}$$

$$P_{n,t}^{CSs,D} = \sum_{i=1}^{N_{n,t}^{V2G}} P_{n,i,t}^{EV,D} \tag{43}$$

$$C_{n,t}^{CSs,V2G} = \sum_{i}^{N_{n,t}^{V2G}} \Pi_{n,i,t}^{EV,V2G} \tag{44}$$

where Δt is the time interval which is taken as 15 minutes in this study with the start/end time stage T_n^{start}/T_n^{end}. $\Pi_{n,t}^{CSs,C}$ is the charging profits while $C_{n,t}^{CSs,E}$ is the cost for purchase of electricity with purchase price $w_{n,t}^{E}$. $\Pi_{n,t}^{CSs,V2G}$ is the CSs profits from selling energy to the power grid with the incentive electricity price $w_{n,t}^{I}$ paid by the DSO. $C_{n,t}^{CSs,P}$ is the penalty if the actual output does not reach 70% bidding volume $P_{n,t}^{CSs,B}$ and $w_{n,t}^{P}$ is the penalty price. $C_{n,t}^{CSs,V2G}$ is the CSs V2G cost for buying the users' discharged energy. $P_{n,t}^{CSs,D}$ is the CSs total discharged power. (41) ensures the profits of public sector of the grid with the maximum price of generation cost $w_{n,t}^{G,MAX}$ according to [25].

We assume that the marginal price of each bus has been set by the day-ahead market and the non-EV loads can be predicted. In this case, the profits of DSO can be formulated as (46) ignoring the other profits components for computing convenience, which is calculated as:

$$\Pi_{t}^{DSO} = \sum_{n \in \boldsymbol{B}} (C_{n,t}^{CSs,E} + C_{n,t}^{CSs,P} - C_{n,t}^{DSO,E} - \Pi_{n,t}^{CSs,V2G}) \tag{45}$$

$$C_{n,t}^{DSO,E} = w_{n,t}^{G} P_{n,t}^{CSs} \tag{46}$$

where $w_{n,t}^{G}$ is the DSO electricity purchase price from the grid.

4) *Weber-Fechner Law and Profits-Discharging Willingness Model*

The Weber-Fechner Law can properly express the functional relationship between the EV user's response and the objective environment stimulus. It was first applied in the fields of psychology and acoustics.

Weber-Fechner Law points out that the magnitude of sensation is directly proportional to the logarithm of stimulus intensity, which is defined as follows:

$$s = k \log(I) + s_0 \tag{47}$$

where s is the human sensory intensity, I is the stimulus intensity of the external environment. k is the Weber's index, which is sense-specific and should be determined according to the sense and type of stimulus. s_0 is the integral constant of stimulus. According to [26], the natural way that number is encoded in a brain is logarithmic scaling rather than the linear model or power function based on the rigorous biological experiments. Thus, the benefit of the proposed method is that the adopted Weber-Fechner Law is more suitable to describe human reflection against the numerical stimulation (like price) than linear models since it fits the human natural cognitive mode. The Weber-Fechner Law has been successfully applied for measuring the human reflection against the price incentives in similar cases, such as dynamic pricing demand response of regenerative electric heating [27], EV user charging decision prediction [28].

It is assumed that each EV charging post has the same parameters. The user's expected unit profits for V2G service can be formulated as:

$$\Pi_{n,i,t}^{EV,Unit} = w_{n,i,t}^{D} - w_{n,i,t_a}^{T} - w_{n,i,t}^{L}. \tag{48}$$

The expected average unit profits can be calculated as:

$$\overline{\Pi}_{n,t}^{EV,Unit} = w_{n,t}^{D} - \overline{w}_{n,t}^{T} - \overline{w}_{n,t}^{L} \tag{49}$$

$$\bar{w}_{n,t}^{T} = \sum_{i}^{N_{n,t}} w_{n,i,t_a}^{T} / N_{n,t} \tag{50}$$

$$\bar{w}_{n,t}^{L} = \sum_{i}^{N_{n,t}} w_{n,i,t}^{L} / N_{n,t} \tag{51}$$

where $\bar{w}_{n,t}^{T}$ is the average total charging price and $\bar{w}_{n,t}^{L}$ is the average unit battery loss price. Because $\bar{w}_{n,t}^{L}$ is generally very small, the common EV model's battery loss price w^{L} can be taken as its value.

In this study, the stimulus intensity I is the average unit profits $\bar{\Pi}_{n,t}^{EV,Unit}$, the EV user's sensory intensity s represents the average willingness $\bar{W}_{n,t}^{D}$ of users for participating in V2G service.

According to the Weber-Fechner Law, we can obtain the relationship between the user profits and the discharging willingness as follows:

$$\bar{W}_{n,t}^{D}(\bar{\Pi}_{n,t}^{EV,Unit}) = \begin{cases} 0, & \bar{\Pi}_{n,t}^{EV,Unit} \le \bar{\Pi}_{n,t}^{Lower} \\ k_{n,t}\log(\bar{\Pi}_{n,t}^{EV,Unit} / \bar{\Pi}_{n,t}^{Upper}) + s_{0n,t}, & \bar{\Pi}_{n,t}^{Lower} < \bar{\Pi}_{i,t}^{EV,Unit} \le \bar{\Pi}_{n,t}^{Upper} \\ s_{0n,t}, & \bar{\Pi}_{n,t}^{EV,Unit} > \bar{\Pi}_{n,t}^{Upper} \end{cases} \tag{52}$$

$$\bar{\Pi}_{n,t}^{Lower} / \bar{\Pi}_{n,t}^{Upper} = 10^{-s_{0n,t}/k_{n,t}} \tag{53}$$

$$0 < k_{n,t}, 0 < s_{0n,t} < 1 \tag{54}$$

where $\bar{\Pi}_{n,t}^{Lower} / \bar{\Pi}_{n,t}^{Upper}$ is the lower/upper bounds of effective profits incentive range. (53) shows the boundary conditions where $s_{0n,t}$ is the maximum willingness and $k_{n,t}$ is the coefficient reflecting user sensitivity to average profits. When $\bar{\Pi}_{n,t}^{EV,Unit} \le \bar{\Pi}_{n,t}^{Lower}$, the average unit profits fall in the dead zone so that the users would not participate in the V2G service. With the increase of $\bar{\Pi}_{n,t}^{EV,Unit}$, the willingness strengthens gradually until it is saturated at $s_{0n,t}$.

E. Objective Function

We formulate the EV management problem as follows:

$$\begin{aligned} &\text{DSO}: && \max_{w_{n,t}^{F}} \sum_{t\in\boldsymbol{\Gamma}} \Pi_{t}^{DSO} \\ &&& s.t.\ \text{Voltage constraint: (5)} \\ &&& \quad \text{Peak shaving constraint: (7)} \\ &&& \quad \text{Grid constraints: (1)-(4),(6)} \\ &\text{CS aggregator}: && \max_{w_{n,t}^{T}, w_{n,t}^{D}, N_{n,t}^{V2g}, N_{n,t}^{rpc}} \sum_{t\in\boldsymbol{\Gamma}} \sum_{n\in\boldsymbol{B}} \Pi_{n,t}^{CSs} \\ &&& s.t.\ \text{Voltage constraint: (5)} \\ &&& \quad \text{EV charging constraints: (8)-(15)} \\ &&& \quad \text{Charging demand constraint: (16)} \\ &&& \quad \text{Willing constraints: (17)-(27), (48)-(54)} \end{aligned} \tag{55}$$

The above objective function is difficult to solve by using an optimization algorithm because of the strong nonlinearity of the complex EV charging process and willingness evaluation models. Moreover, applying the optimization algorithm to solve the problem will be time-consuming and not suitable for fast response. In view of this, a data-efficient bi-level SAC approach is proposed to solve the real-time optimal control problem.

