# Dynamic Incentive Pricing on Charging Stations for Real-Time Congestion Management in Distribution Network: An Adaptive Model-Based Safe Deep Reinforcement Learning Method

Hongrong Yang, *Student Member, IEEE*, Yinliang Xu, *Senior Member*, *IEEE*, Qinglai Guo, *Senior Member, IEEE*

***Abstract*—This paper concerns the pricing strategy for real-time distribution network congestion management, aiming at maximizing the total social welfare while eliminating congestion. The difficulty lies in dealing with the complex time-varying relationship between price and charging demand in the integrated power system and transportation network, and ensuring the safe application of the proposed method with limited real historical data. To address these challenges, we first model the EV user charging decision-making process to indirectly reflect the unquantifiable price and demand relationship, and then formulate the pricing problem as a bi-level model with a constrained Markov Decision Process (CMDP). After that, we propose a model-based safe DRL framework and develop an adaptive model-based safe deep reinforcement learning (AMSDRL) algorithm to solve the CMDP problem. AMSDRL learns the environment transition model and uses a strict and adaptive cost constraint to offset potential modeling errors. Compared to the state-of-the-art safe DRL methods, AMSDRL can be deployed with security guarantees by training on limited historical data, which is more practical for applications. The numerical results on modified IEEE 33-bus and 118-bus systems and a transportation network with real-world EV data demonstrate the effectiveness of the proposed method.**

***Index Terms*—Congestion management, model-based safe deep reinforcement learning, EV charging pricing, power and transportation system, constrained Markov decision process.**

## NOMENCLATURE

### *Abbreviations*

| | |
|---|---|
| AMSDRL | Adaptive model-based safe deep. reinforcement learning. |
| CCEM | Constrained cross entropy method. |
| CMDP | Constrained Markov decision process. |
| CPO | Constrained policy optimization. |
| CS | Charging Station. |
| DDPG | Deep Deterministic Policy Gradient. |
| DRL | Deep reinforcement learning. |
| DSO | Distribution system operator. |
| EV | Electric vehicle. |
| FOCOPS | First order constrained optimization in policy space. |
| GMM | Gaussian mixed method. |
| SOC | State of charge. |

### *Sets*

| | |
|---|---|
| $\boldsymbol{B}/n$ | Set/index of buses in network. |
| $\boldsymbol{E}_{n,t}/i$ | Set/index of EVs at bus $n$ during time stage $t$. |
| $\boldsymbol{\Gamma}/\Delta t$ | Discrete operating time set/Operating. interval of power grid. |
| $\boldsymbol{H}/\boldsymbol{H}_n/j$ | CSs Set/set of CSs at bus $n$/index of CSs set. |
| $\boldsymbol{\mathcal{J}}/\mathbf{E}$ | Set of intersections/roads. |
| $\boldsymbol{E}_{n,t}^{c/ed/rd}$ | EVs set with charging/elastic charging/rigid charging demand. |

### *Parameters*

| | |
|---|---|
| $t_{n,i}^{a/l/ac}$ | Arrival/leaving/maximum acceptable time of EV user. |
| $SOC_{n,i}^{min/max}$ | Minimum/maximum SOC of EV battery. |
| $SOC_{n,i,t_l}$ | EV SOC at departure. |
| $SOC_{n,i,t_l}^{E}$ | EV expected energy state after charging |
| $\upsilon_{i,t}$ | Weight coefficient for inclinations. |
| $\omega$ | Conversion factor. |
| $v_{ab}^{0}/f_{ab}/q_{ab}$ | Free flow speed/current traffic flow/ capacity of the road $e_{ab}$. |
| $\vartheta_1, \vartheta_2, \vartheta_3$ | Road adaptability coefficients. |
| $L(e_{ab})$ | Length of road $e_{ab}$. |
| $k_{n,j}$ | Number of charging piles of CSs $j$ at bus $n$. |
| $L_{n,i}^{u/m/s}(j)$ | Length of urban expressway/main road/ secondary road from the position of EV $i$ to the target CS $j$. |
| $N_{n,j}^{C}$ | Number of available charging piles. |
| $N_{n,j,t}^{EV,A/L}$ | Number of arrival/leaving EVs. |
| $b_{n,i,t}^{A/L}(j)$ | EV binary arrival/leaving status. |
| $V_{min/max}$ | Minimum/maximum voltage magnitude. |
| $I_{max}$ | Maximum line current. |
| $P_{mn}^{L}$ | Capacity of line (m, n). |
| $\phi_{mn,t}^{max}$ | Maximum load percentage of line (m, n). |

### *Variable*

| | |
|---|---|
| $E_{n,i,t}$ | Energy pf EV $i$ at bus $n$ at time stage $t$. |
| $P_{n,i,t}^{EV,C/D}$ | EV charging/discharging power. |
| $P_{n,i,max}^{EV,C/D}$ | EV maximum charging/discharging power. |
| $b_{n,i,t}^{EV,C/D}$ | EV binary charging/discharging status. |
| $C_{n,i,t}(j)$ | Total cost of charging at target CS $j$ for EV user $i$ of bus $n$ at time stage $t$. |
| $T/FC_{n,i,t}(j)$ | Time consumption/financial cost. |
| $T_{n,i,t}^{D/W/C}(j)$ | Driving/waiting/charging time of EV user $i$. at bus $n$ who decides to charge at CS $j$. |
| $\chi_{n,j,t}$ / $S_{n,j,t}$ | Average/average service time of CS $j$. at bus $n$ during time stage $t$. |
| $E_{n,i,t}^{W}(j)$ | Energy consumed by EV $i$ at bus $n$ for a one-way trip to the target CS $j$. |
| $\Delta E_{u/m/s,t}$ | Energy consumption per unit mile in urban. expressway/main roads/secondary roads. |
| $\lambda_{n,j,t}^{C/P/S}$ | Charging price/electrical power price/ service fee of CS $j$ at bus n at time stage $t$. |

| | |
|---|---|
| $\lambda_{n,j,t}^{I}$ | Incentive price determined by DSO. |
| $\lambda_{n,j}^{I,min/max}$ | Minimum/maximum incentive price. |
| $\lambda_{n,t}^{DSO,M/D}$ | DSO electricity purchase price from the main grid/distributed generators at bus $n$. |
| $P_{n,t}/Q_{n,t}$ | Active/reactive power of bus $n$. |
| $P_{mn,t}/Q_{mn,t}$ | Active/reactive power injecting from bus $m$ to $n$. |
| $P_{n,j,t}^{CS,M/D}$ | Power from the main network/distributed generator to CS $j$ at bus $n$. |
| $P_{n,j,t}^{CSs}/P_{n,t}^{O}$ | Power consumed by CSs and non-EV loads. |
| $P_{n,t}^{DG}$ | Output of local distributed generator. |
| $r_{mn,t}/x_{mn,t}$ | Impedance/inductive resistance of line ($m$, $n$). |
| $V_{n,t}/l_{mn,t}$ | Square of node voltage magnitude/line current. |

## I. Introduction

The scale of electric vehicles (EVs) has been growing rapidly in recent years. According to the Global EV Outlook of the International Energy Agency, the number of EVs is expected to reach 1.81 billion in 2030, consuming 43000 GWh of electricity every year. The large amount of power drawn by random EV charging imposes great burdens on the distribution network and causes local congestion, which may lead to widespread blackouts associated with serious economic losses and negative social impacts [1].

Although updating the grid facilities could address the above issues directly, and is also essential for the increasing the penetration of sustainable energy resources and EVs, it would take a long time due to the policy and high costs involved. To solve this problem, many researchers have worked on the day-ahead market for EV aggregators to alleviate congestion by scheduling the grid assets wisely. In [2], a price-based EV scheduling method was applied to dispatch distributed energy resources to alleviate the congestion in heavily loaded feeders. In [3], a dynamic tariff was introduced to motivate aggregators to shift their flexible demands while the subsidy-based method in [4] encouraged users to change their planned schedules. Hu, *et. al* [5] proposed a power trading framework to coordinate EV aggregators for congestion management. Although the above works are efficient in relieving heavy congestion, planning ahead cannot fully avoid the congestion caused by time-varying demand, especially for EVs with high mobility and charging uncertainty. Therefore, real-time congestion management for EVs is essential for the economical and safe operation of the distribution network.

Real-time congestion management for EVs can be summarized into two types, namely charging management and pricing incentives. In the field of EV charging management, several methods based on smart charging technology have been proposed. In [6], a rule-based model predictive control method was proposed to regulate the output of PV generators and EV inverters to mitigate line congestion in an active distribution network. In [7], a coordinated scheme for EV charging and PV operation was proposed to prevent line overload and system energy loss. In [8], a field test validation was conducted in a real Danish distribution grid to verify the effectiveness of real-time EV charging strategies for congestion management. However, the above methods have security risks due to the uncertainty of the EV user trip and EV energy level. In addition, charging management methods are generally needed to remunerate EV users for acquiring EV flexible energy, but the amount that should be paid for incentives is complex to determine. In view of this, EV charging management is an auxiliary or emergency means for real-time congestion management, while pricing incentives are more suitable as a stable treatment.

In terms of pricing incentive approaches, Soltani *et al*. [9] used the conditional random field method for real-time optimal price adjustment to prevent local congestion, and in [10], a real-time pricing and scheduling strategy for EVs was proposed to eliminate the disruptive impact of overloading on the distribution network. However, these two studies did not consider the transportation network in EV scheduling, which makes price strategies impossible in reality due to traffic jams, road closures, and other traffic factors. Although the authors of [11]-[13] proposed dynamic pricing strategies for EV charging stations (CSs) under a holistic modeling framework for the transportation network and power distribution network, the user decision models in these studies were crude and could not accurately reflect different users' decisions regarding price incentives; examples include the rough classification of user groups into refusing to join and leaving in [11], and the lack of travel energy constraints in [12]. In addition, the simple representation of the traffic system makes the strategies impractical, for example, the lack of distinct road types in [13] has a negative impact on calculating the travel time and expected energy consumption, which directly affects the accuracy of user decisions. The studies in [14]-[15] achieve efficient coordination between EVs and CSs in an integrated electricity-transportation system, however, the proposed methods may not be suitable for real-time pricing for congestion management due to the long solution time of the evolution algorithms. In addition to the above methods, the work of [16] was creative, and the authors designed a real-time swapping market framework to incentivize flexible demand swap of EVs and heat pumps. However, the trading mechanism was not universal due to the harsh operating conditions.

Nevertheless, the aforementioned model-based approaches exhibit two fundamental limitations under a dynamic and stochastic environment. First, they highly rely on complete and accurate information about EVs, CSs and the power grid, which is difficult to obtain in practice, especially information under privacy limitations, such as user trips. Second, a fully model-based method may not solve the complex optimization problem effectively due to the EV user's uncertain behavior and time-varying operating conditions. In fact, most works have greatly simplified the proposed models with impractical assumptions to facilitate the solution.

