

# Federated Reinforcement Learning for Electric Vehicles Charging Control on Distribution Networks

Junkai Qian, *Graduate Student Member, IEEE*, Yuning Jiang, *Member, IEEE*,

Xin Liu, *Member, IEEE*, Qiong Wang, Ting Wang, *Senior Member, IEEE*,

Yuanming Shi, *Senior Member, IEEE*, and Wei Chen, *Senior Member, IEEE*

**Abstract**—With the growing popularity of electric vehicles (EVs), maintaining power grid stability has become a significant challenge. To address this issue, EV charging control strategies have been developed to manage the switch between vehicle-to-grid (V2G) and grid-to-vehicle (G2V) modes for EVs. In this context, multi-agent deep reinforcement learning (MADRL) has proven its effectiveness in EV charging control. However, existing MADRL-based approaches fail to consider the natural power flow of EV charging/discharging in the distribution network and ignore driver privacy. To deal with these problems, this paper proposes a novel approach that combines multi-EV charging/discharging with a radial distribution network (RDN) operating under optimal power flow (OPF) to distribute power flow in real time. A mathematical model is developed to describe the RDN load. The EV charging control problem is formulated as a Markov Decision Process (MDP) to find an optimal charging control strategy that balances V2G profits, RDN load, and driver anxiety. To effectively learn the optimal EV charging control strategy, a federated deep reinforcement learning algorithm named FedSAC is further proposed. Comprehensive simulation results demonstrate the effectiveness and superiority of our proposed algorithm in terms of the diversity of the charging control strategy, the power fluctuations on RDN, the convergence efficiency, and the generalization ability.

**Index Terms**—V2G, Electrical Vehicle, Optimal Power Flow, Reinforcement Learning, Federated Learning

## NOMENCLATURE

### Parameters

$\beta_i^1, \beta_i^2$	Parameter of anxiety SoC
$\delta$	Learning rate for Q network, policy network and temperature coefficient updates

The first two authors contributed equally. This work was supported in part by the Natural Science Foundation of Shanghai under Grant 21ZR1442700, in part by the Shanghai Sailing Program under Grant 22YF1428500, and in part by the Swiss National Science Foundation under the NCCR Automation (grant agreement 51NF40\_180545). (Corresponding authors: Ting Wang and Xin Liu.)

Junkai Qian and Ting Wang are with the MoE Engineering Research Center of Software/Hardware Co-design Technology and Application, the Shanghai Key Lab. of Trustworthy Computing, East China Normal University, China (e-mail: 51255902137@stu.ecnu.edu.cn, twang@sei.ecnu.edu.cn)

Yuning Jiang is with the Automatic Control Laboratory, EPFL, Switzerland (e-mail: yuning.jiang@ieee.org).

Xin Liu and Yuanming Shi are with the School of Information Science and Technology, ShanghaiTech University, China (e-mail: liuxin7, shiym@shanghaitech.edu.cn).

Qiong Wang is with the State Grid Beijing Electric Power Company, China (email: wangqiong@bj.sgcc.com.cn)

Wei Chen is with the Department of Electronic Engineering, Tsinghua University, China (e-mail: wchen@tsinghua.edu.cn).

$\eta_c/\eta_d$	Energy transfer efficiency for EV charging and discharging
$\gamma$	Reward discount parameter
$\kappa$	Weight coefficient for balancing TA and RA
$\lambda$	Weight coefficient for power, anxiety and grid reward
$\bar{a}_i^{\text{G2V}}/\bar{a}_i^{\text{V2G}}$	Upper bound of charging and discharging rate for EV $i$
$\bar{l}_{se}$	Upper bound of the square of current magnitude from bus $s$ to $e$
$\bar{P}_i^{\text{G2V}}/\bar{P}_i^{\text{V2G}}$	Upper bound of charging and discharging power for EV $i$
$\tau$	Learning rate for target Q network updates
$\underline{v}_s/\bar{v}_s$	Lower and upper bound of the square of voltage magnitude at bus $s$
$b_{se}$	Susceptance on the line $(s, e)$
$C_i$	Battery capacity of EV $i$
$p_{s,t}$	Active power injection at bus $s$ at time slot $t$
$q_{s,t}$	Reactive power injection at bus $s$ at time slot $t$
$r_{se}$	Resistance on the line $(s, e)$
$x_{se}$	Reactance on the line $(s, e)$
<i>Sets</i>	
$\mathcal{L}$	Set of transmission lines in RDN
$\mathcal{V}$	Set of bus in RDN
$\mathcal{Z}_s$	Set of EVs' indexes connected to bus $s$
$\Omega_s$	Set of children buses of bus $s$ in RDN
<i>Random Variables</i>	
$\text{SoC}_{i,t}^x$	Expected SoC for EV $i$ at time slot $t$
$\text{SoC}_i^d$	Estimation SoC of the coming travel for EV $i$
$\xi_t$	Electricity price at time slot $t$
$t_i^a/t_i^d$	Arriving time and departure time of EV $i$
$v_{s,t}$	Square of voltage magnitude at bus $s$ at time slot $t$
<i>Decision Variables</i>	
$\alpha$	Temperature coefficient
$\epsilon$	Reparameterization parameter for policy network
$\phi$	Parameter of actor network
$\sigma_g$	Standard deviation of power change in the test
$\text{SoC}_{i,t}$	State of charge for EV $i$
$\theta^k/\hat{\theta}^k$	Parameter of critic and target critic network
$a_{i,t}$	Charging/discharging rate for EV $i$ at time slot $t$
$p_{0,t}$	Active power obtained from substation at time slot $t$
$P_{agg,s,t}$	Active power on aggregator $s$ at time slot $t$
$P_{i,t}$	Active charging/discharging power of EV $i$ at time slot $t$

$P_{se,t}$	Active power from bus $s$ to bus $e$ at time slot $t$
$Q_{se,t}$	Reactive power from bus $s$ to bus $e$ at time slot $t$
$s_{i,t}$	State of EV <i>i</i> at time slot $t$

## I. INTRODUCTION

ELCTRIC vehicles (EVs) offer a promising alternative to conventional fossil fuel-powered vehicles, with their potential to mitigate greenhouse gas (GHG) emissions and air pollution [1]. Recent projections from the International Energy Agency (IEA) indicate that the EV market share has substantially increased in recent years and is predicted to reach approximately 30% by 2030 [2]. However, the increase in EVs also poses new challenges for the power grid, including burdening the electricity loads and amplifying the peak electricity demands [3]–[5]. [6] demonstrated that even a 10% increase in EVs could result in significant fluctuations in the grid's voltage curves.

Vehicle-to-grid (V2G) and grid-to-vehicle (G2V) are considered effective techniques to minimize the negative impact of EV charging on the grid [7], [8]. V2G involves discharging an EV back into the power grid, while G2V refers to charging an EV from the grid. The V2G/G2V mode in a distribution network can help stabilize the power curve [9]–[11] and provide peak shaving/valley filling services [12]. However, the uncertain factors, including dynamic electricity prices, grid load, and uncertainty in human behavior, increase the difficulty of generating a generalized charging policy [13]–[15]. Consequently, the critical challenge lies in effectively coordinating EVs' charging and discharging behavior in a dynamic environment to enable their active participation in electricity delivery [16].