Fig. 2 illustrates the detailed DRL-based real-time peak shaving and voltage regulation scheme, where the dashed/solid lines denote the communication/ electrical link. The DSO firstly gives the incentive price to the CS aggregator according to its online policy of DRL-based controller, then the controller of CS aggregator sets the charging price to adjust the number of arrival EVs and sets the discharging price to incentive users to participate in V2G service. For allocating charging/discharging power to each EV at a station, as shown in Fig, 2, the CS uses K-means clustering method to classify EVs into m groups based on the energy availability criterion. Thereafter, the CS sequentially selects the EVs in cluster EV_1, EV_2, EV_k until the action number is reached, i.e., EVs with relatively more slack time and less energy to be charged, will be chosen to participate in V2G mode. For EVs in V2G mode, the discharging power is set as a constant. For the other EVs (including EVs in RPC mode), they charge normally as the charging trajectory shown in Figure 1. The DSO and CS aggregator shares the current state and mutual actions, and their controllers are pre-trained in the offline stage using mixed data through a Gaussian mixed method (GMM)-based simulator and fast-trained in the online stage with real-time data.

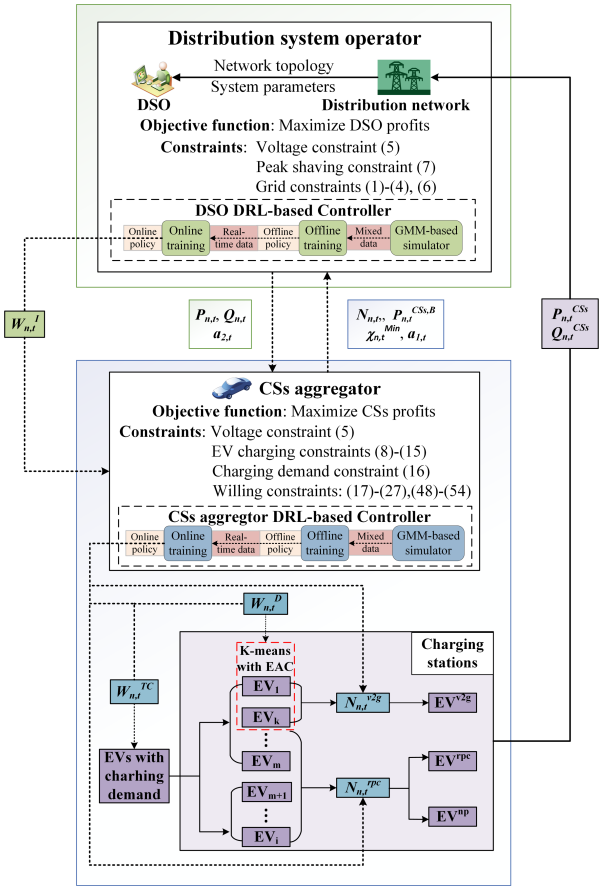

Fig. 2 DRL-Based real-time peak shaving and voltage regulation scheme.

III. Data-Efficient Bi-Level SAC Approach

For the proposed problem in the last section, there are three salient requirements for the solution:1) the approach should take into account the interests of both the DSO and CSs, corresponding to the mathematical formulation; 2) collecting a large number of training data is expensive, the algorithm should be data-efficient; 3) the methods should balance the violations and explorations,

and avoid the drastic violations in the start-up process. Safe DRL seems to meet this requirement, but in practice secure DRLs are not appropriate for this situation. The constraints in CMDP (constrained Markov decision process) for safe reinforcement learning algorithms are generally considered as the hard constraints. That is, the policy should absolutely not violate the constraint or just violate the constraint with a very small probability. Nevertheless, the suitable method for this study should meet the soft constraint. For example, suppose that the cost function in safe DRL is just the SOC constraint and we set the threshold of cost function as 3 (represents the maximum extra for the EV user is 3 min). Now a likely condition is that if the EV user waits 3.1 minutes, the CS will not pay the penalty for not reaching 70% bidding volume. The safe DRL agent absolutely sacrifices the CS profits to meet the SOC constraints, however, a little bit of additional waiting time is acceptable. For DRL, we can avoid such problem by balancing the factor of each item in reward function. When the SOC cost increases a little and profit reward increases a lot, the total reward function still increases. And thus, we can get a moderate and suitable policy; 4) due to the uncertainty of EV status, it is challenging to satisfy the charging demand constraint. The key of the problem is how to measure the EV scheduling potential.

To solve the all issues above, we develop an EV energy availability criterion for the K-means clustering method and propose a data-enhanced bi-level SAC approach in this section.

A. K-means with EV Energy Availability Criterion

Traditional methods usually use the real-time SOC and departure/parking time as the criteria for clustering algorithms to select EVs for V2G service, which does not reflect well the scheduling priority of EVs. For EV with high battery capacities, a high SOC does not mean short charging time, and a short parking time may not indicate the urgent charging tasks for EVs with small capacity. To solve this issue, an EV energy availability criterion for clustering algorithm is developed as follows:

$$E_{n,i,t}^{EV,R} = B_{n,i}^{EV}(SOC_{n,i}^{EV,D} - SOC_{n,i,t}^{EV}) \tag{56}$$

$$\chi_{n,i,t} = t_{n,i}^{l} - t_{n,i}^{a} - \frac{E_{n,i,t}^{EV,R}}{P_{n,i,t}^{EV,C}} \tag{57}$$

where $E_{n,i,t}^{EV,R}$ and $\chi_{n,i,t}$ are the remaining charging energy and slack time, respectively. $SOC_{n,i}^{EV,D}$ is the expected SOC of EV user at departure. The reason why this paper doesn't classify EVs based on only the slack time is to reduce the risk of the user's trip uncertainty, including the user's behavior of premature departure and the fluctuation of other loads. The minimization objective function of the K-means algorithm is written as:

$$J = \sum_{k=1}^{K_{n,t}} \sum_{i=1}^{N_{n,t}} \left\| x_i(E^{EV,R}, \chi) - u_k(E^{EV,R}, \chi) \right\|^2 \tag{58}$$

$$u_k(E^{EV,R}, \chi) = \frac{1}{N_{n,t}} \sum_{i=1}^{N_{n,t}} x_i(E^{EV,R}, \chi) \tag{59}$$

where x_i and u_k denotes the samples and clustering centers. $K_{n,t}$ is the number of clustering centers.

B. Data-Efficient Bi-Level SAC Algorithm

1) *Bi-Level Soft Actor Critic Algorithm*

The MDP tuple in multi-agent RL can be defined as a tuple (S, A_i, T, r_i, γ), where S denotes the state space, A denotes the joint action space while A_i denotes the action space of agent i, r_i denotes the reward function of agent i with the discount factor $\gamma \in (0,1)$. $T: S \times A \to D(S)$ is the transaction function. A policy $\pi: S \to D(A)$ is a mapping from state to distribution over actions.

The purpose of SAC-based agents is to maximize the cumulative reward with entropy [29]. Corresponding to the bi-level optimization (55), bi-level SAC algorithm can be formulated as follows:

$$\begin{aligned} &\max_{\pi_1} \mathop{\mathrm{E}}_{(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)\sim(\rho_{\pi_1}, \rho_{\pi 2})} \sum_{t=0}^{T} \gamma^t [r_1(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) + \alpha_1^T \mathrm{H}(\pi_1(\cdot \mid \boldsymbol{s}_t, \boldsymbol{a}_{2,t}^{'}))] \\ &\text{s.t.} \quad \pi_1 \in \Pi_1 \\ &\quad \max_{\pi_2} \mathop{\mathrm{E}}_{(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)\sim(\rho_{\pi_1}, \rho_{\pi 2})} \sum_{t=0}^{T} \gamma^t [r_2(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) + \alpha_2^T \mathrm{H}(\pi_2(\cdot \mid \boldsymbol{s}_t, \boldsymbol{a}_{1,t}))] \\ &\quad \text{s.t.} \quad \pi_2 \in \Pi_2 \end{aligned} \tag{60}$$

where $\vec{\boldsymbol{a}}_t = (\boldsymbol{a}_{1,t}, \boldsymbol{a}_{2,t})$ is the joint vector of actions. $\rho_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$ denotes the state-action marginal trajectory distribution caused by $\pi(\boldsymbol{a}_t|\boldsymbol{s}_t)$, α_i^T is the temperature coefficient that determines the importance of entropy $\mathcal{H}$ to reward.