These shortcomings of model-based methods motivate the application of the deep reinforcement learning (DRL) method. DRL is a data-driven control method that learns and outputs a near-optimal policy rather than a deterministic optimal solution, and this has made DRL methods popular in recent years for

solving real-time optimal control problems with uncertainty. In [17], a DRL-based holistic framework was proposed to stabilize the operation of coordinated power distribution network and transportation network through EV scheduling. In [18], a soft actor-critic method was developed to address continuous charging under the uncertainty of EV user behaviors. In [19], a DRL method with a prioritized experience replay strategy was proposed to solve the examined EV pricing problem. In [20], a graph DRL method was proposed for real-time CS recommendation considering coupled information of power and transportation networks. In [21], an optimal DRL-based EV driving model was proposed for EV charging scheduling in integrated power and transportation systems. Although the aforementioned methods achieved promising results, these methods can only satisfy soft constraints and cannot guarantee the security of the application, which is fatal in practice. In view of this, a few researchers have worked on safe DRL methods to solve the problem of poor applicability. In [22], the authors applied a safe DRL approach to solve the constrained EV charging issue for participating in demand response. In [23], a safe DRL method was presented to achieve the optimal operation of distribution networks considering the uncertainty of renewable resources. In [24], a multi-agent safe DRL approach was presented for the optimal energy management of networked microgrids in distribution systems. However, most of the existing safe DRL methods learn the strategy through a large amount of offline data due to the very low sample efficiency, while data collection has high cost or is dangerous in real-world applications. In addition, the large difference between the policy distribution and data distribution in the offline dataset leads to errors in the estimated values of security constraints, which make the trained policy unsafe.

To address all the problems mentioned above, this paper proposes an **a**daptive **m**odel-based **s**afe DRL (AMSDRL) method for the real-time EV congestion management problem to achieve social welfare maximization while eliminating distribution network congestion. The main contributions of this paper are summarized as follows:

(1) To the best of our knowledge, this paper is the first time to propose a pricing strategy for the real-time congestion management considering EV users' charging decisions in an integrated power and transportation network. We model the EV user charging decision-making process to indirectly obtain the unquantifiable relationship between price and charging demand, and then formulate the pricing problem as a bi-level model with constrained Markov decision process.

(2) We propose a model-based safe DRL framework and develop the AMSDRL algorithm, which uses a strict and adaptive cost constraint to eliminate potential environment transition modeling errors to ensure safe training results with limited historical data. Compared to the state-of-the-art safe DRL algorithms FOCOPS (first order constrained optimization in policy space) [25] and CPO (constrained policy optimization) [26], AMSDRL is much safer, with only 0.89% and 0.27% of the congestion cost during convergence, and more stable, with 90.60% and 77.97% fewer violations.

(3) The constrained cross-entropy method is applied to solve the pricing problem with a continuous action space rather than using neural networks to output the policy. This makes AMSDRL more sample efficient for scalability. Compared to the FOCOPS and CPO, AMSDRL reduced the training time by 51.92% and 40.44% on the IEEE 33-bus distribution network and by 57.12% and 45.30% on the IEEE 118-bus distribution network, respectively.

The rest of this paper is organized as follows. The mathematical formulation of congestion management for EVs is given in Section II. The AMSDRL algorithm is proposed in Section III. Case studies are presented in Section IV and the conclusion is given in Section V.

## II. Mathematical Models

### A. EV Model

Let $\boldsymbol{B} = \{0,1,2,\dots,n\}$ denote the set of buses and $\boldsymbol{E}_{n,t}$ denote the set of EVs at bus $n$ at time stage $t$. Then EV charging and discharging characteristics can be generally modeled as:

$$E_{n,i,t+1} = E_{n,i,t} + \left(P^{EV,C}_{n,i,t}\eta^{EV,C}_{n,i,t} - P^{EV,D}_{n,i,t}/\eta^{EV,D}_{n,i,t}\right)\Delta t \quad \forall n \in \boldsymbol{B}, i \in \boldsymbol{E}_{n,t}, t \in \boldsymbol{\Gamma}^{ev}_{n,i} \tag{1}$$

$$SOC_{n,i,t+1} = SOC_{n,i,t} + (E_{n,i,t+1} - E_{n,i,t})/B_{n,i} \quad \forall n \in \boldsymbol{B}, i \in \boldsymbol{E}_{n,t}, t \in \boldsymbol{\Gamma}^{ev}_{n,i} \tag{2}$$

$$0 \le P^{EV,C}_{n,i,t} \le b^{EV,C}_{n,i,t} \cdot P^{EV,C}_{n,i,\max} \quad \forall n \in \boldsymbol{B}, i \in \boldsymbol{E}_{n,t}, t \in \boldsymbol{\Gamma}^{ev}_{n,i} \tag{3}$$

$$0 \le P^{EV,D}_{n,i,t} \le b^{EV,D}_{n,i,t} \cdot P^{EV,D}_{n,i,\max} \quad \forall n \in \boldsymbol{B}, i \in \boldsymbol{E}_{n,t}, t \in \boldsymbol{\Gamma}^{ev}_{n,i} \tag{4}$$

$$0 \le b^{EV,D}_{n,i,t} + b^{EV,C}_{n,i,t} \le 1 \quad \forall n \in \boldsymbol{B}, i \in \boldsymbol{E}_{n,t}, t \in \boldsymbol{\Gamma}^{ev}_{n,i} \tag{5}$$

$$SOC^{\min}_{n,i} \le SOC_{n,i,t} \le SOC^{\max}_{n,i} \quad \forall n \in \boldsymbol{B}, i \in \boldsymbol{E}_{n,t}, t \in \boldsymbol{\Gamma}^{ev}_{n,i} \tag{6}$$

$$SOC_{n,i,t_l} \ge SOC^{E}_{n,i,t_l} \quad \forall n \in \boldsymbol{B}, i \in \boldsymbol{E}_{n,t} \tag{7}$$

$$\boldsymbol{\Gamma} = \left\{ t_j \middle| t_j = j\Delta t, j \in Z^* \right\} \tag{8}$$

where (1)-(6) describe the EV charging/discharging process and establish the EV energy storage model. (1)-(2) are the update formulations of the EV energy and SOC, respectively. (3)-(5) are the constraints for the EV charging and discharging status. (6) limits the EV energy level and (7) indicates that the SOC at departure should meet the EV user's energy demand. (8) defines the discrete time stage and the time interval is taken as 15 minutes in this study. The constraints are embedded in the environment, which are satisfied through the learning process between the agent and environment.

### B. EV User Cost Model

The total cost of charging at target CS $j$ for EV user $i$ with charging demand at bus $n$ at time stage $t$ is formulated as:

$$C^{EV}_{n,i,t}(j) = \upsilon_{i,t}\omega TC^{EV}_{n,i,t}(j) + (1-\upsilon_{i,t})FC^{EV}_{n,i,t}(j) \quad \forall n \in \boldsymbol{B}, i \in \boldsymbol{E}^{c}_{n,t}, j \in \boldsymbol{H}_n, t \in \boldsymbol{\Gamma} \tag{9}$$

where the CS set is $\boldsymbol{H} = \boldsymbol{H}_1 \cup \boldsymbol{H}_2 \cup \dots \cup \boldsymbol{H}_n = \{1,2,\dots,J\}$ and $J$ is the total number of CSs in the distribution networks. $\upsilon_{i,t}$ is the weight coefficient which reflects the different inclinations of EV users toward the two types of costs and $\omega$ is the coefficient that converts the time consumption into cost and its distribution can be obtained by means of questionnaires.

*1) Time Consumption*

The time consumption is estimated as:

$$T^{W}_{n,i,t}(j) \approx \frac{\chi_{n,j,t}^{\,k_{n,j}} E[S_{n,j,t}{}^{2}](E[S_{n,j,t}])^{k_{n,j}-1}}{2(k_{n,j}-1)!(k_{n,j}-\chi_{n,j,t}E[S_{n,j,t}])^{2}\left[\sum_{x=0}^{k_{n,j}-1}\frac{(\chi_{n,j,t}E[S_{n,j,t}])^{y}}{y!}+\frac{(\chi_{n,j,t}E[S_{n,j,t}])^{k_{n,j}}}{(k_{n,j}-1)!(k-\chi_{n,j,t}E[S_{n,j,t}])}\right]} \quad \forall n\in \boldsymbol{B}, j\in \boldsymbol{H}, t\in \boldsymbol{\Gamma} \tag{14}$$

$$TC^{EV}_{n,i,t}(j) = T^{D}_{n,i,t}(j) + T^{W}_{n,i,t}(j) + T^{C}_{n,i,t}(j) \quad \forall n\in \boldsymbol{B}, i\in \boldsymbol{E}^{c}_{n,t}, j\in \boldsymbol{H}, t\in \boldsymbol{\Gamma} \tag{10}$$

•*Driving Time:*

The transportation network can be described as an underlying undirected graph $\boldsymbol{\Gamma}=(\boldsymbol{\mathcal{J}},\mathbf{E})$. We applied the precise speed-flow model in [27] to simulate the EV driving process in dynamic traffic conditions. The driving speed of an EV on a directly connected road section $e_{ab}$ is expressed as follows:

$$\begin{cases} v_{ab}(t) = v^{0}_{ab} / \left(1+\left(f_{ab}(t)/q_{ab}\right)^{\xi}\right) \\ \xi = \vartheta_1 + \vartheta_2 \left(f_{ab}(t)/q_{ab}\right)^{\vartheta_3} \end{cases} \quad \forall e_{ab}\in \mathbf{E}, t\in \boldsymbol{\Gamma} \tag{11}$$

The driving time needed for EV user $i$ of bus $n$ to charge at CS $j$ is given as:

$$T^{D}_{n,i,t}(j) = \sum_{e_{ab}\in \mathbf{E}^{n,i}_{j}} T_{ab}(t) \quad \forall n\in \boldsymbol{B}, i\in \boldsymbol{E}^{c}_{n,t}, j\in \boldsymbol{H}, t\in \boldsymbol{\Gamma} \tag{12}$$

$$T_{ab}(t) = L(e_{ab}) / v_{ab}(t) \quad \forall e_{ab}\in \mathbf{E}^{i}_{j}, t\in \boldsymbol{\Gamma} \tag{13}$$

where $\mathbf{E}^{i}_{j}$ is the set of transportation paths from EV user $i$ at bus $n$ to CS $j$. The transportation parameters, including the free flow speed, traffic flow, and capacity of the road, are calculated from the measured sample data through the fixed-point detector. The path selection for the EV is obtained by using the Dijkstra algorithm based on the principle of the shortest time.

•*Waiting Time:*

Here we apply the theory of the M/G/k queue in [28] to estimate the waiting process. M/G/k presents an approximation for the mean time a customer waits in the queue in a $k$-server system that assumes Poisson arrivals but allows for a general service distribution. Then the mean waiting time can be calculated as (14), where in this study, $k_{n,j}$ is the number of charging piles of CSs $j$ at bus $n$. $E[S_{n,j,t}]$ denotes the average service time of CS $j$ at bus $n$ during time stage $t$.