Numerous approaches have been proposed to develop optimal EV charging control policies. Traditional studies have formulated charging control as an optimization problem that aims to maximize the drivers' profits [17]. For instance, the work [18] proposes a two-stage charging optimization model to reduce the operation cost at the workplace. In [19], the capacity limitation in distribution transformers is considered, and the continuous EV charging is formulated as linear programming. The authors of [20] investigate energy trading in coordination with V2G and propose a solution using mixed-integer stochastic linear programming. [21] designs a cost function with battery energy constraints, and the optimal EV charging control is computed through dynamic programming. The work [22] introduces the uncertainty in wind energy supply and formulates the EV charging control as a Markov Decision Process (MDP). Reference [23] considers the customers' dynamic preference for charging parameters and conducts a stochastic game to address the uncertainty. Reference [24] introduces time anxieties to address uncertain events and solves the EV charging problem using a generalized Nash equilibrium (NE) game. In summary, all the algorithms mentioned above maintain an explicit optimization model under the assumption of a fully observable environment called model-based algorithms. The optimal charging control strategy is generated based on an accurate system model. However, fully observable environments rely on a priori knowledge

about state transitions, and uncertain factors in the environment can interfere with state transitions, which further affects the model's accuracy.

Model-free deep reinforcement algorithm (DRL) has demonstrated great potential in sequential decision-making problems without prior knowledge of the environment's dynamics [25]–[27]. In DRL, the agent makes a decision for each state to obtain the reward sequence and then learns to optimize its policy to maximize expected future rewards. DRL algorithms have been widely adopted in grid management and have achieved excellent performance [28], including in EV charging control. In [29], discriminative features of electricity prices are extracted through a representation network, and Q-learning is developed to obtain the optimal discrete EV charging control strategy. Reference [30] utilizes the long short-term memory (LSTM) method to extract the trend of electricity prices. The estimated electricity price is applied in a deep deterministic policy gradient (DDPG) algorithm for continuous EV charging control policy. In [31], the traffic conditions are considered to minimize the travel time and charging costs during charging navigation. [32] formulates the EV charging control problem as a constrained MDP and develops a safe DRL method to learn the optimal scheduling strategy. In [33], the authors mathematically model the driver's anxiety using statistical principles and propose using a soft actor-critic (SAC) to learn the optimal policy. However, EV charging behaviors in the same system tend to affect each other, where such mutual influence is often ignored in single-agent DRL-based approaches. Multi-agent deep reinforcement learning (MADRL) algorithms are more suitable for addressing EV charging control problems than single-agent ones [34]. For example, the work [35] promotes the type of tariff and develops a multi-agent multi-objective reinforcement learning framework to optimize the charging schedule. In [36], a multi-agent SAC-based framework is proposed to learn the environment's dynamics, and its performance is verified in a simulated distribution network. In [37], the authors study the transformer constraints in a residential system and approximate the collective behaviors of EVs through a collective-policy model, using multi-agent SAC to learn the charging scheduling strategy. Overall, model-free DRL-based methods have proven to be superior in addressing uncertainty in EV charging control.

Nevertheless, the approaches mentioned above have two drawbacks. Firstly, while some advanced methods (e.g., [19], [36], [37]) combine EV charging control with the distribution network and consider a systematic optimization process, they usually use simplified constraints as a substitute for complex power flow in the distribution network, which is insufficient and unreliable. Moreover, the substitute electricity constraints are considered an optimization objective, which is often optional, leading to the possibility of irreversible damage to the distribution network during the EV charging control. Secondly, the charging and travel data of each EV is privacy-sensitive, raising concerns about driver privacy when sharing massive data in MADRL. To address this, some latest MADRL algorithms (e.g., [37]) maintain a charging control strategy with individual charging experience to avoid privacy breaches.

However, the isolated data of each EV limit the generalization of the strategy [38]–[40].

To tackle the above issues, we propose an approach that leverages optimal power flow (OPF) [41] to model radial distribution networks (RDNs) [42] and federated learning (FL) [43]–[47] for charging control policy training. Specifically, we study a multi-EV charging and discharging system on an RDN, which operates under OPF and aims to allocate power flows to minimize power loss. Each EV is controlled by an agent that deploys a soft actor-critic (SAC) model inspired by [33]. SAC has achieved excellent effects in continuous action space. It uses an entropy regularization method, which can effectively control the exploration and stability of the policy. Each agent receives regulation information from the RDN and executes local model training to maximize rewards, including V2G profits, grid idle degree, and driver satisfaction. The training process follows the FL mechanism, where agents upload their local model parameters to the server via a communication link. The server aggregates these parameters to refine a global model that enhances the generalization of the charging control strategy. As only model parameters are transferred during transmission, and local data is stored on each EV, drivers' privacy is guaranteed [48]–[50].

To summarize, this paper makes the following main contributions:

- We introduce an RDN that operates under OPF to combine with a multi-EV charging control system. By considering electric component constraints, we simulate the EV charging/discharging power flow on the RDN, resulting in more solid and realistic constraints than quantitative approaches.
- We develop a grid reward with OPF-based RDN, introducing a new optimization objective, i.e., grid idle degree, in the case of unbalanced charging and discharging rates. The EV agents trained through the grid reward will sequentially charge/discharge when the grid is idle, significantly reducing the power fluctuations on the RDN.
- We model the EV charging control problem as a multi-agent MDP and propose a continuous federated SAC algorithm called FedSAC. Compared to previous standard DRL methods, our FedSAC approach achieves the best control performance.

The rest of this paper is organized as follows. Section II presents the EV charging/discharging model on an OPF-based RDN and models the EV charging/discharging process as an MDP. Section III introduces the overall framework of the proposed FedSAC. In Section IV, simulations are conducted to verify the performance of the proposed algorithm. Section V concludes the paper.

## II. THE DISTRIBUTION NETWORK MULTI-EV G2V/V2G MODEL

This paper considers the radial distribution network (RDN) integrated multi-EV G2V/V2G model. Each EV is equipped with a G2V/V2G management module, which works as an agent to conduct model training and control the charging process. In this section, we first combine EV charging/discharging

with a radial distribution network and introduce OPF to model the system. Then, the driver's charging anxiety is modeled to consider the different habits of the driver. Finally, the charging control process is formulated as an MDP with hourly dynamic electricity prices.

### A. The Radial Distribution Network Model

The radial distribution network refers to the power network that receives electricity from the transmission network or regional power plants. Through distribution facilities, it distributes electricity to various users regionally or level by level according to voltage.

As discussed in [51, Section II.A], we consider the AC power flow on RDN. To this end, we define the topology of RDN by  $(\mathcal{V}, \mathcal{L})$ , where  $\mathcal{V} := \{0, \dots, v\}$  denotes the set of buses. Without loss of generality, the substation is denoted as bus 0,  $\mathcal{L} \subseteq \mathcal{V} \times \mathcal{V}$  denotes the set of transmission lines. Then, let  $p_{s,t}$  and  $q_{s,t}$  denote the active and reactive power injection at bus  $s$  at time slot  $t$ . For the transmission line  $(s, e) \in \mathcal{L}$ , we define by  $r_{se}$  and  $x_{se}$  the resistance and the reactance, respectively and by  $b_{se}$  the total charging susceptance. Moreover,  $P_{se,t}$  and  $Q_{se,t}$  denote the active and reactive power from bus  $s$  to  $e$  during  $[t, t+1]$ , respectively. Thus, the AC power equation in a radial distribution network can be described as

$$P_{se,t} = p_{s,t} + P_{agg_{s,t}} + \sum_{h \in \Omega_s} (P_{hs,t} - r_{hs}l_{hs,t}), \quad (1a)$$

$$Q_{se,t} = q_{s,t} + \sum_{h \in \Omega_s} (Q_{hs,t} - x_{hs}l_{hs,t}), \quad (1b)$$

where  $\Omega_s$  denotes the index set of children neighbors of bus  $s$ , and  $P_{agg_{s,t}}$  is the charging power of the aggregator connected to bus  $s$  during  $[t, t+1]$ . Moreover,  $l_{hs}$  is the square of current magnitude, i.e.,

$$l_{se,t} = \frac{P_{se,t}^2 + Q_{se,t}^2}{v_{s,t}}, \quad (2)$$

where  $v_{s,t} \in [\underline{v}_s, \bar{v}_s]$  is the square of voltage magnitude at bus  $s$  during  $[t, t+1]$ . Then, the equations of power flow from  $s$  to  $e$  are represented as

$$v_{s,t} - v_{e,t} = 2(r_{se}P_{se,t} + x_{se}Q_{se,t}) - l_{se,t}(r_{se}^2 + x_{se}^2). \quad (3)$$