The Bellman function that denotes the iterative relationship between the soft state-values $V_i(\boldsymbol{s}_t)$ and soft action-values $Q_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)$ is given as follows:

$$Q_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) = r_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) + \gamma \mathrm{E}_{\boldsymbol{s}_{t+1}\sim p}[V_i(\boldsymbol{s}_{t+1})] \tag{61}$$

$$V_i(\boldsymbol{s}_t) = \mathrm{E}_{\boldsymbol{a}_{i,t}\sim\pi_i}\left[Q_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) - \alpha_i^T \log \pi_i(\boldsymbol{a}_{i,t} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) \right] \quad i = 1,2 \tag{62}$$

where $\widetilde{\boldsymbol{a}_t}$ denotes the action of the other agent. Agent i simultaneously uses a policy network/two critic networks/two critic networks to learn the policy π_{φ_i}/Q-function Q_{θ_i}/target Q-function $Q_{\bar{\theta}_i}$ with the parameter set $\varphi_i/\theta_i/\bar{\theta}_i$.

The critic network takes the state and joint action as its input and the soft Q-function parameters can be learned by minimizing the Bellman residual as:

$$\begin{aligned} J_Q(\theta_i) = \mathrm{E}_{(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)\sim \mathrm{D}} [\frac{1}{2}(Q_{\theta_i}(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) - \\ (r_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) + \gamma \mathrm{E}_{\boldsymbol{s}_{t+1}\sim p}[V_{\bar{\theta}_i}(\boldsymbol{s}_{t+1})]))^2] \quad i = 1,2 \end{aligned} \tag{63}$$

where $\mathcal{D}$ is the experience replay buffer. The parameters $\bar{\theta}$ of Q-function are obtained as an exponential moving average of the soft Q-function weights:

$$\begin{aligned} \nabla_{\theta_i} J_Q(\theta_i) &= \nabla_{\theta_i} Q_{\theta_i}(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)(Q_{\theta_i}(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) - (r_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) \\ &+ \gamma(Q_{\bar{\theta}_i}(\boldsymbol{s}_{t+1}, \vec{\boldsymbol{a}}_{t+1}) - \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t+1} \mid \boldsymbol{s}_{t+1}, \tilde{\boldsymbol{a}}_{t+1}))) \quad i = 1,2 \end{aligned} \tag{64}$$

The policy network takes the state and the opponent's action as its input. The policy parameters can be learned by minimizing the Kullback-Leibler divergence as:

$$J_\pi(\varphi_i) = \mathrm{E}_{(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)\sim \mathrm{D}} \left[\mathrm{E}_{\boldsymbol{a}_{i,t}\sim\pi_i} \left[\alpha_i^T \log \pi_i(\boldsymbol{a}_{i,t} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) - Q_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) \right] \right] \quad i = 1,2 \tag{65}$$

We use the reparameterization trick to optimize φ_i as follows:

$$\begin{aligned} \nabla_\varphi J_{\pi_i}(\varphi_i) &= \nabla_{\varphi_i} \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)) \\ &+ (\nabla_{\boldsymbol{a}_{i,t}} \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)) \\ &- \nabla_{\boldsymbol{a}_{i,t}} Q(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)) \nabla_{\varphi_i} f_{\varphi_i}(\varepsilon_{i,t}; \boldsymbol{s}_t, \boldsymbol{a}_t) \quad i = 1,2 \end{aligned} \tag{66}$$

where $\varepsilon_{i,t}$ is an input noise vector which is sampled from a Gaussian distribution in this study.

Then we have the following bi-level updating rules:

$$\boldsymbol{a}'_{2,t+1} = \pi_{\varphi_2}(\cdot \mid \boldsymbol{s}_{t+1}, \boldsymbol{a}_{1,t}) \tag{67}$$

$$\boldsymbol{a}_{1,t+1} = \pi_{\varphi_1}(\cdot \mid \boldsymbol{s}_{t+1}, \boldsymbol{a}'_{2,t+1}) \tag{68}$$

$$\boldsymbol{a}_{2,t+1} = \pi_{\varphi_2}(\cdot \mid \boldsymbol{s}_{t+1}, \boldsymbol{a}_{1,t+1}) \tag{69}$$

$$\theta_i = \theta_i - \beta_Q \nabla_{\theta_i} J_Q(\theta_i) \quad i = 1,2 \tag{70}$$

$$\bar{\theta}_i = \tau\theta_i + (1-\tau)\bar{\theta}_i \quad i = 1,2 \tag{71}$$

$$\varphi_i = \varphi_i - \beta_{\pi_i} \nabla_{\varphi} J_{\pi_i}(\varphi_i) \quad i = 1,2 \tag{72}$$

$$\alpha_i^{T^*} = \arg\min_{\alpha_i^T} \mathrm{E}_{\boldsymbol{a}_{i,t}\sim\pi_i}\left[-\alpha_i^T \log \pi_i(\boldsymbol{a}_{i,t} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) - \alpha_i^T \overline{\mathrm{H}}_i\right] \quad i = 1,2. \tag{73}$$

The DSO (upper level) first anticipates the CSs aggregator's actions confronting the new state, and then sample the actions based on the predicted opponent action and new state according to (67) and (68). After that, CS aggregator samples its actions based on the identified DSO's actions and the new state according to (69). The weights of critic networks are updated by having them slowly track the learned networks as (70) and (71), while the parameters of actor network is directly calculated by gradient descent method as (72). The adjusted temperature hyperparameter is calculated as (73) where $\bar{\mathcal{H}}$ is the desired minimum expected entropy.

To ensure the practical application of the proposed method, we prove that the proposed algorithm converges to the optimal policy, which is attached in the Appendix. And the detailed bi-level SAC algorithm is presented as follows:

Algorithm 1: Proposed Bi-level SAC Algorithm

Input: $\theta_1^{\mathrm{I}}, \theta_1^{\mathrm{II}}, \bar{\theta}_1^{\mathrm{I}}, \bar{\theta}_1^{\mathrm{II}} / \theta_2^{\mathrm{I}}, \theta_2^{\mathrm{II}}, \bar{\theta}_2^{\mathrm{I}}, \bar{\theta}_2^{\mathrm{II}}$ -initial parameters of four critic networks of leader/follower; φ_1 / φ_2 -initial parameters of the policy network of leader/follower; $\mathrm{D} \leftarrow \varnothing$ -initialize an empty replay buffer; γ -discount factor; $\beta_{\varphi_1}, \beta_{Q_1}, \beta_{\alpha_1} / \beta_{\varphi_2}, \beta_{Q_2}, \beta_{\alpha_2}$ -the learning rate of the policy network, the critic networks, temperature coefficient of upper-level agent/ lower-level agent.

for each iteration **do**

for each environment step **do**

Upper-level agent estimates the policy of lower-level confronting the new state $\pi_{\varphi_2}(\cdot \mid \boldsymbol{s}_{t+1}, \boldsymbol{a}_{1,t})$ and then samples the actions based on (67) and (68).

Lower-level agent samples the action in face of the current status and upper-level agent's action based on (69).

Sample transition from the environment $\boldsymbol{s}_{t+1} \sim p(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)$

Store the transition in replay buffer $\mathrm{D} \leftarrow \mathrm{D} \cup \{(\boldsymbol{s}_t, \boldsymbol{a}_{1,t}, \boldsymbol{a}_{2,t}, \boldsymbol{s}_{t+1}, \boldsymbol{r}_{1,t+1}, \boldsymbol{r}_{2,t+1})\}$

end for

for each gradient step **do**

Sequentially update critic parameters for the two agents based on (63), (64), (70).

Sequentially update policy weights for the two agents based on (65), (66), (72).

Sequentially update critic network weights for target Q-function based on (63), (64) (70), (71).