•*Charging Time:*

The charging time $T^{C}_{n,i,t}(j)$ can be estimated as follows:

$$T^{C}_{n,i,t}(j) = (E^{C}_{n,i,t} + E^{W}_{n,i,t}(j)) / P^{EV,C}_{n,i,t} \quad \forall n\in \boldsymbol{B}, i\in \boldsymbol{E}^{c}_{n,t}, j\in \boldsymbol{H}, t\in \boldsymbol{\Gamma} \tag{15}$$

$$E^{C}_{n,i,t} = B^{EV}_{n,i}(SOC^{E}_{n,i} - SOC_{n,i,t}) \quad \forall n\in \boldsymbol{B}, i\in \boldsymbol{E}^{c}_{n,t}, t\in \boldsymbol{\Gamma} \tag{16}$$

$$\begin{cases} \Delta E_{u,t} = 0.247 + 1.52/v_{ab,t} - 0.004 v_{ab,t} \\ \qquad + 2.992\times 10^{-5} v_{ab,t} \\ \Delta E_{m,t} = -0.179 + 0.004 v_{ab,t} + 5.492/v_{ab,t} \\ \Delta E_{s,t} = 0.21 - 0.001 v_{ab,t} + 1.531/v_{ab,t} \end{cases} \quad \forall e_{ab}\in \mathbf{E}^{n,i}_{j}, n\in \boldsymbol{B}, i\in \boldsymbol{E}^{c}_{n,t}, j\in \boldsymbol{H}, t\in \boldsymbol{\Gamma} \tag{17}$$

$$E^{W}_{n,i,t}(j) = \Delta E_{u,t} L^{u}_{n,i}(j) + \Delta E_{m,t} L^{m}_{n,i}(j) + \Delta E_{s,t} L^{s}_{n,i}(j) \quad \forall n\in \boldsymbol{B}, i\in \boldsymbol{E}^{c}_{n,t}, j\in \boldsymbol{H}, t\in \boldsymbol{\Gamma} \tag{18}$$

where (16) denotes the expected charging energy and (17) is the estimation formula of energy consumption per unit mile according to [27]. (18) formulates the energy consumed by EV $i$ at bus $n$ for a one-way trip to the target CS $j$.

*2) Financial Cost*

The financial cost of each EV user is calculated as:

$$FC^{EV}_{n,i,t}(j) = E^{C}_{n,i,t}\lambda^{C}_{n,j,t} + E^{W}_{n,i,t}(j)\lambda^{C}_{n,j,t} \quad \forall n\in \boldsymbol{B}, i\in \boldsymbol{E}^{c}_{n,t}, j\in \boldsymbol{H}, t\in \boldsymbol{\Gamma}. \tag{19}$$

The first item of (19) is the cost for the expected charging energy, while the second item represents the cost for a one-way charging trip. The components of the charging price $\lambda^{C}_{n,j,t}$ are shown in (32).

### *C. Congestion Management Model*

Because the migration of EV loads is influenced by the pricing of each CS and the traffic topology, it is not useful to simply increase or decrease the total CS charging price, which could result in the loss of profits of CSs due to excessive load shedding, or even the aggravation of congestion.

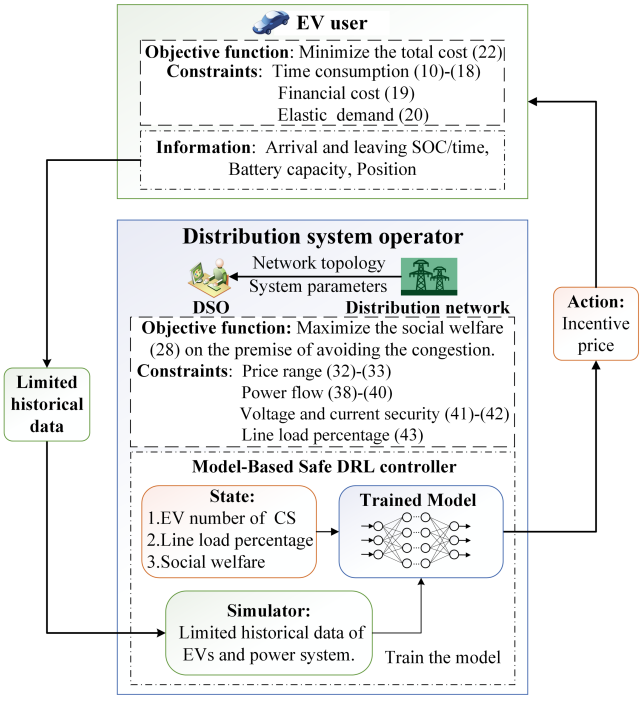



Fig. 1 Structure of the proposed method for congestion management.

For day-ahead congestion management, the common practice is to calculate the distribution locational marginal price (for a CS, this is the electrical power price) by the DSO under the bidding of traders (including CS aggregators) in the day-ahead market, and the CSs determine their own service fees.

However, for real-time congestion management, the DSO will impose an incentive price on the CS to achieve maximum

social welfare while eliminating local congestion. The structure of the proposed method for distribution network real-time congestion management is shown in Fig. 1. The real-time information is first observed by the DSO, including the CS EV number, line load percentage and social welfare, which will be explained in detail in the next section. Then the trained policy is loaded and the pricing strategy for the fast response of real-time congestion management will be provided. The policy is obtained by the model-based safety DRL controller trained using limited historical data on EV information as well as the power and transportation system parameters. The policy is continuously executed and trained until the congestion risk is removed. The EV data only includes the arrival and leaving SOC/time, and battery capacity without private information such as the trip chain.

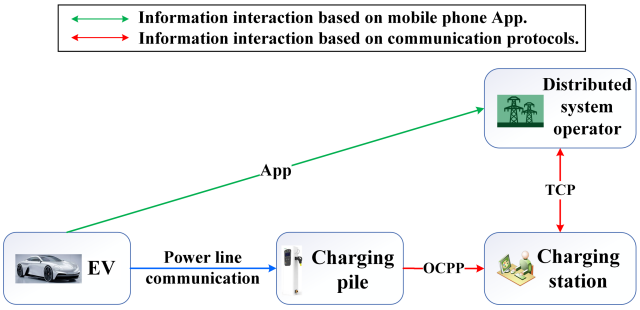


Fig. 2 Information sharing mechanism of the proposed method.

The DSO collects the limited historical information data of EV users through a mobile phone app to simulate the EV users charging decisions for the DRL training environment. The EVs transmit the information of battery management system, such as battery capacity and SOC, to the charging piles through power line communication. For real-time charging information, such as the EV number and EV charging load, the CSs obtain this information from the charging piles by communication interaction using the open charge point protocol (OCPP). Then, the CSs share the information with the DSO and receive the incentive price by communication through the transmission control protocol (TCP).

It is essential to establish a price-demand model for congestion pricing strategies, but it is extremely difficult to model the accurate relationship between price and demand directly by mathematical methods due to the complicated interplay between charging station pricing strategies. However, it is intuitive and easy to model the decision-making behaviors of EV users, which means we can solve the problem indirectly.

*1) EV User Charging Decision*

We classify the EVs with charging demand into the EV set of elastic demand and the EV set of rigid demand, which are shown as follows:

$$\boldsymbol{E}_{n,t}^{ed}=\{\boldsymbol{E}_{n,i,t}\,|\,T_{n,i,t}^{D}(j)<\Delta t, E_{n,i,t}>E_{n,i,t}^{W}(j), TC_{n,i,t}^{EV}(j)\le t_{n,i}^{ac}\} \quad \forall n\in\boldsymbol{B}, i\in\boldsymbol{E}_{n,t}^{c}, t\in\boldsymbol{\Gamma} \tag{20}$$

$$\boldsymbol{E}_{n,t}^{rd}=\{\boldsymbol{E}_{n,i,t}\,|\,\boldsymbol{E}_{n,t}^{c}/\boldsymbol{E}_{n,t}^{ed}\}\quad n\in\boldsymbol{B}, t\in\boldsymbol{\Gamma}. \tag{21}$$

The first constraint in (20) is set to avoid the risk of violating the real-time pricing timeline, the second constraint ensures that the EV has enough power to reach the target CS, and the third constraint indicates that the total charging time consumption should be satisfied within the maximum time acceptable to EV users. Price incentives only work for EVs with elastic charging demand and each EV user has only one choice of the target CS. EVs with rigid demand choose the nearest CS (i.e., only the time cost is considered).

The objective of the EV user decision strategy is to choose the CS $j$ to minimize the total cost, which is formulated as:

$$\min_{j} C_{n,i,t}^{EV} \quad \forall n\in\boldsymbol{B}, i\in\boldsymbol{E}_{n,t}^{c}, j\in\boldsymbol{H}, t\in\boldsymbol{\Gamma} \tag{22}$$

s.t. (8)-(18)

Then the update formula for the number of EVs of CS $j$ at bus $n$ can be obtained by (23):

$$N_{n,j,t}^{EV}=\begin{cases}N_{n,j,t-1}^{EV}+N_{n,j,t}^{EV,A}-N_{n,j,t}^{EV,L} & N_{n,j,t}^{EV}\le N_{n,j}^{C}\\ N_{n,j}^{C} & N_{n,j,t}^{EV}>N_{n,j}^{C}\end{cases} \quad \forall n\in\boldsymbol{B}, j\in\boldsymbol{H}, t\in\boldsymbol{\Gamma} \tag{23}$$

$$N_{n,j,t}^{EV,A}=\sum_{n}^{|\boldsymbol{B}|}\sum_{i}^{|\boldsymbol{E}_n|}b_{n,i,t}^{A}(j)\quad \forall n\in\boldsymbol{B}, i\in\boldsymbol{E}_{n}, j\in\boldsymbol{H}, t\in\boldsymbol{\Gamma} \tag{24}$$

$$b_{n,i,t}^{A}(j)=\begin{cases}1, & \arg\min\left\{C_{n,i,t}^{EV}\right\}=j\\ 0, & otherwise\end{cases} \quad \forall n\in\boldsymbol{B}, i\in\boldsymbol{E}_{n,t}, j\in\boldsymbol{H}, t\in\boldsymbol{\Gamma} \tag{25}$$

$$N_{n,j,t}^{EV,L}=\sum_{n}^{|\boldsymbol{B}|}\sum_{i}^{|\boldsymbol{E}_n|}b_{n,i,t}^{L}(j)\quad \forall n\in\boldsymbol{B}, i\in\boldsymbol{E}_{n,t}, j\in\boldsymbol{H}, t\in\boldsymbol{\Gamma} \tag{26}$$

$$b_{n,i,j,t}^{L}=\begin{cases}1, & SOC_{n,i,j,t}\ge SOC_{n,i,j,t_l}^{E}\,||\,t\ge t_{n,i}^{a}+t_{n,i}^{ac}\\ 0, & otherwise\end{cases} \quad \forall n\in\boldsymbol{B}, i\in\boldsymbol{E}_{n,t}, j\in\boldsymbol{H}, t\in\boldsymbol{\Gamma} \tag{27}$$

where (24)/(26) is the arrival/leaving EV number, and (25)/(27) defines the EV binary arrival/leaving status.