In addition, flows in the branch's charging susceptance should also be considered. Thus, the capacity limits on all branches during  $[t, t+1]$  are defined as

$$l_{se,t} + \frac{1}{4}v_{s,t}b_{se}^2 + b_{se}Q_{se,t} \leq \bar{l}_{se}, \quad (4a)$$

$$l_{se,t} + \frac{1}{4}v_{e,t}b_{se}^2 + b_{se}(x_{se}l_{se,t} - Q_{se,t}) \leq \bar{l}_{se}. \quad (4b)$$

### B. EV Charging Model on RDN

Let us consider  $N$  EVs with the same battery and charging infrastructures. Each EV can work in both G2V and V2G modes. We assume that EV $i$  connects the charging pile at time  $t_i^a$  and departs at time  $t_i^d$ . The State of Charge (SoC) of EV $i$  at time  $t$  is denoted as  $\text{SoC}_{i,t}$ , which satisfies  $\text{SoC}_{i,t} \in [0, 1]$ .

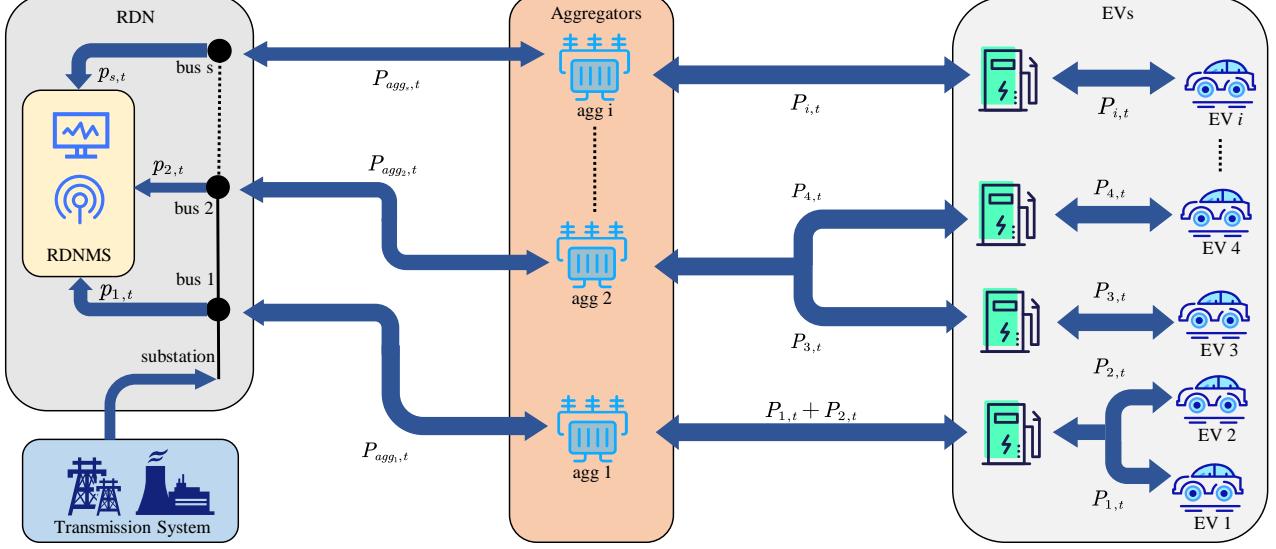


Fig. 1. The structure of EVs connected to RDN

The charging and discharging process of  $EV_i$ 's battery at time slot  $t$  is modeled as

$$\text{SoC}_{i,t+1} = \begin{cases} \text{SoC}_{i,t} & t < t_i^a, t \geq t_i^d, \\ \text{SoC}_{i,t} + a_{i,t} & t_i^a \leq t < t_i^d, \end{cases} \quad (5)$$

where  $a_{i,t}$  is the continuous charging/discharging rate at time slot  $t$ , i.e.,  $a_{i,t} > 0$  when  $EV_i$  work in G2V mode and  $a_{i,t} < 0$  when  $EV_i$  work in V2G mode.

The actual power  $EV_i$  absorbed/released from RDN [36] during  $[t, t+1]$  is calculated as

$$P_{i,t} = \begin{cases} P_{i,t}^{\text{G2V}} & a_{i,t} > 0 \\ -P_{i,t}^{\text{V2G}} & a_{i,t} \leq 0 \end{cases} \quad (6)$$

$$= \begin{cases} \frac{a_{i,t} C_i}{\eta_c} & a_{i,t} > 0 \\ -\eta_d a_{i,t} C_i & a_{i,t} \leq 0, \end{cases}$$

where  $\eta_c(\eta_d)$  is the (dis-)charging efficiency, and  $C_i$  denotes the actual battery capacity of  $EV_i$ .

The aggregator is a middleware between RDN and EVs responsible for receiving charging/discharging power and transmitting control signals. Thus, the power of the aggregator  $s$  is equal to the sum of EVs managed by it:

$$P_{agg_s,t} = \sum_{i \in \mathcal{Z}_s} P_{i,t} \quad (7)$$

$$= \sum_{i \in \mathcal{Z}_s} P_{i,t}^{\text{G2V}} - \sum_{i \in \mathcal{Z}_s} P_{i,t}^{\text{V2G}}$$

$$= P_{agg_s,t}^{\text{G2V}} - P_{agg_s,t}^{\text{V2G}},$$

where  $\mathcal{Z}_s$  denotes the set of EVs' indexes connected to bus  $s$  and  $P_{agg_s,t}^{\text{G2V}}/P_{agg_s,t}^{\text{V2G}}$  denotes the total charging/discharging rate of EVs on the aggregator  $s$ . Thus,  $P_{agg_s,t}$  can be calculated by the charging and discharging rate of EVs connected to aggregator  $s$ . Additionally, the constraint of charging/discharging power is denoted as

$$-\bar{P}_i^{\text{V2G}} \leq P_{i,t} \leq \bar{P}_i^{\text{G2V}}, \quad (8)$$

where the  $\bar{P}_i^{\text{V2G}}$  and  $\bar{P}_i^{\text{G2V}}$  are the maximum discharging and charging power, respectively.

### C. OPF Formulation on RDN

When the structural parameters and load conditions of RDN are given, the objective of the OPF is to minimize power loss by rationally distributing the power flow. The power loss at time  $t$  is defined as

$$P_t^{\text{loss}} = P_t^{\text{sub}} + P_t^{\text{V2G}} - P_t^{\text{G2V}} - P_t^{\text{others}},$$

where  $P_t^{\text{sub}}$  denotes the power obtained from substation and  $P_t^{\text{others}}$  denotes the other load on RDN.  $P_t^{\text{V2G}}$  denotes the power absorbed from the discharging EVs, and  $P_t^{\text{G2V}}$  denotes the power cost by the charging EVs, i.e.,

$$P_t^{\text{V2G}} = \sum_{s \in \mathcal{V}} P_{agg_s,t}^{\text{V2G}}, \quad P_t^{\text{G2V}} = \sum_{s \in \mathcal{V}} P_{agg_s,t}^{\text{G2V}}.$$

In the article, we assume the other load  $P_t^{\text{others}}$  remains constant over time for simplification. The charging/discharging power  $P_t^{\text{G2V}}$  and  $P_t^{\text{V2G}}$  are managed by individual EVs. They are fixed for RDN at every timeslot. Therefore, for optimization problems in RDN, the  $P_t^{\text{G2V}}$  and  $P_t^{\text{V2G}}$  can be treated as parametric input. In other words, the OPF problem is equivalent to minimizing the energy obtained from the substation at time  $t$ . The objective function to be optimized can be written as

$$\text{minimize } f_{\text{obj},t}(P_{agg,t}) := P_t^{\text{loss}} \quad (9a)$$

$$\text{subject to } P_{agg,t} \in \mathcal{F}_t := \{(1) - (8) \text{ are satisfied}\}, \quad (9b)$$

where  $P_{agg,t}$  stacks  $P_{agg_s,t}$  for all  $s \in \mathcal{V}$ . We assume the RDN works under OPF all the time in the paper.