Sequentially adjust temperature for the two agents based on (73).

end for

end for

Output: $\theta_1^{\mathrm{I}}, \theta_1^{\mathrm{II}}, \bar{\theta}_1^{\mathrm{I}}, \bar{\theta}_1^{\mathrm{II}} / \theta_2^{\mathrm{I}}, \theta_2^{\mathrm{II}}, \bar{\theta}_2^{\mathrm{I}}, \bar{\theta}_2^{\mathrm{II}}$, φ_1 / φ_2

2) *GMM-Based Data-Enhanced Method*

To generate data that reflects the regular characteristics of EVs, GMM is used to fit the distribution of each EV user charging behavior feature. GMM can decompose the features into several Gaussian-based probability density functions. The formulation of GMM model is presented as:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k) \tag{74}$$

where x is the feature, $N(x|\mu_k, \Sigma_k)$ is the kth component of the mixed model with weight π_k. Because GMM is an unsupervised learning method, its optimal cluster number cannot be set automatically. The silhouette coefficient is calculated for each feature to choose the optimal cluster number, which is given as:

$$S(i) = \frac{a(i) - b(i)}{\max\{a(i), b(i)\}} \tag{75}$$

where $a(i)$ is the average distance between sample i and all the other points in the same cluster. $b(i)$ is the average distance between the sample and all the other points in the next nearest cluster. The value range of silhouette coefficient is [-1,1]. The higher silhouette coefficient, the better cluster performance.

Battery capacity, charging duration and start-end SOC are chosen as the three key features. The average silhouette coefficients under different clusters of historical data of CSs in a certain area, Shenzhen in 2021 are shown in Fig. 3. The results indicate that the optimal cluster number for each key feature is 2. The GMM model parameters are shown in the Table I. Based on the fitting results in Table I, we build an EV simulator that can generate 8 (2x2x2) feature data distributions. The clustering results of real/simulation data t are shown in Fig 4.

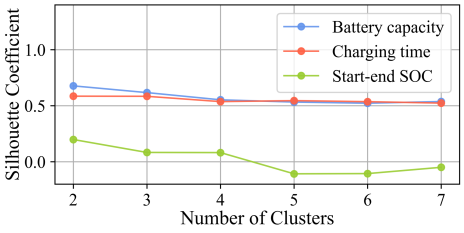

Fig. 3 Average silhouette coefficient of features

TABLE I GMM MODEL PARAMETERS

Features	Mean value (μ)	Variance (σ^2)
Battery capacity (kWh)	{[53.13], [85.35]}	{[79.01], [200.01]}
Charging time (min)	{[66.65], [36.47]}	{[281.69], [170.88]}
Start-end SOC	{[51.86→98.62], [43.32 → 81.69]}	{[[369.62, −0.01], [−0.01,0.60]], [[390.77,146.56], [146.56,258.27]]}

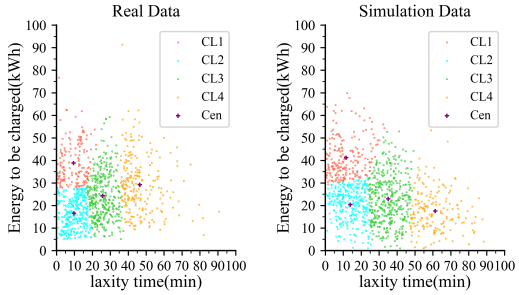

Fig. 4 Clustering results based on EV energy available criterion

3) *Improved Two-Staged Training Process*

To improve the policy robustness against to the uncertainty of

the user's trip and avoid the danger caused by the instability of exploration in the early stage of training, a simulator is established to execute a two-staged training process, but it will incur two additional problems: how to ensure the validity of the generated data for training, and to avoid the over centralized distribution and constant proportion of the training data sets.

To solve the issues above, an improved two-staged training method for SAC algorithm is proposed as shown in Fig. 5. To ensure the validity of simulation data, GMM method is adopted and the real data is divided into three sets for offline training, online training and test process, respectively. The offline training dataset consists of a 1:1 mixture of real and simulation data. In the online training process, only the real data is used to achieve a rapid training. After the two-staged training process, untrained real data is applied to test the proposed method. To improve the robustness of the model, isotropic truncated Gaussian noise is added for each feature in 20% of the simulated data and replace data of certain types in random proportion to avoid over-concentration of the distribution. The offline training is completed far before the implement. After that, the policy will be continuously updated online at each time stage synchronized with the grid's 15-minute operating interval.

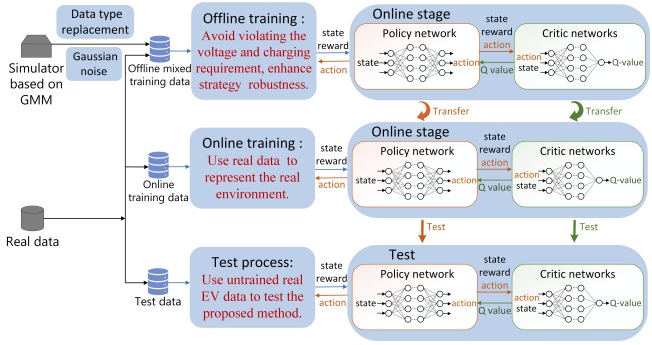

Fig. 5 Improved two-staged training process

C. Formulate Ancillary Service Provision Problem as an Markov Decision Process

In this section, the peak shaving and voltage regulation problem is formulated as an MDP.

1) *Definition of State*

s_t is the observed status information of CSs and buses during time stage t, including the active/reactive power $P_{n,t}$/ $Q_{n,t}$, the total EV number $N_{n,t}$, the CSs bidding volume $P_{n,t}^{CSs,B}$ and the minimum slack time $\chi_{n,i,t}^{Min}$.

2) *Definition of Action*

For DSO, $a_{1,t}$ is the incentive price $w_{n,t}^{I}$; for CS aggregator, $a_{2,t}$ includes the total charging price $w_{n,t}^{T}$, the discharging price $w_{n,t}^{D}$, the number of EVs $N_{n,t}^{V2G}$ and $N_{n,t}^{RPC}$ in V2G and RPC modes. The values of the action variables are pre-constrained according to prior knowledge.

3) *Definition of Reward*

Agent 1 (DSO):

$$r_{1,t} = \xi_1 \Pi_t^{DSO} + \xi_2 p_t + \xi_3 u_t \tag{76}$$

$$p_t = \begin{cases} \sum_{n\in B}(P_{n,t} - P_n^{MAX}), & P_{n,t} > P_n^{MAX} \\ 0, & other \end{cases} \tag{77}$$

$$u_t = \begin{cases} \sum_{n\in B}(0.95pu - V_{n,t}), & V_{n,t} < 0.95pu \\ r_u, & V_{n,t} \in [0.95,1.05]pu \\ \sum_{n\in B}(V_{n,t} - 1.05pu), & V_{n,t} > 1.05pu \end{cases} \tag{78}$$

Agent 2 (CS aggregator):

$$r_{1,t} = \xi_4 \sum_{n\in B} \Pi_{n,t}^{CSs} + \xi_5 u_t + \xi_6 d_t \tag{79}$$

$$d_t = \begin{cases} \chi_{n,i,t}^{Min}, & \chi_{n,i,t}^{Min} < 0 \\ 0, & \chi_{n,i,t}^{Min} \geq 0 \end{cases} \tag{80}$$

where (77), (78) and (80) denote the violation errors of peak shaving, voltage and user's charging demand, respectively. ξ_1 to ξ_6 are the positive weight factors. We typically set negative constraint term (e.g., $\xi_3 u_t$) to be 5-10 times the positive target term (e.g., $\xi_3 \Pi_t^{DSO}$). And to make it easier for the neuron networks to calculate the gradient quicker, we suggest adjusting ξ_1 to ξ_6 to make the reward values of both agents do not exceed 200.

IV. Case Study

A. Test System and Parameter Settings

The effectiveness of the proposed scheme is validated on the modified IEEE 33-bus distribution network. All the tests are conducted using Python on a computer with 4.4 GHz CPU, 1.6 GHz GPU and 16 GB RAM. Pytorch is applied to formulate the neural networks framework of DRL algorithm and Pandapower to establish the modified IEEE 33-bus distribution network.