*2) Pricing Strategy of the DSO for Congestion Management*

Our goal is to maximize the social welfare $\Pi^{SW}$ under the premise of avoiding congestion, i.e., considering the profits of both the generation and demand sides. Let $\boldsymbol{L}\subset\boldsymbol{B}\times\boldsymbol{B}$ denote the set of lines. For each $(m,n)\in\boldsymbol{L}$, $m$ is the unique parent bus of bus $n$, then the objective of the DSO for congestion management is formulated as follows:

$$\max_{\lambda_{n,j,t}^{I}}\Pi^{SW}=\sum_{t\in\boldsymbol{\Gamma}}\left(\sum_{n\in\boldsymbol{B}}\sum_{j\in\boldsymbol{H}_n}\Pi_{n,j,t}^{CS}+\Pi_{t}^{DSO}\right) \tag{28}$$

CS constraints:

$$\Pi_{n,j,t}^{CS}=\Pi_{n,j,t}^{CS,C}-C_{n,j,t}^{CS,P}\quad \forall n\in\boldsymbol{B}, j\in\boldsymbol{H}_n, t\in\boldsymbol{\Gamma} \tag{29}$$

$$\Pi_{n,j,t}^{CS,C}=\sum_{i}^{N_{n,j,t}^{EV}}\lambda_{n,j,t_a}^{TC}P_{n,i,t}^{EV}\Delta t\quad \forall n\in\boldsymbol{B}, i\in\boldsymbol{E}_{n,t}, j\in\boldsymbol{H}_n, t\in\boldsymbol{\Gamma} \tag{30}$$

$$C_{n,j,t}^{CS,P}=\lambda_{n,j,t}^{P}P_{n,j,t}^{CS}\Delta t\quad \forall n\in\boldsymbol{B}, j\in\boldsymbol{H}_n, t\in\boldsymbol{\Gamma} \tag{31}$$

$$\lambda_{n,j,t}^{C}=\lambda_{n,j,t}^{P}+\lambda_{n,j,t}^{S}+\lambda_{n,j,t}^{I}\quad \forall n\in\boldsymbol{B}, j\in\boldsymbol{H}, t\in\boldsymbol{\Gamma} \tag{32}$$

$$\lambda_{n,j}^{I,\min}\le\lambda_{n,j,t}^{I}\le\lambda_{n,j}^{I,\max}\quad \forall n\in\boldsymbol{B}, j\in\boldsymbol{H}, t\in\boldsymbol{\Gamma} \tag{33}$$

DSO constraints:

$$\Pi_{t}^{DSO}=\sum_{n\in\boldsymbol{B}}\sum_{j\in\boldsymbol{H}_n}C_{n,j,t}^{CS,P}-C_{n,j,t}^{DSO,M}-C_{n,j,t}^{DSO,D}\quad \forall t\in\boldsymbol{\Gamma} \tag{34}$$

$$C_{n,j,t}^{DSO,M}=\lambda_{n,t}^{DSO,M}P_{n,j,t}^{CS,M}\Delta t\quad \forall n\in\boldsymbol{B}, j\in\boldsymbol{H}_n, t\in\boldsymbol{\Gamma} \tag{35}$$

$$C_{n,j,t}^{DSO,D}=\lambda_{n,t}^{DSO,D}P_{n,j,t}^{CS,D}\Delta t\quad \forall n\in\boldsymbol{B}, j\in\boldsymbol{H}_n, t\in\boldsymbol{\Gamma} \tag{36}$$

$$P_{n,j,t}^{CS} = \sum_{i=1}^{N_{n,j,t}^{EV}} P_{n,i,t}^{EV} = P_{n,j,t}^{CS,M} + P_{n,j,t}^{CS,D} \quad \forall n \in \boldsymbol{B}, i \in \boldsymbol{E}_{n,t}, j \in \boldsymbol{H}_n, t \in \boldsymbol{\Gamma} \tag{37}$$

Grid constraints:

$$\begin{cases} \sum_{mn\in\boldsymbol{L}} (P_{mn,t} - r_{mn} l_{mn,t}) - \sum_{nq\in\boldsymbol{L}} P_{nq,t} = \sum_{j\in\boldsymbol{H}_n} P_{n,j,t}^{CS} + P_{n,t}^{O} - P_{n,t}^{DG} \\ \sum_{mn\in\boldsymbol{L}} (Q_{mn,t} - r_{mn} x_{mn,t}) - \sum_{nq\in\boldsymbol{L}} Q_{nq,t} = Q_{n,t} \end{cases} \tag{38}$$

$$v_{n,t} = v_{m,t} + 2(P_{mn,t} r_{mn} + Q_{mn,t} x_{mn}) - l_{mn,t}(r_{mn}^2 + x_{mn}^2) \tag{39}$$

$$\left\| \begin{matrix} 2P_{mn,t} \\ 2Q_{mn,t} \\ l_{mn,t} - v_{n,t} \end{matrix} \right\|_2 \le l_{mn,t} + v_{m,t} \tag{40}$$

$$V_{\min}^2 \le v_{n,t} \le V_{\max}^2 \tag{41}$$

$$0 \le l_{mn,t} \le I_{\max}^2 \tag{42}$$

$$\phi_{mn,t} = P_{mn,t} / P_{mn}^{L} \le \phi_{mn,t}^{\max} \quad \forall (m,n) \in \boldsymbol{L} \tag{43}$$

*for* $\forall n, j, t$ *in* (38)-(43), $n \in \boldsymbol{B}, j \in \boldsymbol{H}, t \in \boldsymbol{\Gamma}$

where (29) denotes the CS profits, which consist of the EV charging profits $\Pi_{n,j,t}^{CS,C}$ defined by (30) and the CS electricity purchase cost $C_{n,j,t}^{CS,P}$ define by (31). The electrical power price (locational marginal price) is set in the day-ahead market and the service fee is determined by the CSs while the incentive price can be adjusted by the DSO for real-time congestion management.

The energy loss rate and the output of the distributed generator are often assumed to be a constant and a predicted power output, which does not increase the difficulty of solving the congestion management problem. Therefore, we can assume that the marginal price of each bus has been set by the day-ahead market and ignore the energy loss. In this case, maximizing the profits of the DSO is equivalent to maximizing the total profits of the DSO from electricity sales to CSs, so the total profits of the DSO in (28) can be formulated as (34) ignoring the other components of DSO profits for convenience. (35)/(36) is the cost to the DSO of purchasing power from the main grid/distributed generators to sell to CSs.

Finally, the grid constraints for (28) are given by (38)-(43). (38) formulates the linearized power balance. (39) describes the relationship between the node voltage and branch current. (38) is the convex relaxation of the power flow equality constraints based on second order cone programming. (40)-(43) are the constraints on the node voltage amplitude, line current and line congestion.

Although (22) and (28) constitute a bi-level optimization problem, they are not suitable to be solved by traditional optimization methods for the following reasons. First, the problem involves both discrete and continuous actions, which makes it challenging to solve the algorithm. Second, model-based methods require a real-time explicit physical model with accurate parameters to formulate the behavior of EV users, which is not possible to obtain. Moreover, as the problem scale grows, the problem will become more time-consuming because of the dramatic increase in the dimensionality of the optimal variable with many complex constraints.

## III. Methodology

DRL is a data-driven method and can be used to solve a dynamic, real-time control problem. However, a suitable DRL method for the congestion management problem should be efficient and safe, and can be deployed in the real world by training on only limited historical data. On the one hand, certain exploratory trial-and-error methods are unacceptable for congestion management, as they may cause serious safety issues such as line burning and equipment damage. In addition, there are usually differences between data distributions in offline and online datasets due to the insufficiency and uncertainty of the sample data (i.e., the offline dataset only contains part of the data distributions) which will make the conventional model-based DRL unsafe. On the other hand, unlike in the field of computer games, we are not able to generate a large number of pricing scenarios with low cost. In fact, the datasets for most scenarios are insufficient to support real applications of even off-policy reinforcement learning methods, such as the state-of-the-art soft actor-critic (SAC) algorithm. The ideal approach for congestion management is to train out the safe pricing strategy based on a simulator established by limited historical data collected by CS aggregations and the DSO.

In view of this, inspired by the work in [29] and [30], we propose the AMSDRL algorithm to solve the congestion management problem in the form of a CMDP in this section. The proposed bi-level model acts as the environment of the joint distribution and transportation networks, and the DRL agent can learn the optimal charging strategy through direct interactions with the environment. Specifically, based on the observed information, the agent takes actions (incentive prices) under the constraints of the distribution network, which in turn influences the EV users' decisions. The simulated EV users then choose the optimal charging station under the transportation constraints, which also changes the EV load distribution in the power network.

### A. Model-Based Safe DRL Framework

The infinite-horizon CMDP tuple in safe reinforcement learning can be defined as a tuple $(S, A, T, r, c, \gamma, \mu_0)$, where $S$ and $A$ are the spaces of states and actions, $T: S \times A \to D(S)$ is the transition distribution and $P_0 \in D(S)$ is the initial distribution. $r(S, A)$ is the reward function, and $c(S, A)$ is the cost function. $\gamma \in (0,1)$ is the discount factor. A policy $\pi$ is a mapping from states to distributions over actions. Let $\mathbb{P}_{T,t}^{\pi}$ denote the probability of being in state $s$ at timestep $t$ when actions are sampled according to policy $\pi$ and transition $T$. Then we let $\rho_{s,a}^{\pi}(s,a) := (1-\gamma)\,\pi(a|s) \sum_{t=0}^{\infty} \mathbb{P}_{T,t}^{\pi}(s)$ be the state-action occupancy distribution of policy $\pi$ under dynamic transition $T$, which is a properly normalized probability distribution. The value function in reinforcement learning is formulated as $V^{\pi}(s) := \mathbb{E}_T^{\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s]$ with an expectation over $s_0 \sim \mu_0(\cdot)$, $s_t \sim T(s_t \mid s_{t-1}, a_{t-1})$, and $a_t \sim \pi(\cdot|s_t)$. Similarly, we define the cost value function as $V_c^{\pi}(s) := \mathbb{E}_T^{\pi}[\sum_{t=0}^{\infty} \gamma_c^t c(s_t, a_t) \mid s_0 = s]$. Then, the goal of safe reinforcement learning is to find the optimal feasible policy $\pi^*$ that solves the constrained optimization problem with a cumulative constraint threshold as follows:

$$\max_{\pi} \quad J(\pi) := V^{\pi}(s) = \mathrm{E}\left[\sum_{t=0} \gamma^t r\left(s_t, a_t\right)\right]$$
$$\text{s.t.} \quad J_c(\pi) := V_c^{\pi}(s) = \mathrm{E}\left[\sum_{t=0} \gamma_c^t c\left(s_t, a_t\right)\right] \le C \tag{1}$$

However, the transition function $T$ of a real environment is typically difficult to obtain, which means that $\pi^*$ cannot be solved directly. The problem can be solved by model-based deep reinforcement learning which learns an estimated transition function $\hat{T}$ from dataset $\mathcal{D} := \{(s, a, r, c, s')\}$ to replace $T$. Then we can obtain the surrogate model-based objective function $\hat{J}(\pi)$ and cost constraint $\hat{J}_c(\pi) < C$.

According to (45), a basic model-based safe DRL framework can be defined, which iterates over the following three steps: (1) approximately solving for the optimal policy $\hat{\pi}^*$, (2) collecting trajectory data $(s, a, r, c, s')$ from $\hat{\pi}^*$, and (3) updating estimated transition function $\hat{T}$ using all collected data thus far.

### B. CMDP Formulation for Model-Based Safe DRL

In this section, the congestion management problem is formulated as a constrained Markov decision process.