In the RDN, the management system is responsible for power flow distribution with OPF and sends the results to EVs through the communication unit. A charging pile can usually provide charging services for multiple EVs. Fig. 1 shows the structure of the RDN. The aggregator acts as a middleware

between RDN and EVs, collecting EVs' power and working as a communication link. The aggregators are connected to the lower V2G charging piles, which can provide charging services for multiple EVs.

#### D. Charging Anxiety Model

This paper further considers the driver's charging anxiety, including the driver's range anxiety (RA) and time anxiety (TA). RA represents the driver's anxiety of depleting the battery, which can be relieved by a higher estimated SoC level at time  $t_i^d$ . However, a higher SoC at  $t_i^d$  will also reduce profits obtained from the V2G process. The accuracy of the driver's SoC estimation increases with their experience. TA represents the driver's anxiety about uncertain factors, which may interrupt the charging process and thus leave insufficient power for upcoming travel. Different drivers have different tolerance for uncertain factors.

The charging anxiety is defined by the following mathematical model as presented in [33], which maps the driver's anxiety to the expected SoC:

$$\text{SoC}_{i,t}^x = \frac{\beta_{1,i} (e^{-\beta_{2,i}(t-t_{a,i})/(t_{d,i}-t_{a,i})} - 1)}{e^{-\beta_{2,i}} - 1} \quad (10)$$

for  $[t_i^x, t_i^d]$ , where  $t_i^x \in [t_i^a, t_i^d]$  is the time when the driver begins to feel anxious, hyperparameters  $\beta_i^1 \in [0, 1]$  and  $\beta_i^2 \in (-\infty, 0) \cup (0, \infty)$  determines the growth trend of anxiety SoC. Specifically, larger  $\beta_i^1$  determines higher SoC at  $t_i^d$  while larger  $\beta_i^2$  leads to reaching the expected SoC earlier.

#### E. Markov Decision Process Formulation

The EV charging control process is modeled as a Markov decision process (MDP), which considers the diverse behavior of different EV drivers. For agent  $i$ , the individual MDP model is formalized by a tuple  $\mathcal{M}_i = \langle \mathcal{S}_i, \mathcal{A}_i, \mathbb{P}_i, \mathcal{R}_i, \gamma \rangle$ , where  $\mathcal{S}_i$  represents the state set,  $\mathcal{A}_i$  represents the action set,  $\mathbb{P}_i$  represents the state transition probability,  $\mathcal{R}_i$  represents the reward set, and  $\gamma$  represents the discount factor.

1) *State*: For agent  $i$ , the state gained from the environment is utilized as input of charging control policy to generate current action. Specifically, the state at time  $t$  is defined as

$$s_{i,t} = (\xi_{i,t-n+1}, \dots, \xi_{i,t}, t_i^d, t_i^x, \text{SoC}_{i,t}, \text{SoC}_{i,t}^x, \text{SoC}_i^d),$$

where  $(\xi_{i,t-n+1}, \xi_{i,t-n+2}, \dots, \xi_{i,t})$  is the hourly electricity price of past  $n$  hours,  $\text{SoC}_{i,t}$  is the EV*i*'s current SoC and  $\text{SoC}_i^d$  equal to  $\beta_i^1$  is the expected SoC when EV*i* leaves. Each charging station has an independent pricing right, so each EV faces different electricity prices simultaneously. However, the pricing of charging stations depends on the procurement cost of electricity, which follows certain laws. Thus, past electricity prices are used to summarize the trend of electricity prices. The  $\text{SoC}_i^d$  depends on the driver's estimation of the coming travel, and the  $\text{SoC}_{i,t}^x$  shown in (10) represents the drivers expected SoC at time  $t$ . The other variables are related to drivers' different charging and travel habits. The variables  $t_i^x$  and  $\text{SoC}_{i,t}^x$  reflect the  $i$ -th driver's anxiety trend over time. Drivers with higher anxiety expect the battery to maintain higher SoC during charging. The variables  $t_i^d$  and  $\text{SoC}_i^d$  depend on the driver's estimation of the upcoming travel plan.

2) *Action*: The action  $a_{i,t}$  denotes the charging/discharging rate of EV*i* at time slot  $t$  with the given  $s_{i,t}$ . The charging rate  $a_{i,t}$  has continuous action space, which is restricted by

$$-\bar{a}_i^{\text{V2G}} \leq a_{i,t} \leq \bar{a}_i^{\text{G2V}}.$$

Here,  $\bar{a}_i^{\text{G2V}}$  and  $\bar{a}_i^{\text{V2G}}$  obtained from (6) and (7) denote the maximum charging and discharging rate respectively.  $a_{i,t} > 0$  when EV*i* works in G2V mode and  $a_{i,t} < 0$  when EV*i* works in V2G mode. In our formulation, the maximum charging rate  $\bar{a}_i^{\text{G2V}}$  and discharging rate  $\bar{a}_i^{\text{V2G}}$  are modeled as an imbalance to suit reality [7], [52]. We assume all EVs have the same battery and charging infrastructures.

3) *State Transition*: The behaviors of different drivers and random factors result in different environments for each EV in our MDP formulation. Each EV*i* thus has different state transition functions  $\mathbb{P}_i$  affected by personalized drivers' behavior, dynamic electricity price changes, and random factors. We use notation  $\mathbb{P}_i(s_{i,t+1}|s_{i,t}, a_{i,t})$  to denote the probability of transition from state  $s_{i,t}$  to  $s_{i,t+1}$  when  $a_{i,t}$  is taken.

4) *Reward*: Compared to [33], we design a reward  $r_{i,t}$  such that the impact of EV charging/discharging on the RDN can be considered. Besides, it also considers dynamic electricity prices and drivers' anxiety. The reward is obtained from the local environment unique to each EV and the global environment shared by all EVs. The local environment is an individual EV with a G2V/V2G management module. The global environment is the RDN comprising various electrical components and microcomputers.

The G2V costs and V2G profits are calculated in the local environment with electricity price and charging rate. Thus we define the power reward of EV*i* at time  $t$  as

$$r_{i,t}^p = -\xi_t \cdot a_{i,t}, \quad t_i^a \leq t < t_i^d. \quad (11)$$

The mathematical model in (10) maps the driver's anxiety to the expected SoC. Additional costs are spent on the amount of charged power to satisfy the driver's anxiety. The time anxiety reward is calculated in the local environment, i.e.,

$$r_{i,t}^{\text{ta}} = \begin{cases} 0 & t_i^a \leq t < t_i^x, \\ -\max(\text{SoC}_{i,t}^x - \text{SoC}_{i,t}, 0)^2 & t_i^x \leq t < t_i^d. \end{cases} \quad (12)$$

When  $t_i^a \leq t < t_i^x$ , the driver has not started to feel anxious about the battery level. No additional power is charged, so the cost is zero. When  $t_i^x \leq t < t_i^d$ , the term  $-\max(\text{SoC}_{i,t}^x - \text{SoC}_{i,t}, 0)^2$  denotes a penalty to additional charged power in satisfaction of TA. The penalty term is made nonlinear by using the square function. An intuitive statement is that penalty increases at an accelerating rate when the difference between  $\text{SoC}_{i,t}^x$  and  $\text{SoC}_{i,t}$  increases.