Three CS aggregators are set at bus 13, 16 and 31, respectively. The detailed parameters for simulation and algorithm are presented in Table II and Table III. The results of willingness models are shown in Fig. 6. The peak time of urban load profile usually occurs between 11:30 and 12:00, and drops rapidly after 12:00 while the peak time of EV charging during the day is from 12:30-1:00. The peak time of grid power consumption is set to 11:45 based on the real situation and eight operations with an interval of 15 minutes from 10:00 in the morning are carried out. The information about the number of EVs with charging demand and bidding volume of CS aggregators is shown in Fig.7. All the data used in the simulation can be accessed from [30].

TABLE II Simulation Parameters

Parameters	Value		
	Bus13	Bus16	Bus31
$P_{n,i,t}^{R}/P_{n,i,t}^{D}(kW)$	60, 4.05		
ω_i / a_i / b_i / c_i	5/12/13.4/1		
B_i^1 / B_i^2 / B_i^3 (μ,σ^2)	(0.6,0.4)/0.8,0.2)/(0.4,0.2)		
D^1 / D^2 / D^3 (μ,σ^2)	(0.33,0.1)/ (0.33,0.1)/ 1- D^1 - D^2	(0.3,0.05)/ (0.6,0.05)/ 1- D^1 - D^2	(0.6,0.1)/ (0.2,0.1)/ 1- D^1 - D^2
$w_{n,t}^{E}$ / $w_{n,t}^{P}$ / $w_{n,t}^{G}$ (¥)	0.77/0.77/0.45	0.85/0.77/0.45	0.74/0.77/0.45
Charging pile number	700	300	1000

TABLE III ALGORITHM PARAMETERS

Parameters	Value
Optimizer	Adam
Number of hidden layers (All networks)	2
Number of hidden units per layer of policy network	64/512
Number of hidden units per layer of critic networks	256/16
Learning rate of actor network/critic network/temperature coefficient	4e-4/8e-4/8e-4
Discount factor	0.99
Replay buffer size	1e6
Number of samples per mini batch	256
Nonlinearity	Leaky-ReLU

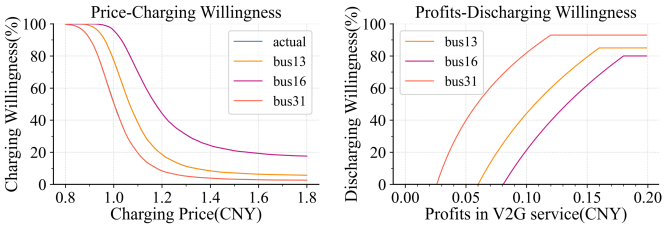

Fig. 6 The relationship between Price/Profit-charging/discharging willingness

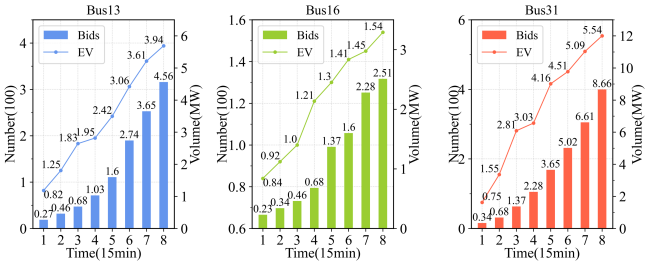

Fig. 7 Number of EVs with charging demand and bidding volume of CSs

B. Results and Analysis

The training results of the two-stage training process are shown in Fig.8, in which the lighter line represents the average value per 50 steps while the darker line represents the actual value of each training step. It can be indicated from Fig. 8 that both the number of training iterations for convergence and deviation of system reward are reduced significantly in the online stage compared with the offline stage.

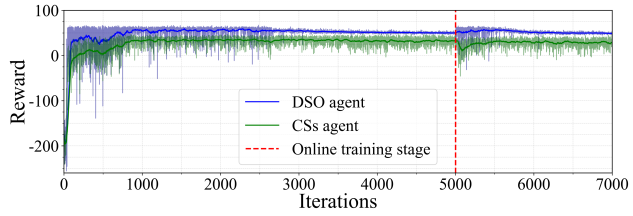

Fig. 8 Reward of training process

The detailed performance comparisons during the start-up process between the conventional single-stage DRL method and the proposed two stage method is presented in Table IV. The single-stage DRL method is the SAC algorithm with the same parameters as the two-stage algorithm. The reason why we only compare the online training time to single-stage method is the offline training can be completed far before the implement. The results in Table IV show that the proposed data-efficient method eliminates the start-up voltage violation, and reduces the average/maximum peak shaving errors, average/maximum number of EVs with unmet charging demand, average/maximum CSs loss by 81.06%/77.72%, 69.63%/53.35%, 4.18%/14.02% compared to the single-stage method. The prior knowledge learned from the GMM-based simulator with mixed data during offline stage can help the agent to converge to the local optimal solution faster and perform better against the uncertainty at the beginning of the online stage.

TABLE IV COMPARISONS OF ONLINE START-UP PROCESS WITH SINGLE-STAGE DRL METHOD OVER THE FIRST 200 STEPS USING REAL DATA

Items	Single-stage method	Two-stage method (online process)
Average reward of agent DSO/CSs	0.92/-37.71	46.67/38.92
Average/Maximum voltage violation (p.u.)	2.97e-5/1.04e-3	0/0
Average/Maximum cumulative peak shaving violation (MW)	3.96/36.62	0.75/8.16
Average/Maximum number of EVs with unmet charging demands	281.82/1029	85.60/480
Average/Minimum CSs total profits (￥)	16501.87/ 13218.43	17190.85/ 15072.16
Time to convergence (min)	43.67	9.34
Loading time for the execution (s)	9.05	9.05

All the following results are obtained by loading the trained model on the test set. The actions of the algorithm are shown in Fig. 9 and Fig. 10. The yellow shadow in Fig. 9 denotes the V2G profits of CSs while the other shadow reflects the V2G profits of EV users. The discharging price must be higher than the average charging price, but can be lower than the discharging price, which means that the arrival EVs do not participate in the dispatch for peak shaving. Fig 10 presents the different EVs scheduling strategies of three buses. It can be observed that CSs controller schedules more EVs into V2G and RPC modes to shave the peak load and regulate the voltage.

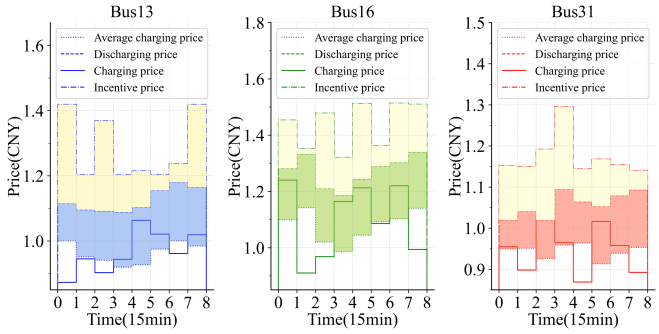

Fig. 9 Prices of CSs at three buses

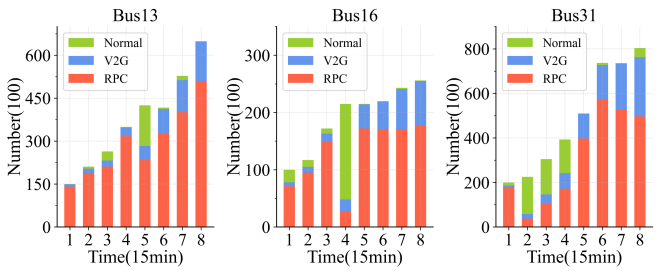

Fig. 10 EVs scheduling status of CSs at three buses

For peak shaving, the simulation results are shown in Fig 11 to Fig 15. Fig 11 presents the peak shaving result and Fig 12 shows

the active/reactive power of CSs. The proportions of peak shaving amount by V2G service in power load of three buses are 3.17%, 4.01% and 3.51%, respectively. The reward function of control group in Fig. 13 and 14 considers voltage constraints and agent profits. According to the proposed CSs profit model in section II, the CSs can obtain high profits under the designed scheme in a condition of reasonable electricity price and bidding quantities. Fig. 13 shows the total CSs profits after the completion of charging for all EVs arriving at CSs during the peak shaving process. The CSs profits at bus 13, 16 and 31 are improved by 71.66%, 1.06%, 20.01% and the total profits are improved by 29.6%. The corresponding results of DSO profits are given in Fig. 14, where the red shadow is the cost of V2G incentives. The DSO profits at bus 13 and 31 are improved by 193.92%, 64.51% while the DSO profits at bus 16 have decreased by 23.00%, the total DSO profits are improved by 151.27%. The results indicate that the proposed scheme increases the maximum EV number that can be charged simultaneously and thus, profits of both DSO and CSs increase significantly. However, it should be noticed that the scheme is not suitable for areas with concentrations of price-insensitive EV users.