*1) Definition of States*

$s_t$ is the observed status information of CSs and the DSO at time stage $t$, including the line load percentage $\phi_{n,t}$, the number of EVs at each charging station $N_{n,j,t}^{EV}$, and the social welfare $\Pi_t^{SW}$. Congestions in distribution networks usually occur at several fixed lines at regular time periods of peak load. In addition, there are many efficient forecasting methods to predict congestion. It is practical to choose congestion lines or lines with high congestion risk based on historical data. The common method of selecting associated CSs is to calculate the power transfer distribution factors with respect to the congested lines. For details on the CSs selection method, see the decentralized submarket identification method in Section III in [31].

*2) Definition of Actions*

$a_t$ is the control action given by the DSO during time stage $t$, which is the CS service fee $\lambda_{n,j,t}^{S}$. A continuous action space has an infinite variety of action vectors. To reduce the computation time for exploration, the values of the action variables should be pre-constrained according to prior knowledge.

*3) Definition of Reward and Cost*

The reward and cost functions are given as follows:

$$r_t = \tau_1 \Pi_t^{SW} \quad c_t = \tau_2 \phi_t^E \quad \forall t \in \boldsymbol{\Gamma} \tag{45}$$

$$\phi_t^E = \begin{cases} \sum_{n \in \boldsymbol{B}} (\phi_{mn,t} - \phi_{mn,t}^{\max}), & \phi_{mn,t} > \phi_{mn,t}^{\max} \\ 0, & else \end{cases} \quad \forall (m,n) \in \boldsymbol{L}, t \in \boldsymbol{\Gamma} \tag{46}$$

where $\phi_t^E$ is the violation error of congestion. $\tau_1$ and $\tau_2$ are positive weight coefficients. To make it easier for the neural network to calculate the gradient to improve the algorithm training efficiency, we suggest adjusting $\tau_1$ and $\tau_2$ so that the reward and cost values do not exceed 200. We usually set $\tau_2$ to 100, which means that $c_t$ represents the total percentage of congestion violations, while $\tau_1$ is mainly adjusted according to the EV fleet size.

### C. Adaptive Model-Based Safe DRL

*1) Strict and Adaptive Cost Constraint*

The model-based safe DRL framework has a high training efficiency; however, it cannot fully guarantee safety due to the modeling error. We have the following conclusion regarding the difference between the expected costs of $T$ and $\hat{T}$:

$$\begin{aligned} & J_c(\pi) - J_c(\pi) \\ &= \frac{1}{1-\gamma_c}\left(\sum_{s,a} \rho_{\hat{T}}^{\pi}(s,a) c(s,a) - \sum_{s,a} \rho_T^{\pi}(s,a) c(s,a)\right) \\ &\le d_{\mathrm{F}}\left(T(s,a), \hat{T}(s,a)\right). \end{aligned} \tag{47}$$

The proof of the conclusion is given in the appendix. (47) shows that satisfying $\frac{1}{1-\gamma}\sum_{s,a} \rho_{\hat{T}}^{\pi}(s,a) c(s,a) \le C$ does not guarantee that the resulting optimal policy $\pi^*$ would not violate the safety constraint in the real CMDP (i.e., $\frac{1}{1-\gamma}\sum_{s,a} \rho_T^{\pi}(s,a) c(s,a) \le C$) because of the modeling error $\mathrm{d}_{\mathcal{F}}(\hat{T}(s,a), T(s,a))$. To guarantee that policy $\pi^*$ is feasible for $T$, we rewrite the safety model-based framework (44) with the strict and adaptive cost function $J_S(\pi)$ as follows:

$$\begin{aligned} & \max_{\pi} \quad J(\pi) := \mathrm{E}\left[\sum_{t=0} \gamma^t r\left(s_t, a_t\right)\right] \\ & \text{s.t.} \quad J_S(\pi) := \mathrm{E}\left[\sum_{t=0} \gamma_c^t c\left(s_t, a_t\right) + \beta u_T(s_t, a_t)\right] \le C \\ & \quad d_{\mathrm{F}}\left(T(s,a), \hat{T}(s,a)\right) \le u_T(s_t, a_t) \end{aligned} \tag{48}$$

where $u_{\hat{T}}$ is a heuristic cost penalty based on the statistics of the transition model. We use a neural ensemble that outputs a Gaussian distribution $\hat{T}_\theta = \mathcal{N}(\mu_\theta(s_t, a_t), \Sigma_\theta(s_t, a_t))$ to formulate the transition and learn an ensemble of $N$ models $\{\hat{T}_\theta^i = \mathcal{N}(\mu_\theta^i, \Sigma_\theta^i)\}_{i=1}^N$. In other words, we use a neural network with parameter $\theta$ to learn and predict the transitions in CMDP. Then we take the maximum Frobenius norm of the standard deviation of the learned models in the ensemble as $u_{\hat{T}}$, which is done for offline reinforcement learning as follows:

$$u_T(s_t, a_t) = \max_i^N \left\| \sum\nolimits_\theta^i (s_t, a_t) \right\|_{\mathrm{F}}. \tag{49}$$

However, we find that using the cost penalty $u_{\hat{T}}$ directly is inflexible, leading to a poor performance in practice, so we set an adaptive scalability factor $\beta$ to balance safety and exploration, which is shown in (48). We adopt the PI control method for updating $\beta$ as follows:

$$\beta_{t+1} = \beta_t + \alpha(J_C(\pi_t) - C) \tag{50}$$

where α is the learning rate. When $J_C(\pi) > C$, $\beta$ will be updated to a larger value to tighten the cost constraint, and vice versa.

*2) Constrained Cross-Entropy Method*

The action and state space of the pricing strategy for congestion management is continuous, therefore, it is not feasible to solve (48) by rewriting it as its dual problem. (48) is also not suitable for solution by formulating it in terms of state-action pairs, as in discrete environments, due to the curse of dimensionality.

To solve the pricing problem with a continuous action space, we apply the constrained cross-entropy method. We use the distributions of the natural exponential family (NEF) $F_v = \{f(\cdot$

$;v), v \in \mathcal{V} \subseteq \mathbb{R}^{d_v}\}$ over action space $A$ to cover the distributions on policy space Π. The NEF contains many useful distributions, such as the Gaussian distribution and Bernoulli distribution. Then the optimal solution of (48) can be reformulated as follows:

$$\upsilon^* = \arg\max_{\upsilon \in \mathcal{V}} \mathrm{E}_{a\sim f_\upsilon}[J(a) \mid a \in A_S] \quad f_v \in F_v \tag{51}$$

where $A_S$ is the set of feasible actions for the strict and adaptive constraint $J_S(\pi) < C$. Drawing on the cross-entropy method, we first need to search for a surrogate function with solvable form to replace (51). The surrogate function is the conditional expectation of $J(a)$ over the sample policies $\pi$ with sampling distribution $f_\upsilon$.

Here, we introduce the concept of the $\rho$-quantile to express the sample policy: the $\rho$-quantile of random variable $X$ is defined as a scalar σ such that $\mathbb{P}(X \le \sigma) \ge \rho$ and $\mathbb{P}(X \ge \sigma) \le 1-\rho$.

We use $\varphi_G(v,\rho)$ to denote the $\rho$-quantile of function $G$: $A \to \mathbb{R}$ for $a\sim f_\upsilon$, $\upsilon \in \boldsymbol{\mathcal{V}}$. Generally, $\rho$ represents the proportion of highly ranked actions. If $G(a) \le \varphi_G(v,\rho)$, the G-value of the action is smaller than at most $1-\rho$ of all actions with sampling distribution $f_\upsilon$; the opposite implies that the G-value of the action is greater than at least $\rho$ of all actions.

We define the set of operations $\mathcal{O} \coloneqq \{\ge, >, \le, <, =\}$, and let $\delta: \mathbb{R} \times \mathcal{O} \times \mathbb{R} \to \{0,1\}$ be an indicator function for $\circ \in \mathcal{O}$, $a\&b \in \mathbb{R}$; $\delta(a \circ b) = 1$ if and only if $(a \circ b)$ holds. Similarly, we define the sample indicator function $H: A \times \mathcal{V} \times (0,1) \to \{0,1\}$ as follows:

$$\begin{aligned} H(a,\upsilon,\rho) \coloneqq{}& \delta(\varphi_{J_s}(\upsilon,\rho) > C)\delta(J_s(a) \le \varphi_{J_s}(\upsilon,\rho)) + \\ & \delta(\varphi_{J_s}(\upsilon,\rho) \le C)\delta(K(a) \ge \varphi_K(\upsilon, 1-\rho)) \\ K(a) \coloneqq{}& J(a)\delta(J_s(a) \le C). \end{aligned} \tag{52}$$

Then, the surrogate function for constrained cross-entropy can be expressed as follows:

$$S(\upsilon;\rho) = \mathrm{E}_{a\sim f_\upsilon}[J(a)H(a,\upsilon,\rho)]. \tag{53}$$

The optimal distribution with minimal variance of the original objective function (51) can be calculated as follows:

$$\upsilon^* = \arg\max_{\upsilon \in \mathcal{V}} S(\upsilon;\rho) \tag{54}$$

Finally, instead of updating the distribution parameters directly via the solution of (54), we use the following smooth updating rule with weighted factor $\kappa$:

$$\hat{v}_t = \kappa v_t^* + (1-\kappa)\hat{v}_{t-1}. \tag{55}$$

For a more detailed explanation, (52)-(54), i.e., the updated rules for the sampling distribution of actions in the constrained case, can be expressed as follows: If $\varphi_{J_S}(v,\rho) > C$, which means that the proportion of feasible actions is lower than $\rho$, we select the actions with the lowest cost; i.e., action $a$ is preferred if $J_S(a) \le \varphi_{J_S}(v,\rho)$. If $\varphi_{J_S}(v,\rho) \ge C$, which means that the proportion of feasible actions is at least $\rho$, we select the actions that satisfy the constraint with the highest reward; i.e., action $a$ is preferred if $J_S(a) \le C$ and $J(a) \ge \varphi_J(v, 1-\rho)$. In this study, we use the normal distribution in the NEF as $f_v$, and the full AMSDRL method is described in Algorithm 1.