The range anxiety reward is calculated locally as

$$r_{i,t}^{\text{ra}} = -\max(\text{SoC}_i^d - \text{SoC}_{i,t}, 0)^2, \quad t = t_i^d. \quad (13)$$

At departure time  $t_i^d$ , the term  $-\max(\text{SoC}_i^d - \text{SoC}_{i,t}, 0)^2$  is similary a penalty to RA. We combine time anxiety reward and range anxiety reward to obtain the anxiety reward:

$$r_{i,t}^a = \kappa_{\text{ta}} r_{i,t}^{\text{ta}} + \kappa_{\text{ra}} r_{i,t}^{\text{ra}}, \quad (14)$$

where  $\kappa_{ta}$  and  $\kappa_{ra}$  control sensitivity to different anxieties according to the driver's habit. The anxiety reward reflects the driver's satisfaction with the current SoC.

If EVs are the only flexible loads on RDN, we assume that all EVs indirectly interact with the global environment(RDN) through charging piles connected to aggregators. Thus the optimal power flow in (9a) can be calculated by the management system of RDN. Then, the power change at time  $t$  on the substation is calculated as

$$g_t(a_{1,t}, \dots, a_{N,t}) = f_{\text{obj},t}(P_{agg,t}) - f_{\text{obj},t}(0), \quad (15)$$

which we write as  $g_t$  in the following for notation simplification. Here,  $f_{\text{obj},t}$  in (9a) is a function of  $P_{agg,t}$  set, thus,  $g_t$  is actually a function of  $(a_{1,t}, \dots, a_{N,t})$  according to (7). RDN will broadcast the power change  $g_t$ . Each EV receives power change  $g_t$  from the global environment through communication. Then, the EV*i* can quantify charging/discharging power except itself on RDN, and we regard it as grid reward, i.e.,

$$r_{i,t}^g = \begin{cases} -\max(g_t - g_{i,t}, 0) & a_{i,t} > 0, \\ \min(g_t - g_{i,t}, 0) & a_{i,t} < 0, \\ 0 & a_{i,t} = 0, \end{cases} \quad (16)$$

with  $g_{i,t} = g_t(0, \dots, a_{i,t}, \dots, 0)$ . The grid reward reflects RDN load, i.e., the idle degree of RDN. Specifically, the  $g_{i,t}$  represents the impact of EV*i* charging power on RDN. The  $g_t - g_{i,t}$  is the difference between the charging/discharging trends of the RDN and the charging/discharging power of an individual EV. If  $g_t - g_{i,t} > 0$  when  $a_{i,t} > 0$ , some other EVs are charging besides EV*i*. EVs charging concentrated will burden RDN. So we penalize such action with a negative value. And if  $g_t - g_{i,t} < 0$  when  $a_{i,t} < 0$ , EVs discharging concentrated will also cause an increase in node voltage, thereby affecting the RDN operation. Therefore, the discharging behavior of EV*i* is discouraged in this case. When  $a_{i,t} = 0$ , the EV*i* neither charges nor discharges. Thus, the grid reward is set as 0. In our assumption, charging will impose a greater burden on RDN because EV's charging power is larger than its discharging power.

The sum reward for EV*i* at time  $t$  combines the power reward, anxiety reward, and grid reward:

$$r_{i,t}(s_{i,t}, a_{i,t}) = \lambda_p r_{i,t}^p + \lambda_a r_{i,t}^a + \lambda_g r_{i,t}^g, \quad (17)$$

where weight coefficients  $\lambda_p$ ,  $\lambda_a$  and  $\lambda_g$  depend on sensitivity to electricity price, driver's anxiety, and RDN load. The setting of these coefficients depends on the drivers' habits and the structure of RDN. All these coefficients are non-negative.

#### F. Problem Formulation

The distribution network multi-EV G2V/V2G system aims to coordinate the EVs charging/discharging on RDN subjected to the physical model and the operational constraints. A global policy is learned to map the states of each agent to probability action space. The ultimate optimization objective is to obtain a global charging control policy  $\pi$  to maximize the expected cumulative reward:

$$\max_{\pi} \left( \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{a_{i,t} \sim \pi_i, s_{i,t+1} \sim \mathbb{P}_i} \left[ \sum_{t=1}^T \gamma^t \cdot r_{i,t}(s_{i,t}, a_{i,t}) \right] \right),$$

where  $\gamma \in [0, 1]$  reflects importance of future reward and  $T$  is the horizon of EV charging control process. Note that the OPF-based RDN shields the detailed interaction between the charging/discharging power of each EV. As a result, each agent only requires local observation in our proposed environment, simplifying the problem.

### III. METHODOLOGY

In this section, we first summarize the Soft Actor-Critic (SAC), used to control the charging behavior of a single EV agent. We then introduce FedSAC, a novel approach that combines SAC with Federated Learning to overcome data isolation issues in a privacy-preserving manner.

#### A. Reinforcement Learning for Single EV Charging

Reinforcement Learning enables an agent to learn through interactions with the environment and make near-optimal sequential decisions. SAC is one of the off-policy Reinforcement Learning frameworks based on maximum entropy to address the poor stability of model-free DRL methods. SAC aims to learn a more diverse policy by balancing exploration and exploitation with entropy regularization, which prevents the agent from getting stuck in local optima. The experimental results demonstrate that SAC achieves state-of-the-art performance on various benchmark tasks. Therefore, we adopt it in this paper to make optimal decisions in EV charging control.

The objective of SAC for EV*i* is to determine its associated optimal local policy  $\pi_i$ , which maximizes cumulative reward and policy entropy. This can be achieved through the following equation:

$$\pi_i^* = \arg \max_{\pi} \mathbb{E}_{a_{i,t} \sim \pi(\cdot | s_{i,t})} \left[ \sum_{t=0}^T \gamma^t \left( r_t(s_{i,t}, a_{i,t}) + \alpha_i \mathcal{H}(\pi(\cdot | s_{i,t})) \right) \right], \quad (18)$$

where  $\mathcal{H}(\cdot)$  denotes the entropy regularization, and the  $\alpha_i$  is a temperature coefficient that controls the weight between cumulative reward and policy entropy.

The soft Q-function in SAC for EV*i* is defined as

$$Q(s_{i,t}, a_{i,t}) = r_t(s_{i,t}, a_{i,t}) + \gamma \mathbb{E}_{s_{i,t+1} \sim \mathbb{P}_i} [V(s_{i,t+1})],$$

where  $V$  is a soft state value function approximated by the Q-function as

$$V(s_{i,t}) = \mathbb{E}_{a_{i,t} \sim \pi_i} [Q(s_{i,t}, a_{i,t}) - \alpha_i \log(\pi(a_{i,t} | s_{i,t}))].$$

The critic is approximated by a Deep Neural Network (DNN) parametric over  $\theta_i$ , i.e.,  $Q_{\theta_i}(s_{i,t}, a_{i,t})$ . The soft Q-function is learned as a regression problem to minimize the soft Bellman residual  $J_Q(\theta_i) :=$

$$\mathbb{E}_{(s_{i,t}, a_{i,t}) \sim \mathcal{D}_i} \left[ \frac{1}{2} \left( Q_{\theta_i}(s_{i,t}, a_{i,t}) - Q_{\hat{\theta}_i}(s_{i,t}, a_{i,t}) \right)^2 \right], \quad (19)$$

where

$$Q_{\hat{\theta}_i}(s_{i,t}, a_{i,t}) = r_t(s_{i,t}, a_{i,t}) + \gamma \cdot \mathbb{E}_{s_{i,t+1} \sim p} V_{\hat{\theta}_i}(s_{i,t+1})$$

sampling mini-batches from a replay buffer  $\mathcal{D}_i$ . The  $V_{\hat{\theta}_i}$  is the estimated soft state value defined by a target DNN parametric over  $\hat{\theta}_i$ . The parameter  $\hat{\theta}_i$  can be updated by moving average method  $\hat{\theta}_i = \tau\theta_i + (1 - \tau)\hat{\theta}_i$ .