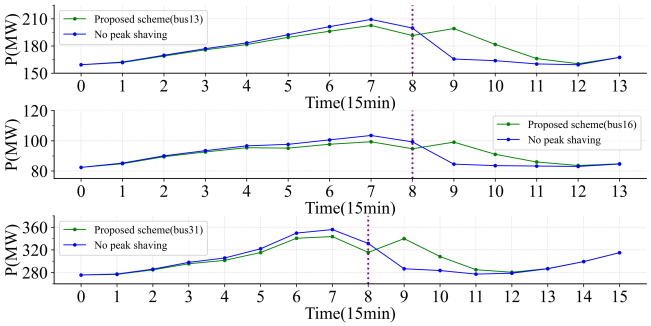

Fig. 11 Active power of three buses during peak shaving

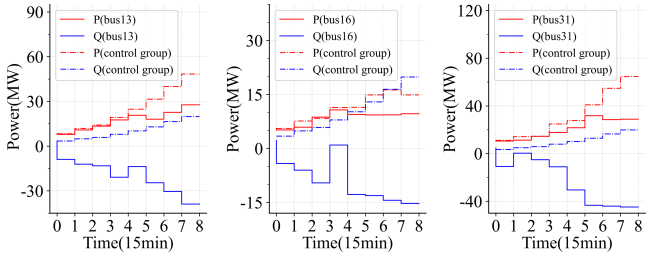

Fig. 12 Active/Reactive power of CSs at three buses

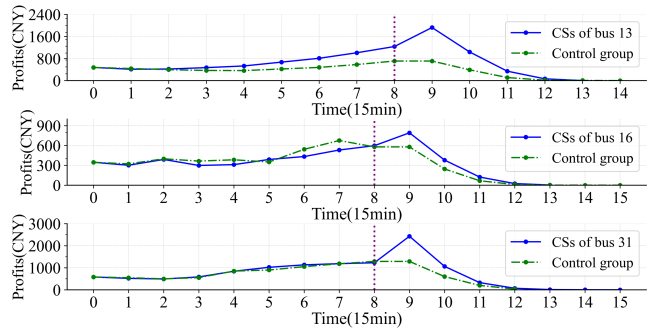

Fig. 13 Profits of CSs at three buses

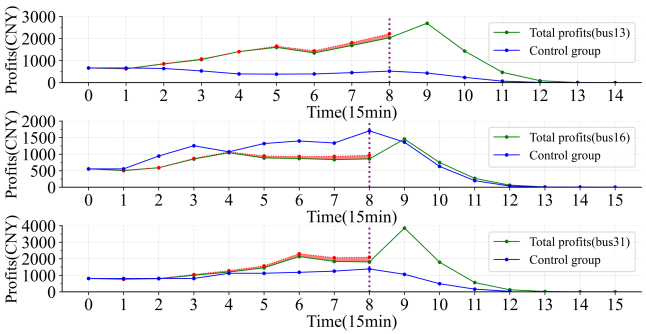

Fig. 14 Profits of DSO at three buses.

Fig 15 presents the change of EVs minimum slack time. The detailed information of EVs with unmet charging demand is shown in Table V. Compared with the method in [4], the proposed minimum slack time constraint ensures the benefits of EV users, reducing the percentage of non-compliant charging EVs from 27.04% to 0.089%, with an average extra waiting time of 1.00 minutes and maximum extra waiting time of 3.12 minutes.

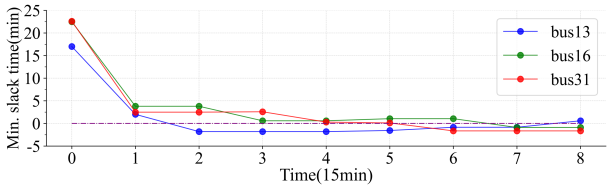

Fig. 15 Minimum slack time of EVs at three buses.

TABLE V INFORMATION OF CHARGING REQUIREMENT VIOLATION EVS

Items	Value			
	Bus13	Bus16	Bus31	Total
Number of unsatisfied EVs	0	1	3	4
Extra waiting time (min)	0	0.89	3.12	2.28
Average/maximum extra waiting time (min)	0/0	0.89/0.89	1.04/1.65	1.00/1.65

The performance of the proposed method in terms of voltage regulation is validated by simulation results in Fig 16 and 17, which present the voltage profile of all buses and the voltage variation at buses 13, 16, 31, respectively. The reward function of control group in Fig. 17 only considers the profits of the two agents. However, the effect of voltage adjustment is limited because of the constraint (15) and (16), the proposed scheme is more suitable as a regulatory approach for voltage fluctuations caused by the collective charging behavior of EV clusters rather than as a primary voltage control method.

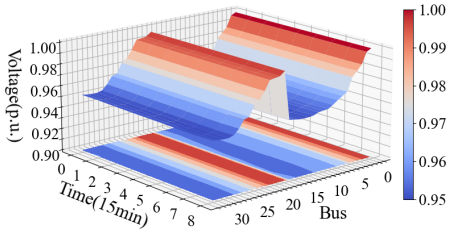

Fig. 16 Voltage profile of the whole 33-IEEE network

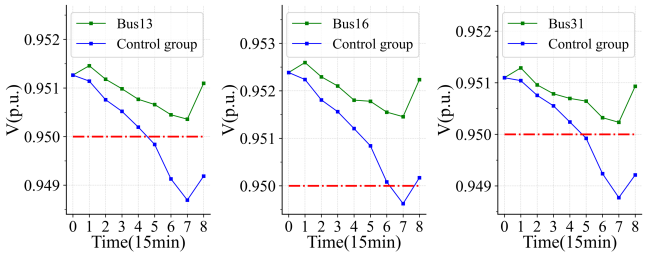

Fig. 17 Voltage variation of three buses

C. Comparisons with different DRL methods.

Proximity policy optimization (PPO) and twin delayed deep deterministic (TD3) policy gradient algorithm are two state-of-the-art DRL-based algorithms. PPO is often used as the benchmark for reinforcement learning-based algorithm in computer science due to its strong robustness, while TD3 is suitable for the case where the optimal policy is close to the boundary value. Comparisons of the detailed results are presented in Table VII. The execution time of all three algorithms is acceptable for implementation. The proposed method outperforms PPO in peak load reduction and satisfying the users' charging demand with much lower training time. Although TD3 has advantages in the peak load reduction and training speed, it overly violates the constraints of voltage and charging demand. Besides, TD3 is very sensitive to its hyper parameters and requires a long time to adjust.

TABLE VII Results of COMPARISON WITH DIFFERENT DRL ALGORITHMS

Items	Proposed method	TD3	PPO
Peak load reduction (%)	3.17~4.01	3.51~4.26	3.28~3.88
Total profits increase of DSO/CSs (%)	151.27/29.6	157.33/31.24	112.91/21.94
Voltage violation (p.u.)	0	1.21e-7	0
Number of unsatisfied EVs	4	25	7
Time to convergence of online stage (min)	9.34	7.66	28.75
Loading time for execution (s)	9.05	7.21	7.92

D. Scalability Verification

The simulations were also performed on the IEEE 118-bus distribution system (Fig. 18) to verify the scalability of the proposed method with similar conclusions. There are seven CS aggregators at nodes 49, 53, 69, 73, 96, 106, 110. The size of the state is 42 and the action size of agent DSO/CSs increases to 7/28. The results of the simulation are presented in Table VI. Compared with that in the IEEE 33-bus system, similar results and conclusions can also be concluded in the IEEE 118-bus distribution system with only 29.66% increase in online training time. Although the loading time increases by 31.27%, it is still fast enough for real-time control. It can be seen that the training time is not directly linearly related to the network size and the proposed approach remains valid for large-scale problems with the same action characteristics. However, the method still needs to be improved in term of training time to accommodate shorter optimization time interval.