**Algorithm 1:** AMSDRL Algorithm

**Inputs:** Transition model $\hat{T}_\theta$, an empty experience buff $\mathcal{D}$, cumulative constraint threshold $C$, initial $\beta$ value, $\beta$ learning rate $\alpha$, smooth learning rate κ, initial sampling distribution $\mathcal{N}(\mu_0, \Sigma_0)$, $\rho$-quantile , training episode $I$, CCEM max iteration $J$, population size $N$, planning horizon $H$

**Outputs**: Optimal action sequence $\{a_1^*, \dots, a_m^*\} \coloneqq \mu_J$

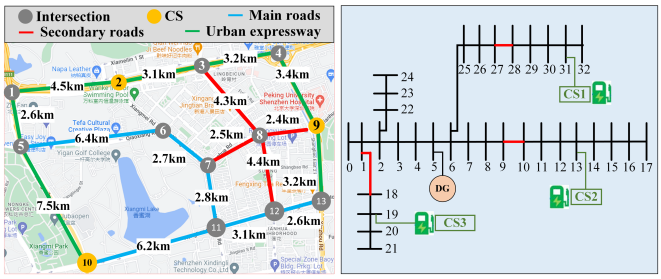
**Process**:

Initialize $\mathcal{D}$ with random policy, $k \leftarrow \lceil \rho N \rceil$
**for** $i$ = 1, …, $I$ **do**
  Train $\hat{T}_\theta$ using data in $\mathcal{D}$
  **for** $j$ = 1, ..., $J$ **do**
    Sample $N$ action sequences $A^1 \coloneqq \{a_t^1\}_{t=1}^H = \{a_1^1, a_2^1, a_3^1, \dots, a_H^1\}, \dots,$ $A^N \coloneqq \{a_t^N\}_{t=1}^H = \{a_1^N, a_2^N, a_3^N, \dots, a_H^N\}$.
    Evaluate the actions sequences by simulating trajectories in $\hat{T}_\theta$ according to (51).
    Construct feasible set $\Psi \coloneqq \{\theta_n | J_S(A^n) \le C, n \in [N]\}$ with (52)-(53)
    **if** $|\Psi| < k$ **then**
      Construct elite set $\varsigma \coloneqq$ {The $k$ sequences out of all $\{A^n\}_{n=1}^N$ with lowest costs}
    **else**
      Construct elite set $\varsigma \coloneqq$ {The $k$ sequences in $\Psi$ with highest rewards}
    **end if**
    Compute $\mu_j, \Sigma_j$ according to (54)-(55)
  **end for**
  Collect trajectory and store to buff $\mathcal{D}$
  Calculate $J_C(\pi_t)$ and update $\beta$ based on (50)
**end for**

## IV. Case Study

### A. Test System and Parameter Settings

All tests are performed on a computer with a 3.40 GHz CPU,1050Ti GTX graphics card and 16 GB RAM. We apply PyTorch to formulate the neural network framework of the AMSDRL algorithm and Pandapower to establish the modified IEEE 33-bus distribution network.




Transportation network topology  Modified IEEE 33-bus distribution network

Fig. 3 Illustration of the transportation/distribution network.

The topologies of the transportation network and modified IEEE 33-bus distribution network are given in Fig. 3 where CS1/2/3 is at node 2/9/10, and detailed information on the transportation network, CSs and EV users can be found in Tables I~II. Power lines $l_{27,28}, l_{9,10}, l_{1,18}$ are prone to congestion with a maximum load percentage limit of 90% and the corresponding line capacities are 15.2, 5.6, and 4.9(MW). The detailed parameters of the AMSDRL method are given in Table III, where we set the learning rate of the adaptive scalability factor higher than those of the ensemble networks to improve the initial training speed. The population size and $\rho$-quantile determine the stability and speed of convergence. Too large a population size and $\rho$-quantile can lead to too many elite actions being selected, which will reduce the training speed, and

vice versa, leading to violent training fluctuations. Based on our test, the appropriate coordinated values of the population size and $\rho$-quantile are 0.15 and 500 for this study, and the corresponding value of the maximum iteration number for CCEM is better taken as 3~5.

TABLE I. PARAMETERS OF TRAFFIC SYSTEM

| Parameters | Value | | |
|---|---|---|---|
| | Urban expressway | Main roads | Secondary roads |
| $\vartheta_1/\vartheta_2/\vartheta_3$ | 1.726/3.15/3 | | 2.076/2.870/3 |
| $v^0$ | 70 | 50 | 40 |

TABLE II. PARAMETERS OF CSS AND EV USERS

| Parameters | Value |
|---|---|
| Number of charging piles of CS1/CS2/CS3 | 50/30/20 |
| Maximum/Minimum incentive price (¥/kWh) | 0.5/0 |
| Electricity power price of CSs (¥/kWh) | 0.65 |
| Service fee of CSs (¥/kWh) | 0.35 |
| Electricity cost from the main grid (¥/kWh) | 0.5 |
| Cost of distributed generator (¥/kWh) | 0.3 |
| Distributed generator maximum output (MW) | 9 |
| Distribution of the service time$(\mu, \sigma^2)$ (min) | (48.37, 432.50) |
| Rated charging power (kW) | 60 |
| EV user inclination coefficient $(\mu, \sigma^2)$ | (0.5, 0.1) |
| EV user time cost conversion factor | (15, 3) |

TABLE III ALGORITHM PARAMETERS

| Parameters | Value |
|---|---|
| Optimizer | Adam |
| Number of hidden layers | 5 |
| Number of hidden units per layer | 256 |
| Learning rate of network/ adaptive scalability factor | 1e-3/1e-2 |
| Discount factor/cumulative constraint threshold | 0.99/0 |
| Max iteration number for CCEM/ population size/ $\rho$-quantile/ planning horizon/weighted factor for smooth update | 5/500/0.15/12/0.9 |
| Replay buffer size | 1e6 |
| Number of samples per mini batch | 128 |
| Nonlinearity | Swish |

For convenience, we define the ratio of the current traffic flow and capacity of the road as the road condition indices $I(t)$. The detailed road condition indexes for urban expressways, main roads and secondary roads from 14:00 to 17:00 are shown in Fig. 4 according to the data in [32]. The number of EVs with charging demand is given in Fig. 5. The EVs technical specifications are given in Table IV, which are calculated based on the GMM (Gaussian mixed method). The optimal cluster number of the GMM is determined by the silhouette coefficient. We obtain information on the arrival/departure SOC, battery, and parking time, as well as the number of EVs with charging demand from 2098 sets of historical data of CSs in the interval 2:00 pm-5:00 pm, July 10 2022, in the actual transportation network, Shenzhen.

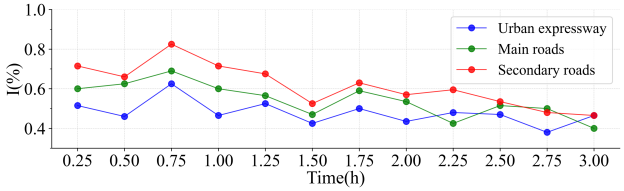



Fig. 4 Road condition index for three types of roads.

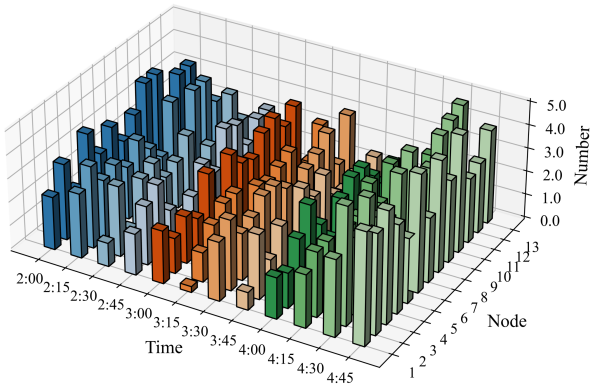



Fig. 5 Number of EVs with charging demand at each traffic node.

TABLE IV GMM MODEL PARAMETERS FOR EV TECHNICAL SPECIFICATIONS

| Features | Mean value ($\mu$) | Variance ($\sigma^2$) |
|---|---|---|
| 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], [43.3 → 81.69]} | {[[369.62, −0.01], [−0.01,0.60]], [[390.77,146.56], [146.56,258.27]]} |

## B. *Effectiveness of the AMSDRL Method*

Fig. 6 compares the training performance of the proposed AMSDRL and the state-of-the-art safe DRL algorithms with the same learning rate, training samples per step and action preset range. FOCOPS [25] and CPO [26] are the representative algorithms of the two dominant paradigms of policy projection and the Lagrangian method in safe DRL. Notably, safe reinforcement learning does not mean that the trained policy will not violate the constraints. Instead, the violations are much less frequent than other DRL algorithms; i.e., safe DRL can have a very small number of mild violations [33]. The dark line represents the average reward/cost per 50 steps while the light line represents the actual reward/cost of each training step. The trend of declining rewards is caused by the conflict between the reward and cost. Specifically, the trained policy has to create a price difference to shift the EVs with elastic demand to eliminate and avoid congestion, which might reduce social welfare compared to the most random policy during the initial training process.

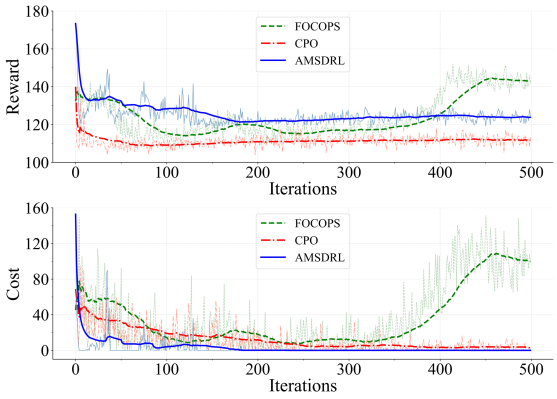



Fig. 6 Comparisons of performance of the state-of-art safe DRL algorithms.

TABLE V STATISTICAL RESULTS OF THE TRAINING PERFORMANCE

| Items | AMSDRL | CPO | FOCOPS |
|---|---|---|---|
| Average reward during convergence | 123.24 | 121.82 | 117.69 |
| Average cost during convergence | 0.036 | 4.05 | 13.52 |
| Violations number of the trained model under the test set (500 iterations) | 13 | 59 | 117 |
| Time to convergence (min) | 70.87 | 118.98 | 147.40 |
| Loading time for the execution (s) | 31.14 | 24.60 | 46.06 |

To further compare the performance of the three algorithms, the statistical results are presented in Table V. The average reward and cost during convergence reflect the effect of the trained model while the training violations and convergence time reveal the stability and efficiency of the algorithms. The loading times for the execution of the three algorithms are acceptable, but the proposed AMSDRL exhibits the highest average reward and is significantly safer maintaining only 0.89% and 0.27% of the cost of FOCOPS and CPO, respectively, which shows the perfect balance between effectiveness and constraint. In addition, AMSDRL outperforms the other two algorithms in terms of efficiency and stability, reducing the training time by 51.92%/40.44% and the number of training violations by 90.60%/77.97% compared to FOCOPS/CPO.

### *C. Results and Analysis of Congestion Management*

The control group in Figs. 7-10 is trained by the SAC DRL method whose reward function is only CS profits with the same action preset range of the proposed AMSDRL. The maximum line load percentage is set to 135%. The SAC algorithm is a state-of-the-art reward-driven algorithm developed for maximum entropy DRL. In theory, SAC has a better performance than the safe DRL methods for the unconstrained problems because SAC focuses on exploring the action space for the highest reward without considering the cost constraints. Therefore, we choose SAC as the algorithm for the control group with only a positive reward function rather than FOCOPS or CPO.