The policy function is also approximated by a DNN parametric over  $\phi_i$ . The policy network is updated by minimizing Kullback-Leibler (KL) divergence  $J_\pi(\phi_i) :=$

$$\mathbb{E}_{\substack{s_{i,t} \sim \mathcal{D}_i \\ a_{i,t} \sim \pi_{\phi_i}}} [\alpha_i \log \pi_{\phi_i}(a_{i,t} | s_{i,t}) - Q_{\theta_i}(s_{i,t}, a_{i,t})]. \quad (20)$$

Because the charging rate of EV is continuous in the charging control process, we set the policy  $\pi_{\phi_i}$  as a Gaussian distribution, i.e.,

$$\pi_{\phi_i}(a_{i,t} | s_{i,t}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a_{i,t} - \mu)^2}{2\sigma^2}\right), \quad (21)$$

where  $\mu$  and  $\sigma$  are outputs of the policy network, which are the mean and standard deviation of  $s_{i,t}$ , respectively. As  $\mu$  and  $\sigma$  are the function of  $s_{i,t}$ , we can rewrite them as  $\mu = \mu_{\phi_i}(s_{i,t})$  and  $\sigma = \sigma_{\phi_i}(s_{i,t})$ . The policy network will output  $\mu$  and  $\sigma$  given the current state  $s_{i,t}$ . Then we randomly sample the current action  $a_{i,t}$  from (21).

To deploy backpropagation in the policy network, we introduce a reparameterization trick to rewrite the policy in (21) as  $f_{\phi_i}(\epsilon; s_{i,t}) = \mu_{\phi_i}(s_{i,t}) + \epsilon\sigma_{\phi_i}(s_{i,t})$  with  $\epsilon \sim \mathcal{N}(0, 1)$ . Specifically, the  $\epsilon$  is sampled from the standard normal distribution and the  $a_{i,t} = f_{\phi_i}(\epsilon; s_{i,t})$  is then generated deterministically. Thus the gradient of (20) is approximated by

$$\begin{aligned} & \nabla_{\phi_i} \log \pi_{\phi_i}(a_{i,t} | s_{i,t}) + \left( \nabla_{a_{i,t}} \log \pi_{\phi_i}(a_{i,t} | s_{i,t}) \right. \\ & \quad \left. - \nabla_{a_{i,t}} Q_{\theta_i^k}(s_{i,t}, a_{i,t}) \right) \nabla_{\phi_i} f_{\phi_i}(\epsilon; s_{i,t}). \end{aligned} \quad (22)$$

The choice of temperature coefficient affects the effectiveness of the policy network. Thus we use an adjustment strategy to update  $\alpha_i$  automatically by minimizing

$$J(\alpha_i) = \mathbb{E}_{\substack{s_{i,t} \sim \mathcal{D}_i \\ a_{i,t} \sim \pi_i}} [-\alpha_i \log \pi_{\phi_i}(a_{i,t} | s_{i,t}) - \alpha_i \mathcal{H}], \quad (23)$$

where  $\mathcal{H}$  is the desired minimum expected target entropy.

In practice, two critic networks  $Q_{\theta_i^k}$  and two target networks  $Q_{\hat{\theta}_i^k}$  where  $k \in \{1, 2\}$  are utilized to diminish the overestimation of Q-value. So the equation in (19) and (20) can be rewritten as

$$J_Q(\theta_i^k) = \mathbb{E}_{(s_{i,t}, a_{i,t}) \sim \mathcal{D}_i} \left[ \frac{1}{2} \left( Q_{\theta_i^k} - \min_k Q_{\hat{\theta}_i^k} \right)^2 \right], \quad (24)$$

$$J_\pi(\phi_i) = \mathbb{E}_{\substack{s_{i,t} \sim \mathcal{D}_i \\ a_{i,t} \sim \pi_{\phi_i}}} \left[ \alpha_i \log \pi_{\phi_i} - \min_k Q_{\theta_i^k} \right]. \quad (25)$$

Algorithm 1 presents charging control of  $EV_i$  based on SAC. Firstly,  $EV_i$  clears its experience replay buffer  $\mathcal{D}_i$  and initializes network parameters. Afterward, the  $EV_i$  trains until a stable charging strategy is obtained. During the environment step, the EV interacts with local and global environments to collect experience. In the global environment, the RDN calculates the power change  $g_t$  at time  $t$  through solving OPF in (9). It sends  $g_t$  and real-time electricity price  $\xi_t$  to all subordinate nodes at fixed intervals. The EV receives  $g_t$ , and  $\xi_t$  from RDN and then executes grid reward  $r_{i,t}^g$  through (16). The

$EV_i$  subsequently calculates the power reward  $r_{i,t}^p$  and anxiety reward  $r_{i,t}^a$  through interaction with the local environment. The sum reward  $r_t(s_{i,t}, a_{i,t})$  is executed, and the obtained state is stored into  $\mathcal{D}_i$  for the local training. In the gradient step, the  $EV_i$  locally updates the parameters, i.e.,  $(\theta_i^k, \phi_i, \alpha_i, \hat{\theta}_i^k)$ , by randomly sampling experience from  $\mathcal{D}_i$ . Here, the  $\delta_Q$ ,  $\delta\pi$  and  $\delta\alpha$  are the learning rates. Finally, the EV generates its charging control strategy.

---

**Algorithm 1** Soft Actor-Critic for Charging Control of Single EV

---

```

1: Initialize the replay buffer, actor network with  $\phi_i$  and
   (target) critic network with  $(\hat{\theta}_i^k)$   $\theta_i^k$ 
2: for each training episode do
3:   for each environment step do
4:     Local Environment:
5:       Sample action  $a_{i,t}$  from  $\pi_i$  given current state  $s_{i,t}$ 
6:       Execute  $a_{i,t}$  and obtain next state  $s_{i,t+1}$ 
7:     Global Environment:
8:       Solve OPF in (9) and obtain  $g_t$  through (15)
9:       Broadcast the result  $g_t$  on a communication link
10:      Broadcast the real-time electricity price  $\xi_t$  on a
        communication link
11:      Local Environment:
12:        Receive  $g_t$  and execute  $r_{i,t}^g$  through (16)
13:        Calculate the power reward  $r_{i,t}^p$  and anxiety reward
         $r_{i,t}^a$  through (11) and (13)
14:        Obtain  $r_t(s_{i,t}, a_{i,t})$  through (17)
15:        Store the experience  $(s_{i,t}, a_{i,t}, r_t, s_{i,t+1})$  into  $\mathcal{D}_i$ 
16:    end for
17:    for each gradient step do
18:      Update critic network parameters  $\theta_i^1$  and  $\theta_i^2$ :
19:       $\theta_i^k \leftarrow \theta_i^k + \delta_Q \nabla_{\theta_i^k} J_Q(\theta_i^k)$ ,  $k \in \{1, 2\}$ 
20:      Update actor network parameter  $\phi_i$ :
21:       $\phi_i \leftarrow \phi_i + \delta_\pi \nabla_{\phi_i} J_\pi(\phi_i)$ 
22:      Update temperature parameter  $\alpha_i$ :
23:       $\alpha_i \leftarrow \alpha_i + \delta_\alpha \nabla J(\alpha_i)$ 
24:      Update target critic network with moving average
        method:
25:       $\hat{\theta}_i^k = \tau\theta_i^k + (1 - \tau)\hat{\theta}_i^k$ ,  $k \in \{1, 2\}$ 
26:    end for
27:  end for

```

---

### B. FedSAC Framework for Multi-EV Charging

The SAC mentioned above is shared among all agents, thus providing high autonomy for each EV's charging control. However, the limited diversity of charging experiences within individual EVs can weaken the generalization performance of the agent. Moreover, due to the privacy-sensitive nature of charging/discharging behavior, sharing data among different EVs may pose a risk. To overcome these limitations, FL is proposed to train a highly generalized agent while preserving privacy [53].