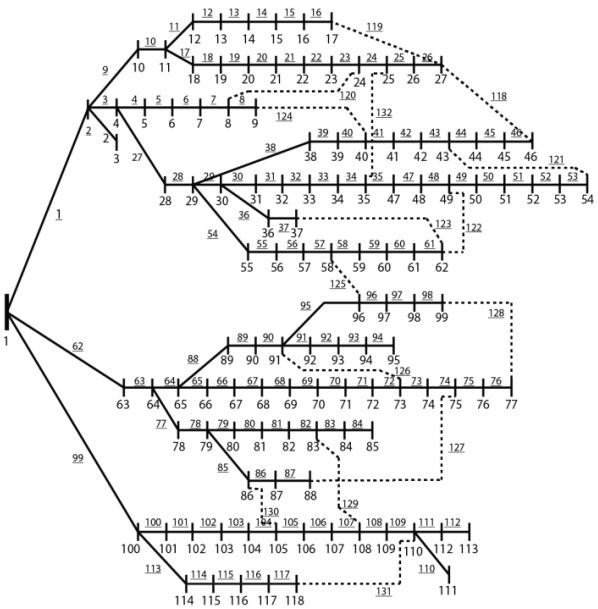

Fig. 18 Illusion of the transportation network and IEEE 118-bus distribution system for scalability verification.

TABLE VI SIMULATION RESULTS ON IEEE 118-BUS DISTRIBUTION NETWORK

Items	Value
Peak load reduction (%)	3.28~4.46
Maximum voltage drop reduction (p.u.)	0.08
Total profits increase of DSO/CSs (%)	143.22/24.7
Total number of EVs with unmet charging requirement	13
Average extra waiting time for charging (min)	1.26
Time to convergence of online stage (min)	12.11
Loading time for the execution (s)	11.88

V. CONCLUSION

This article proposes a data-efficient bi-level DRL-based peak shaving and voltage regulation scheme for EVs in the distribution network. A new approach considering the willingness of EV users and incentive price is proposed and the slack time is defined to establish the EV energy availability criterion. The effectiveness of the proposed scheme has been verified in simulation case studies on the modified IEEE 33-bus and 118-bus distribution networks with real historical data. The following conclusions can be drawn:

1) The proposed bi-level scheme reduces the peak load by 3.17%~ 4.01%, and avoids the voltage violation. Moreover, the CSs total profits and the DSO total profits are increased by 7.07% and 151.27%, respectively.

2) Compared to the single-stage method, the implementation of the GMM-based data-enhanced method eliminates the start-up voltage violation, and reduces the average/maximum peak shaving errors, average/maximum number of EVs with unmet charging demand, average/maximum CSs loss by 81.06%/77.72%, 69.63%/53.35%, 4.18%/14.02%.

3) The proposed energy available criterion and SOC constraint in the reward function ensure the benefits of EV users by reducing the percentage of non-compliant charging EVs from 27.04% to 0.089% compared with the method in [4], with an average extra waiting time of 1.00 minutes and maximum extra waiting time of

1.65 minutes.

4) The scalability of the proposed method has been validated. The training/loading time of the proposed method on the IEEE 118-bus large distribution network increases only 29.66%/31.27% compared to those of the IEEE 33-bus system.

APPENDIX

Proof of the algorithmic convergence to an optimal policy:

Assumption 1: *Define the learning rate $\beta_t(s_t, a_{1,t}, a_{2,t})$ as the inverse of the number of times that the state-action pair $(s_t, a_{1,t}, a_{2,t})$ has been visited for learning. Then κ_t satisfies the following conditions:*

1) $0 \le \beta_t(s_t, a_{1,t}, a_{2,t}) < 1$, $\sum_{t=0}^{\infty} \beta_t(s_t, a_{1,t}, a_{2,t}) = \infty$, $\sum_{t=0}^{\infty} [\beta_t(s_t, a_{1,t}, a_{2,t})]^2 < \infty$, *the latter two hold uniformly and with probability* 1.

2) $\beta_t(s, a_1, a_2) = 0$ *if* $(s, a_1, a_2) \ne (s_t, a_{1,t}, a_{2,t})$, *which means the agent updates through Q-function.*

Our convergence proof is based on *Lemma 1* by Szepesvari and Littman [31] as follows:

Lemma 1: *Assume that* β_t *satisfies Assumption 1 and the mapping* $P^t: \mathbb{Q} \to \mathbb{Q}$ *satisfies the following conditions:*

1) There exists a number $0 < \gamma < 1$ *and a sequence* $\lambda_t \ge 0$ *converging to zero with probability* 1 *such that* $\| P^t Q - P^t Q^* \| \le \gamma \| Q - Q^* \| + \lambda_t$ *for all* $Q \in \mathbb{Q}$ *and*

2) $Q^* = E[P^t Q^*]$,

then the iteration defined by

$$Q_{t+1} = (1 - \beta_t) Q_t + \beta_t [P^t Q_t] \tag{a-1}$$

converges to Q^* *with probability* 1.

First we prove that the convergence point satisfies the *condition 1)* of *Lemma 1*. We give two definitions:

Definition 1: *Let* $Q = (Q_1, Q_2)$, *where* $Q_1 \in \mathbb{Q}_1, Q_2 \in \mathbb{Q}_2$, *and* $\mathbb{Q} = \mathbb{Q}_1 \times \mathbb{Q}_2$. P^t *is a mapping on the complete metric space* $\mathbb{Q} \to \mathbb{Q}$, $P^t Q = (P^t Q_1, P^t Q_2)$, *where*

$$P^t Q_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) = r_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) + \gamma Q_i(\boldsymbol{s}_{t+1}, \overrightarrow{\boldsymbol{a}_{t+1}}) \quad i = 1, 2. \tag{a-2}$$

Definition 2:

$$\begin{aligned} \left\| Q - Q \right\| &\equiv \max_j \max_{\boldsymbol{s}_t} \left\| Q^j(\boldsymbol{s}_t) - Q^j(\boldsymbol{s}_t) \right\|_{(i, \boldsymbol{s}_t)} \\ &\equiv \max_j \max_{\boldsymbol{s}_t} \max_{\overrightarrow{\boldsymbol{a}_t}} \left| Q^j(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) - Q^j(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) \right|. \end{aligned} \tag{a-3}$$

Lemma 2 (Hu and Wellman [32], Lemma 16): $\| P^t Q - P^t \hat{Q} \| \le \gamma \| Q - \hat{Q} \|, \forall Q, \hat{Q} \in \mathbb{Q}$.

According to *Lemma 2*, $\| P^t Q - P^t \hat{Q} \| \le \gamma \| Q - \hat{Q} \| \le \gamma \| Q - Q^* \| + \lambda_t$, $\lambda_t \ge 0$, where P^t is a contraction operator. The proof of *condition 1)* is complete.

Then we prove the algorithm meets the *condition 2)* of *Lemma 1*. when the algorithm converges, the optimal Q-function can be formulated as:

$$\begin{aligned} & Q_i^*(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) \\ &= r_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) + \gamma \sum_{\boldsymbol{s}_{t+1} \in \boldsymbol{S}} p(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) (Q_i^*(\boldsymbol{s}_{t+1}, \overrightarrow{\boldsymbol{a}_{t+1}}) + \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t+1} \mid \boldsymbol{s}_{t+1}, \tilde{\boldsymbol{a}}_{t+1}))) \\ &= \sum_{\boldsymbol{s}_{t+1} \in \boldsymbol{S}} p(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) (r_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) + \gamma (Q_i^*(\boldsymbol{s}_{t+1}, \overrightarrow{\boldsymbol{a}_{t+1}}) + \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t+1} \mid \boldsymbol{s}_{t+1}, \tilde{\boldsymbol{a}}_{t+1})))) \\ &= \mathrm{E}_{\theta_i}[P^t Q_i^*(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t})] \quad i = 1, 2. \end{aligned} \tag{a-4}$$

Then, we have:

$$\begin{aligned} Q^*(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) &= (Q_1^*(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}), Q_2^*(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t})) \\ &= \mathrm{E}_{\theta_i}[P^t (Q_1^*(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}), Q_2^*(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}))] \\ &= \mathrm{E}[P^t Q^*(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t})] \end{aligned} \tag{a-5}$$

so that the proof of *condition 2)* is completed.