Fig. 7 presents the real-time charging prices of the three CSs. The pricing strategy of the AMSDRL presents significant stratification between the three CSs. Fig. 8 shows the load percentages of target lines $l_{27,28}, l_{9,10}, l_{1,18}$, where the red dashed line represents the line load percentage limit of 90% and the blue fill represents the congestion error. Congestion is prone to occurring at $l_{9,10}$ in the whole process and at $l_{1,18}$ during the end of the congestion management process. Compared to the control group, the proposed method reduces congestion by up to 66.64% and all the load percentages lie within the range of the operating congestion limits with lower fluctuation. Figs. 9 and 10 show the arrival EV number and active powers of three CSs, which provide an explanation of the results in Fig. 8. We find that the price of CS 2 is always higher than CS1 and 3, which makes EVs shift from CS 2 to CS 1 or 3 during 2:00-4:45 to avoid congestion at $l_{9,10}$. During 4:45-5:00, with the decrease in the price of CS 1 and the increase in the prices of CS 2 and 3, some EVs are encouraged to charge at CS 1 to eliminate the congestion violation at $l_{1,18}$. Note that the CS load includes the charging power of arriving EVs and previously arrived EVs with unsatisfied charging demand, which means that the impact of price is superimposed and the decision process is more complex than in the straightforward analysis.

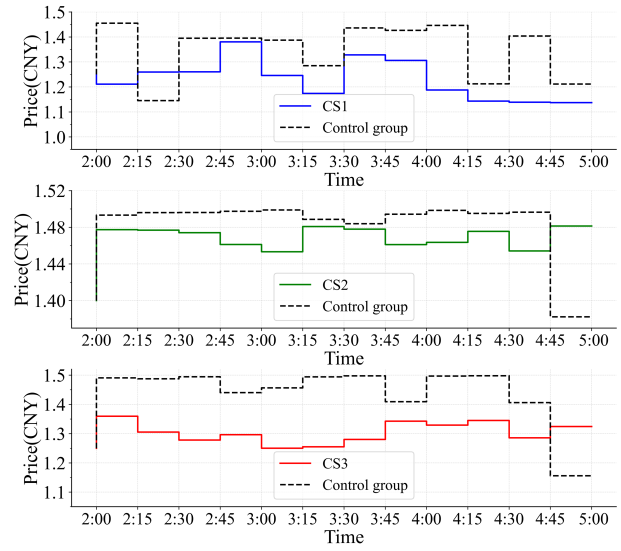



Fig. 7 Real-time price of three CSs.

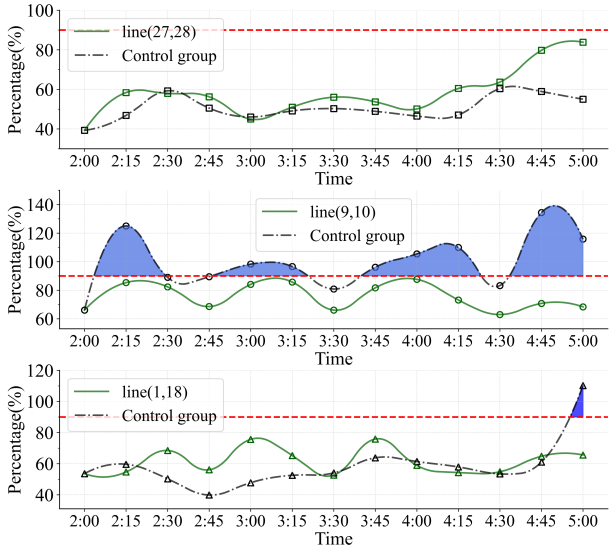



Fig. 8 Load percentage of target lines.

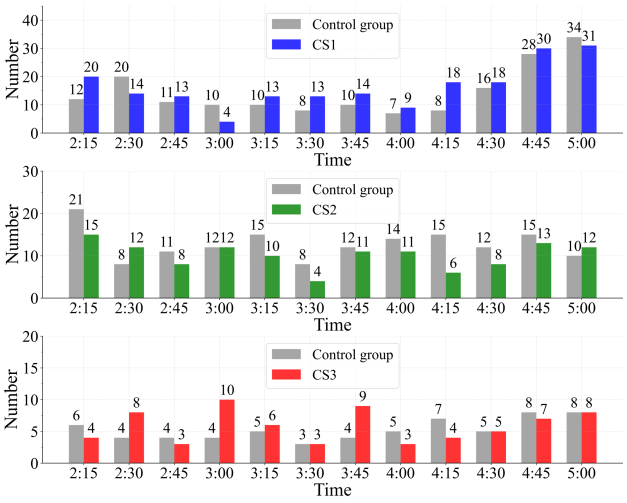



Fig. 9 Number of EVs arriving in each time stage.

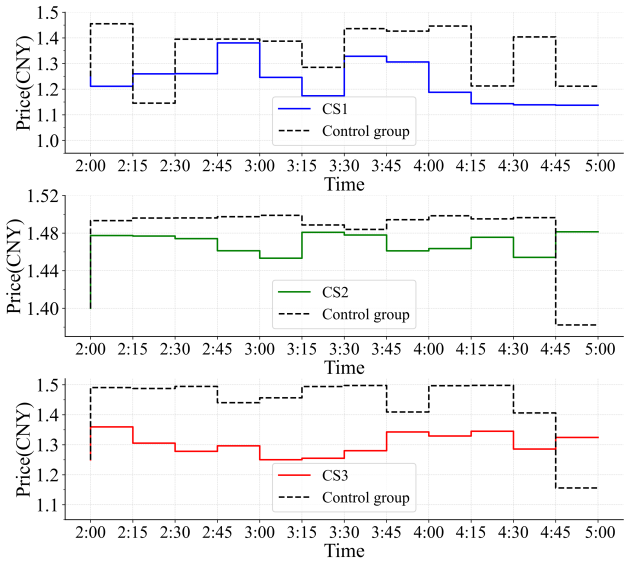



Fig. 10 Active power of three CSs.

The results in terms of social welfare are shown in Fig. 11, where the control groups are trained by the SAC and DDPG algorithms with the same settings but the reward function includes social welfare and a line load percentage limit of 90%. As benchmark unconstrained algorithms, the SAC and DDPG should yield better results in social welfare than AMSDRL. However, for the constrained congestion management problem, the proposed approach achieves total social welfare that is higher by 15.06% and 17.74% than that of the SAC and DDPG, respectively. The reason is that the SAC and DDPG need a high negative reward to achieve a strict congestion constraint, which weakens the ability to explore and causes the conservative pricing to prevent congestion.

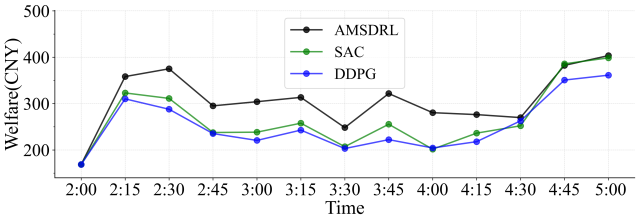



Fig. 11 Comparisons of different DRL methods in terms of social welfare..

### *D. Scalability Verification*

The EV dispatching capacity is mainly limited by the energy and time constraints, and EV planning is generally applicable to relatively small regions in the transportation network. So we test the method performance for larger systems in terms of nodes and EVs. The simulations were performed on the IEEE 118-bus distribution system to verify the scalability of the proposed method. As shown in Fig. 12, there are six CSs at the buses 49, 53, 70, 77, 106, and 113. The number of electric vehicles with charging demand quadruple. Power lines $l_{35,47}, l_{66,67}, l_{102,103}$ are prone to congestion with corresponding line capacities of 30, 26, and 60 (MW). The detailed results are presented in Table VI.

Compared with the simulation results on the IEEE 33-bus system, similar conclusions can also be drawn for the IEEE 118-bus system with only a 13.12%/7.93% increase in training/loading time. Compared to the FOCOPS and CPO, AMSDRL reduces the training time by 57.12% and 45.30%, respectively, on the IEEE 118-bus distribution network. This is because AMSDRL uses the constrained cross-entropy method to directly calculate the distributions of actions rather than using a policy network to output the sample policy. The computational time mainly depends on population size. An increase in the numbers of EVs and traffic nodes will not significantly increase the training time or loading time in theory.

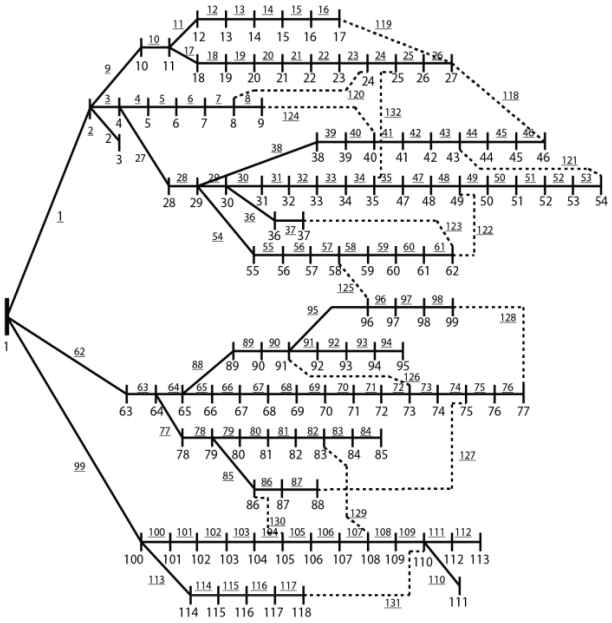


Fig. 12 Illusion of the transportation network and IEEE 118-bus distribution system for scalability verification.

TABLE VI RESULTS OF SIMULATIONS ON IEEE 118-BUS DISTRIBUTION NETWORK

| Items | Value |
|---|---|
| Average/ Maximum congestion reduction of the three lines(kW) | 15.25/34.32 |
| Maximum reduction of line load percentage (%) | 57.17 |
| Social welfare increase (%) | 17.61 |
| Time to convergence of AMSDRL/CPO/FOCOPS (min) | 80.17/146.56/186.97 |
| Loading time for the execution (s) | 33.61 |

## V. CONCLUSION

A model-based safe DRL method is proposed to solve the real-time EV congestion management problem with coupled power and transportation networks. The congestion management problem is formulated as a CMDP and solved by the AMSDRL algorithm. Numerical studies conducted on modified IEEE 33-bus and 118-bus distribution systems with real historical data demonstrate the effectiveness of the proposed method. The following conclusions can be drawn:

1) The proposed method significantly prevents all line congestion with lower fluctuations and reduces the maximum line load percentage by 66.64%, while improving the total social welfare by 15.06% compared with the control group.

2) Compared with FOCOPS/CPO, the proposed AMSDRL algorithm achieves a better convergence reward, safer constraint satisfaction with only a 0.89%/0.27% cost during

convergence, higher training efficiency with a 51.92%/40.44% reduction in training time and greater model stability with a 90.60%/77.97% reduction in violations.

3) The method is scalable. The training/loading time of the proposed method on the IEEE 118-bus large distribution network increases by only 13.12%/7.93% compared to the training results of the IEEE 33-bus system. Compared to the FOCOPS and CPO, AMSDRL reduces the training time by 57.12% and 45.30%, respectively, on the IEEE 118-bus distribution network.