Federated learning is a distributed machine learning framework for exchanging intermediate parameters free of privacy disclosure during training. It consists of two participants: various data-owned clients and the central server. The clients

update the parameters of the local model based on decentralized data and upload updates to a central server. The central server aggregates the parameters from the clients to update the global model. The server then broadcasts the updated global model to all clients and updates their parameters. This process is repeated multiple times until the model converges.

The algorithm procedure of FedSAC is shown in Algorithm 2. At the beginning of the training process, the server builds a global SAC model with parameters  $\theta_G^k$  and  $\phi_G$ . The  $N$  clients train their own SAC models on local devices in parallel by following Algorithm 1 from lines 4 to 29. The training on each client will iterate for  $N_e$  episodes. The server receives the updated parameters from the clients, and aggregates them through  $\theta_G^k = \frac{1}{N} \sum_{i=1}^N \theta_i^k$ ,  $k \in \{1, 2\}$ , and  $\phi_G = \frac{1}{N} \sum_{i=1}^N \phi_i$ . The new  $\theta_G^k$  and  $\phi_G$  replace the original parameters. The updated parameters are then broadcasted to the clients. The procedure of aggregation and broadcast repeats enough epochs until the global model converges.

**Algorithm 2** Federated Soft Actor-Critic for Multi-EVs Charging Control

```

1: Initialize the global actor network with  $\phi_G$  and critic
   network with  $\theta_G^k$ 
2: for global epoch do
3:   for client  $i$  from 1 to  $N$  in parallel do
4:     Update  $\theta_i^k$ ,  $\phi_i$  and  $\alpha_i$  locally: execute Algorithm 1
        from lines 4 to 29 with  $N_e$  looping episodes
5:     Upload  $\theta_i^k$  and  $\phi_i$  to the server
6:   end for
7:   Receive the parameters from each client
8:   Aggregate parameters of model through  $\theta_G^k = \frac{1}{N} \sum_{i=1}^N \theta_i^k$ ,  $k \in \{1, 2\}$  and  $\phi_G = \frac{1}{N} \sum_{i=1}^N \phi_i$ 
9:   for client  $i$  from 1 to  $N$  in parallel do
10:    Broadcast  $\theta_G^k$  and  $\phi_G$  to clients:  $\theta_i^k \leftarrow \theta_G^k$ ,  $\phi_i \leftarrow \phi_G$ 
11:   end for
12: end for
Output:  $\phi_G$ 
```

Fig. 2 visualizes the overall framework of FedSAC applied in EV charging control process. The EV agent is a client responsible for local model training and communication with the server and the RDN. Local models are stored on individual EVs considering mobility and privacy. When an EV is plugged into the V2G charging pile, the embedded computer will take charge of the charging process. The onboard embedded computer generates a charging/discharging rate for the local EV for the next period at a specific time. The RDN manages its power flow by solving the OPF problem and executing  $g_t$  at fixed intervals. Then, it broadcasts  $g_t$  and real-time electricity price  $\xi_t$  to all EVs (i.e., charging stations). The charging stations specify their unique price strategies according to  $\xi_t$ . Each EV calculates the individual rewards based on the global information from the RDN. After collecting enough rewards and states, such information will be used to update the parameters of SAC. The process is repeated enough times to generate a personalized charging policy for each EV. The EVs upload the updated parameters, and the server aggregates these parameters to generate a global model. The EVs download the

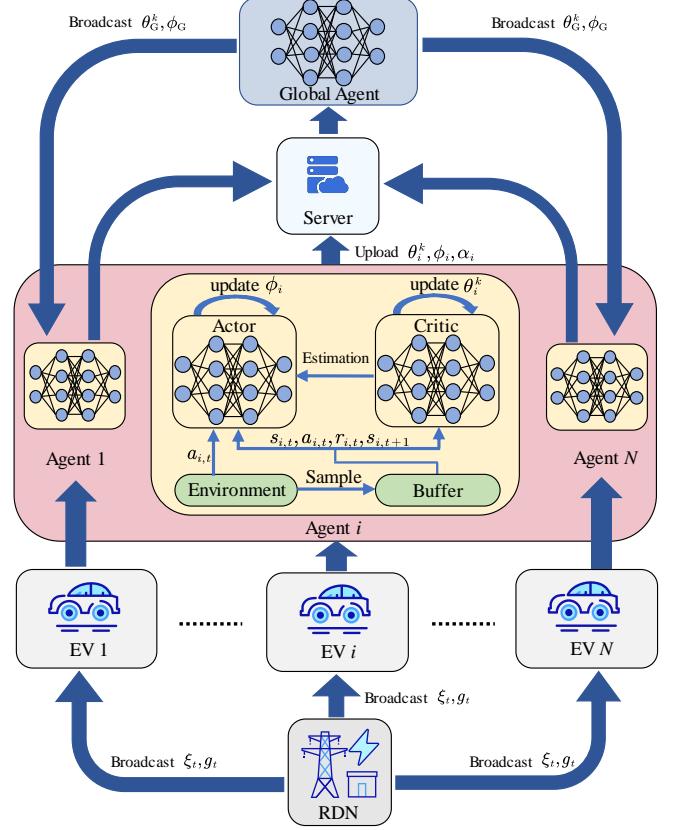


Fig. 2. The overall framework of FedSAC algorithm

global model to update the personalized charging policy. Training episodes repeat until the global model converges. Then a real-time charging control policy considering different driver behaviors is well-trained. As a standard federated learning framework, the FedSAC can train a highly generalized agent in a privacy-preserving manner.

#### IV. EXPERIMENT

In this section, we present several experiments to verify the performance of the proposed FedSAC applied to EV charging control. First, in Section IV-A, we describe the training settings and simulation setup. Then, in Section IV-B, we present the training performance of the proposed algorithm. In Section IV-C and IV-D, we conduct two simulations to verify the performance of charging control and charging power on RDN. Finally, in Section IV-E, we compare the performance of the proposed algorithm with different algorithms.

##### A. Experiment Setting

To demonstrate the effectiveness and performance of the proposed algorithm, we conducted our experiments on a platform with an Intel Xeon Silver 4310 CPU @ 2.10GHz × 12 and 1 NVIDIA RTX4090 GPU. All algorithms are implemented by Python 3.8.15 with Pytorch 1.7.1. The hourly electricity price of weekdays between July.1, 2021, and June.30, 2022, from ISO New England [54] is used. We divide the extracted dataset into two parts. The

first part includes the hourly electricity price of 100 weekdays used for simulation, and the second part is used to train the algorithm. The past 48 hours' electricity prices are input into the network to learn the trend of electricity prices.