Finally, rewrite (71) to the formulation expressed by soft Q-function:

$$\begin{aligned} Q_i(\boldsymbol{s}_{t+1}, \overrightarrow{\boldsymbol{a}_{t+1}}) = (1 - \beta_t) Q_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) + \beta_t (r_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) + \\ \gamma (Q_i^*(\boldsymbol{s}_{t+1}, \overrightarrow{\boldsymbol{a}_{t+1}}) + \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t+1} \mid \boldsymbol{s}_{t+1}, \tilde{\boldsymbol{a}}_{t+1})))) \quad i = 1, 2. \end{aligned} \tag{a-6}$$

(a-6) is consistent with the format of (a-1) and satisfies the *condition 1)* and *2)* in *Lemma 1*, so the Q value of the bi-level algorithm converges to Q^* with probability 1. Then the policy of the actor networks is trained to converge to the optimal distribution under the optimal Q function.

REFERENCES

[1] C. O'Malley, L. Badesa, F. Teng and G. Strbac, "Frequency response from aggregated v2g chargers with uncertain EV connections," *IEEE Trans. Power Syst.*, vol. 38, no. 4, pp. 3543-3556, July 2023.

[2] Y. Huang, "Day-ahead optimal control of PEV battery storage devices taking into account the voltage regulation of the residential power grid," *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 4154–4167, Nov. 2019.

[3] N. I. Nimalsiri, E. L. Ratnam, D. B. Smith, C. P. Mediwaththe and S. K. Halgamuge, "Coordinated charge and discharge scheduling of electric vehicles for load curve shaping," *IEEE Trans. Intel. Transp. Syst.*, vol. 23, no. 7, pp. 7653-7665, July 2022.

[4] X. Sun and J. Qiu, "A customized voltage control strategy for electric vehicles in distribution networks with reinforcement learning method," *IEEE Trans. Ind. Inform.*, vol. 17, no. 10, pp. 6852-6863, Oct. 2021.

[5] H. Liang, Y. Liu, F. Li and Y. Shen, "Dynamic economic/emission dispatch including PEVs for peak shaving and valley filling," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 2880-2890, April 2019.

[6] V. Gupta, R. Kumar and B. K. Panigrahi, "User-willingness-based decentralized EV charging management in multi-aggregator scheduling," *IEEE Trans. Ind. Appl.*, vol. 56, no. 5, pp. 5704-5715, Sept.-Oct. 2020.

[7] X. Xu et al., "Evaluating multi-timescale response capability of EV aggregator considering users' willingness," *IEEE Trans. Ind. Appl.*, vol. 57, no. 4, pp. 3366-3376, July-Aug. 2021.

[8] M. A. Azzouz, M. F. Shaaban, and E. F. El-Saadany, "Real-time optimal voltage regulation for distribution networks incorporating high penetration of PEVs," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3234–3245, Nov. 2015.

[9] M. Restrepo, C. A. Cañizares and M. Kazerani, "Three-stage distribution feeder control considering four-quadrant EV chargers," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3736-3747, July 2018.

[10] G. E. Mejia-Ruiz, R. Cárdenas-Javier, M. R. Arrieta Paternina, et al., "Coordinated optimal volt/var control for distribution networks via D-PMUs and EV chargers by exploiting the eigensystem realization," *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2425-2438, May 2021.

[11] J. Wang et al., "Coordinated electric vehicle charging with reactive power support to distribution grids," *IEEE Trans. Ind. Inform*, vol. 15, no. 1, pp. 54-63, Jan. 2019.

[12] L. Wang, A. Dubey, A. H. Gebremedhin, A. K. Srivastava and N. Schulz, "MPC-based decentralized voltage control in power distribution systems with EV and PV coordination," *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 2908-2919, July 2022.

[13] M. M. Hoque et al., "Transactive coordination of electric vehicles with voltage control in distribution networks," *IEEE Trans. Sustain. Energy*, vol. 13, no. 1, pp. 391-402, Jan. 2022.

[14] D. Cao, W. Hu, J. Zhao, Q. Huang, Z. Chen, and F. Blaabjerg, "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Trans. Power Syst.*, vol. 35, no. 5, pp. 4120–4123, Sep. 2020.

[15] S. Wang et al., "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Trans. Power Syst.*,

vol. 35, no. 6, pp. 4644–4654, Nov. 2020.
[16] H. T. Nguyen and D. -H. Choi, “Three-stage inverter-based peak shaving and volt-var control in active distribution networks using online safe deep reinforcement learning,” *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 3266-3277, July 2022.
[17] H. Li, Z. Wan and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2427-2439, May 2020.
[18] Y. Cao, H. Wang, D. Li and G. Zhang, “Smart online charging algorithm for electric vehicles via customized actor–critic learning,” *IEEE Internet Things J.*, vol. 9, no. 1, pp. 684-694, Jan. 2022.
[19] B. Wang, Y. Li, W. Ming and S. Wang, “Deep reinforcement learning method for demand response management of interruptible load,” *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3146-3155, July 2022.
[20] Wathieu L. “Consumer habituation,” *Management Science*, pp. 587-596, May 2004.
[21] Kucharavy D, De Guio R. “Application of S-shaped curves,” *Procedia Engineering*, vol. 9, pp. 559-572, 2011.
[22] Homburg C, Koschate N, Hoyer W D. “Do satisfied customers really pay more? A study of the relationship between customer satisfaction and willingness to pay,” *Journal of marketing*, vol. 69, no. 2, pp. 84-96, Apr. 2005.
[23] Cotes-Torres A, Muñoz-Gallego P A, Cotes-Torres J M. “S-shape relationship between customer satisfaction and willingness to pay premium prices for high quality cured pork products in Spain.,” *Meat science*, vol. 90, no. 3, pp. 814-818, Apr. 2012.
[24] M. Sufyan et al., “Charge coordination and battery lifecycle analysis of electric vehicles with V2G implementation,” *Electric Power Syst. Research*, vol. 184, pp. 106307, 2020.
[25] M. Zeng, S. Leng, S. Maharjan, S. Gjessing and J. He, “An incentivized auction-based group-selling approach for demand response management in v2g systems,” *IEEE Trans. Ind. Inform.*, vol. 11, no. 6, pp. 1554-1563, 2015.
[26] S. Dehaene, “The neural basis of the Weber–Fechner law: A logarithmic mental number line,” *Trends Cognit. Sci.*, vol. 7, no. 4, pp. 145–147, Apr. 2003.
[27] S. Zhong, X. Wang, J. Zhao, W. Li, H. Li, Y. Wang, S. Deng, and J. Zhu, “Deep reinforcement learning framework for dynamic pricing demand response of regenerative electric heating,’’ *Appl. Energy*, vol. 288, no. 116623, Apr. 2021.
[28] Zheng W, Li J, Shao Z, Lei K, Li J, Xu Z. “Optimal dispatch of hydrogen/electric vehicle charging station based on charging decision prediction,” *Int. J. Hydrogen Energy*, vol. 48, no. 69, pp. 26964-26978, Aug. 2023.
[29] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel et al., “Soft actor-critic algorithms and applications,” *arXiv:1801.01290*, 2018.
[30] Hongrong Yang, Simulation Data, 2023. [Online]. Available: https://github.com/hongrongyang/Paper_data.
[31] C. Szepesvári and M. L. Littman, “A unified analysis of value-functionbased reinforcement-learning algorithms,” Neural Comput., vol. 11, no. 8, pp. 2017–2060, 1999.
[32] J. Hu and M. P. Wellman, “Nash Q-learning for general-sum stochastic games,” J. Mach. Learn. Res., vol. 4, no. 11, pp. 1039–1069, 2003.