## APPENDIX

Proof of the conclusion in (51):

Let $C_j^E$ be the expected cost from executing the first $j$ steps of $\pi$ and then switching to $T$ for the remainder, as follows:

$$C_j^E = \mathop{\mathrm{E}}_{\substack{a_t\sim\pi(s_t)\\ t<j:s_{t+1}\sim T(s_t,a_t)\\ t\geq j:s_{t+1}\sim T(s_t,a_t)}}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right].$$

Note that $\hat{J}_c(\pi)=C_\infty^E$ and $J_c(\pi)=C_0^E$, then

$$J_c(\pi)-J_c(\pi)=\sum_{j=0}^{\infty}(C_{j+1}^E-C_j^E).$$

$C_{j+1}^E$ can $C_j^E$ be expanded as

$$C_{j+1}^E = C_j + \mathop{\mathrm{E}}_{s_j,a_j\sim\pi,T}\left[\mathop{\mathrm{E}}_{s'\sim T(s_j,a_j)}\gamma^{j+1}[V_c^\pi(s')]\right]$$

$$C_j^E = C_j + \mathop{\mathrm{E}}_{s_j,a_j\sim\pi,T}\left[\mathop{\mathrm{E}}_{s'\sim T(s_j,a_j)}\gamma^{j+1}[V_c^\pi(s')]\right].$$

After substitution, we have the conclusion 1 as:

$$\begin{aligned}
&J_c(\pi)-J_c(\pi)\\
&=\sum_{j=0}^{\infty}\gamma^{j+1}\mathop{\mathrm{E}}_{s_j,a_j\sim\pi,T}\left[\mathop{\mathrm{E}}_{s'\sim T(s_j,a_j)}\gamma^{j+1}[V_c^\pi(s')]-\mathop{\mathrm{E}}_{s'\sim T(s_j,a_j)}\gamma^{j+1}[V_c^\pi(s')]\right]\\
&=\gamma\mathop{\overline{\mathrm{E}}}_{(s,a)\sim\rho_T^\pi}\left[\mathrm{E}_{s'\sim T(s,a)}\left[V_c^\pi(s')\right]-\mathrm{E}_{s'\sim T(s,a)}\left[V_c^\pi(s')\right]\right].
\end{aligned}$$

The CMDP can be solved by considering its dual problem, the dual formulation of (44) can be written as follows:

$$\begin{aligned}
\max_{\rho(s,a)\geq 0}\quad & \frac{1}{1-\gamma}\sum_{s,a}\rho(s,a)r(s,a)\\
\text{s.t.}\quad & \frac{1}{1-\gamma_c}\sum_{s,a}\rho(s,a)c(s,a)\leq C\\
& \sum_{s,a}\rho(s,a)=(1-\gamma_c)\mu_0(s)+\gamma_c\sum_{s',a'}T(s\mid s',a')\rho(s',a'),
\end{aligned}$$

where the second constraint defines the space of valid occupancy distributions by ensuring the flow conservation equilibrium between distributions.

The integral probability metric (IPM) associated with the measurable real-valued functions $\mathcal{F}$ on full set $\chi$, i.e., the difference between the true and learned transitions can be defined as follows:

$$\begin{aligned}
d_{\mathrm{F}}(T(s,a),T(s,a)) &:= \sup_{f\in\mathrm{F}}\left|\int_\chi f dT-\int_\chi f dT\right|\\
&=\sup_{f\in\mathrm{F}}\left|\mathrm{E}_{s'\sim T(s,a)}\left[f(s')\right]-\mathrm{E}_{s'\sim T(s,a)}\left[f(s')\right]\right|
\end{aligned}$$

Using the conclusion 1 and the definition of the integral probability metric, we can obtain the conclusion of (47) as follows:

$$\begin{aligned}
&J_c(\pi)-J_c(\pi)\\
&=\frac{1}{1-\gamma}\sum_{s,a}(\rho_T^\pi(s,a)-\rho_T^\pi(s,a))c(s,a)\\
&=\gamma\sum_{s,a}\rho_T^\pi(s,a)\left[\mathrm{E}_{s'\sim T(s,a)}\left[V_c^\pi(s')\right]-\mathrm{E}_{s'\sim T(s,a)}\left[V_c^\pi(s')\right]\right]\\
&\leq\gamma\sum_{s,a}\rho_T^\pi(s,a)\sup_{f\in\mathrm{F}}\left|\mathrm{E}_{s'\sim T(s,a)}\left[f(s')\right]-\mathrm{E}_{s'\sim T(s,a)}\left[f(s')\right]\right|\\
&=\gamma\sum_{s,a}\rho_T^\pi(s,a)d_{\mathrm{F}}(T(s,a),T(s,a))\\
&\leq d_{\mathrm{F}}(T(s,a),T(s,a)).
\end{aligned}$$

## REFERENCES


[1] Á. Paredes, J. A. Aguado and P. Rodrí-guez, "Uncertainty-aware trading of congestion and imbalance mitigation services for multi-dso local flexibility markets," *IEEE Trans. Sustain. Energy*, doi: 10.1109/TSTE.2023.3257405.

[2] P. Zhuang and H. Liang, "Stochastic energy management of electric bus charging stations with renewable energy integration and B2G capabilities," *IEEE Trans. Sustain. Energy*, vol. 12, no. 2, pp. 1206–1216, Apr. 2021.

[3] S. Huang, Q. Wu, M. Shahidehpour, and Z. liu, "Dynamic power tariff for congestion management in distribution networks," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2148 2157, Mar. 2019.

[4] S. J. Huang and Q. W. Wu, "Dynamic subsidy method for congestion management in distribution networks," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 2140–2151, May 2018.

[5] J. Hu, X. Liu, M. Shahidehpour, and S. Xia, "Optimal operation of energy hubs with large-scale distributed energy resources for distribution network congestion management," *IEEE Trans. Sustain. Energy*, vol. 12, no. 3, pp. 1755–1765, Jul. 2021.

[6] S. Deb *et al*., "Charging coordination of plug-in electric vehicle for congestion management in distribution system integrated with renewable energy sources," *IEEE Trans. Ind. Appl.*, vol. 56, no. 5, pp. 5452–5462, Sep./Oct. 2020.

[7] A. Ali, D. Raisz, K. Mahmoud, and M. Lehtonen, "Optimal placement and sizing of uncertain PVs considering stochastic nature of PEVs," *IEEE Trans. Sustain. Energy*, vol. 11, no. 3, pp. 1647–1656, Jul. 2020.

[8] K. Knezovi´c *et al*., "Enhancing the role of electric vehicles in the power grid: Field validation of multiple ancillary services," *IEEE Trans. Transp. Electrific.*, vol. 3, no. 1, pp. 201–209, Mar. 2017.

[9] N. Y. Soltani, S. -J. Kim and G. B. Giannakis, "Real-Time load elasticity tracking and pricing for electric vehicle charging," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1303-1313, May 2015.

[10] H. Saber, *et al*., "Network-constrained transactive coordination for plug-in electric vehicles participation in real-time retail electricity markets," *IEEE Trans. Sustain. Energy*, vol. 12, no. 2, pp. 1439–1448, Apr. 2021.

[11] S. Lai, J. Qiu, Y. Tao and J. Zhao, "Pricing for electric vehicle charging stations based on the responsiveness of demand," *IEEE Trans. Smart Grid*, vol. 14, no. 1, pp. 530-544, Jan. 2023.

[12] M. Alizadeh *et al*., "Optimal pricing to manage electric vehicles in coupled power and transportation networks," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 4, pp. 863–875, Dec. 2017.

[13] Y. Cui, Z. Hu and X. Duan, "Optimal pricing of public electric vehicle charging stations considering operations of coupled transportation and power systems," *IEEE Trans. Smart Grid*, vol. 12, no. 4, pp. 3278-3288, July 2021.

[14] A. Thangaraj, S.A.E. Xavier, R. Pandian, "Optimal coordinated operation scheduling for electric vehicle aggregator and charging stations in integrated electricity transportation system using hybrid technique", *Sustain. Cities Soc.*, vol. 80, Oct, 2020, Art. no. 103768.

[15] Z. Ding et al., "Optimal coordinated operation scheduling for electric vehicle aggregator and charging stations in an integrated electricity transportation system," *Int. J. Elect. Power Energy Syst.*, vol. 121, pp. 1–11, Oct. 2020.

[16] S. Huang and Q. Wu, "Real-time congestion management in distribution networks by flexible demand swap," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4346–4355, 2018.
[17] T. Qian, C. Shao, X. Li, *et al.*, "Enhanced coordinated operations of electric power and transportation networks via EV charging services," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3019–3030, Jul. 2020.
[18] L. Yan, X. Chen, J. Zhou, *et al.*, "Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5124–5134, 2021.
[19] D. Qiu *et al*., "A deep reinforcement learning method for pricing electric vehicles with discrete charging levels," *IEEE Trans. Ind. Appl.*, vol. 56, no. 5, pp. 5901–5912, Sep./Oct. 2020.
[20] P. Xu et al., "Real-time fast charging station recommendation for electric vehicles in coupled power-transportation networks: A graph reinforcement learning method," *Int. J. Elect. Power Energy Syst.*, vol. 141, Oct. 2022, Art. no. 108030.
[21] H. Cha, M. Chae, M. A. Zamee and D. Won, "Operation Strategy of EV Aggregators Considering EV Driving Model and Distribution System Operation in Integrated Power and Transportation Systems," *IEEE Access*, vol. 11, pp. 25386-25400, 2023.
[22] 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.
[23] 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, Jul. 2020.
[24] Q. Zhang, K. Dehghanpour, Z. Wang, F. Qiu, and D. Zhao, "Multi-agent safe policy learning for power management of networked microgrids," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1048–1062, Mar. 2021.
[25] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," *in Proc. 34th Int. Conf. Mach. Learn.*, pp. 22–31, 2017.
[26] Y. Zhang, Q. Vuong, and K. Ross, "First order constrained optimization in policy space," *in Proc. 34th Neural Inf. Process. Syst.*, vol. 33, pp. 15338–15349, Dec, 2020.
[27] X. Li *et al.*, "Price incentive-based charging navigation strategy for electric vehicles," *IEEE Trans. Ind. Appl.*, vol. 56, no. 5, pp. 5762–5774, Sep./Oct. 2020.
[28] S. M. Ross, Introduction to Probability Models, 10th ed. Amsterdam, The Netherlands: Academic Press, 2010.
[29] T. Yu, G. Thomas, L. Yu, *et al*., "Mopo: Model-based offline policy optimization," *in Proc. 34th Neural Inf. Process. Syst.*, vol. 33, pp. 14129–14142, Dec 2020.
[30] Y. J. Ma, A. Shen, O. Bastani, and J. Dinesh. Conservative and adaptive penalty for model-based safe reinforcement learning. *in Proc. 36th AAAI Conf. Artif. Intell.*, vol. 36, no.5, pp. 5404-5412, Jun, 2022.
[31] A. Asrari, M. Ansari, J. Khazaei, and P. Fajri, "A market framework for decentralized congestion management in smart distribution grids considering collaboration among electric vehicle aggregators," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1147–1158, Mar. 2020.
[32] W. Shu, K. Cai, and N. N. Xiong, "A short-term traffic flow prediction model based on an improved gate recurrent unit neural network," *IEEE Trans. Intell. Transp. Syst.*, early access, Jul., 2021.
[33] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, Y. Yang, and A. Knoll, "A review of safe reinforcement learning: Methods, theory and applications," *arXiv preprintarXiv: 2205.10330*, 2022.