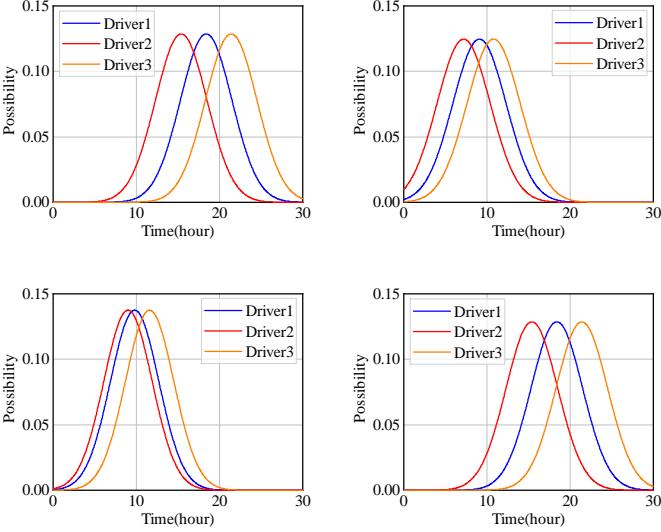


Fig. 3. Statistical distribution of drivers' travel habits: Home Arrival(upper left); Home Departure(upper right); Office Arrival(lower left); (d) Office Departure(lower right).

We study a thirty-car system with three kinds of travel and charging/discharging habits, as illustrated in Fig. 3(a)-(d). The curves in the figure indicate the probability of EV behavior. On weekdays, the kind-1 driver leaves home in the morning and arrives at the office after driving for a while. The driver finishes one day's work in the afternoon and departs the office. After driving for a period, the driver arrives home. We assume the EV is connected to the charging pile while not driving. The travel habit of kind-1 driver refers to the statistical analysis of the 2017 National Household Travel Survey (NHTS2017) [55]. Based on it, we assume the kind-2 driver prefers to return to the office early, and the kind-3 driver prefers to work overtime and return late. We sample the expected SoC  $\beta_i^1$  at departure time  $t_i^d$  for each EV driver from the ranges [0.85, 0.95], [0.85, 0.9] and [0.9, 0.95] respectively. The duration of three EVs is randomly distributed in [1, 4], [1, 2], and [2, 4]. The ratio of three different kinds of EV drivers is 3:1:1. Each charging station sets a different price strategy to attract drivers, which means that each EV faces different electricity prices simultaneously. Despite the unit cost of electricity potentially varying among charging stations, they must still purchase electricity from the RDN. Therefore, the prices conform to the general trend of electricity prices within the grid. To simulate this situation, we randomly sample from [-10, 10] and add it to the electricity price for every charging station at each timeslot. Additionally, we add a time offset of [-4, 4] to simulate any advance or delay in the electricity price trend. The EVs share some common parameter settings as well, including the initial SoC that is uniformly sampled from [0, 0.95], the shape parameter  $\beta_i^2$  that is sampled from  $\mathcal{N}(9, 1^2)$  with bounds [6, 12], and the charging/discharging

efficiency  $\eta_c, \eta_d$  that is set to 0.98. Each EV battery's energy storage capacity  $C_i$  is set to 0.03, meaning it can store 30 kW·h of energy. We restrict the charging rate to  $[-0.2, 1.0]$  to simulate the imbalance between the charging and discharging rates. The weight coefficients in (17) is set to  $\lambda_p = 9, \lambda_a = 1, \lambda_g = 100$  and  $\kappa_{ta} = 36, \kappa_{ra} = 16$ .

Our research uses a 74-bus distribution network based on the IEEE 74-bus [56] to construct the global environment. We assume the presence of G2V/V2G DC charging stations on odd buses. Thirty EVs are connected to them separately when not driving. We assume that the charging piles can provide the maximum charging/discharging rate required by the EVs. In our research, the optimal power flow (OPF) problem is solved through the primal-dual interior point method [57].

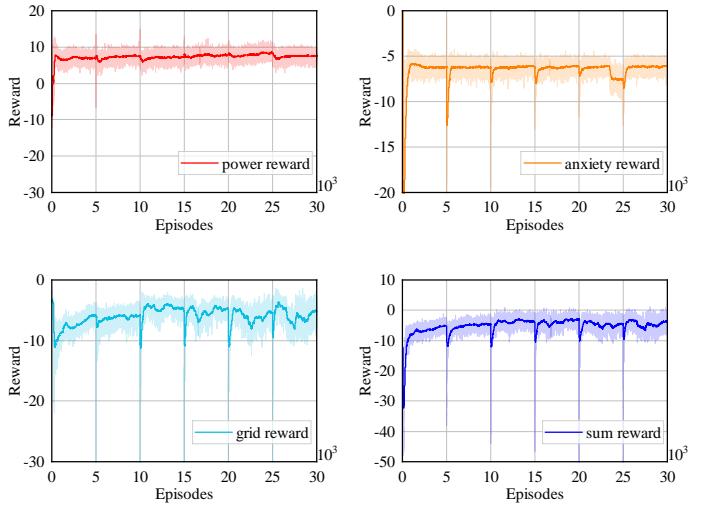


Fig. 4. The aggregation reward curve during the training of FedSAC: power reward (upper left); anxiety reward (upper right); grid reward (lower left); sum reward (lower right).

In the SAC framework, the critic network consists of three fully connected hidden layers with 128 neurons each and rectified linear unit (ReLU) activation functions are applied after each hidden layer to bound the output. The input to the critic network concatenates the state  $s_{i,t}$  and action  $a_{i,t}$ . A linear layer generates the final output Q-value. On the other hand, the actor network consists of four fully connected hidden layers with  $\{128, 128, 128, 128\}$  neurons. The input to the actor network is the current state  $s_{i,t}$ , and two parallel linear layers generate the output mean ( $\mu$ ) and standard deviation ( $\sigma$ ). The size of reply buffer  $\mathcal{D}_i$  is set to 10000, and the learning rate for the actor, critic, and temperature parameters are  $10^{-4}, 3 \times 10^{-4}$  and  $2 \times 10^{-4}$ , respectively. The batch size is 512, and the discount factor  $\gamma$  is 0.99. The training process is separated into six rounds, each conducting 5000 episodes. The FL mechanism is applied after every round.

## B. Training Performance

The reward curves of thirty agents during the training process are shown in Fig. 4. The figure includes four kinds of reward curves: power reward, anxiety reward, grid reward, and sum reward. Each of them is the average of the reward

curves of thirty EVs. The sum reward curve is the sum of the other three parts, which refer to (17). The shadowed regions represent the real training reward per episode, and the deep-colored curves illustrate smoothed episode reward curves. Although the natural average reward curves of the thirty agents exhibit significant fluctuations, the smoothed sum reward curves exhibit gradual growth and tend to stabilize eventually, indicating that our algorithm has learned a stable global policy. Since the FL mechanism averages the parameters of agents, the training reward would dramatically decrease during adjacent rounds, i.e., the reward curve drops at the 5000th, 10000th, etc, episode.

Specifically, the algorithm converges rapidly in the first round. After applying the FL mechanism, the reward curve drops briefly at the beginning of the second round. The power and anxiety rewards tend to be even with those of the first round, and the trend repeats throughout the training process. Overall, the power reward and anxiety reward curves are relatively stable during federated learning. Besides, the grid reward converges quickly in the first two rounds and falls into local optimal convergence at the end of 2nd round. However, with the help of the FL mechanism, the grid reward overcomes this and rises to a higher level in the subsequent rounds. The grid reward tends to be stable after about the 20000th episode. This phenomenon illustrates that the FL mechanism helps break through the limitations of a single EV dataset, thus improving the generalization of the algorithm. The grid reward is related to the other EVs in the RDN, in other words, globally relevant, while power reward and anxiety reward reflect the personalized requirement of each EV. Therefore, the federal mechanism has a more significant effect on improving the grid reward. During the training process, each reward curve shows an upward trend, indicating that our algorithm considers every optimization objective, and the weight coefficients set achieves a balance between several objectives. The sum reward situation is similar to that of grid rewards, and the final stable trend indicates that our proposed algorithm has learned a generalized global policy. In summary, the FL mechanism would not harm the convergence process and could help isolated agents escape from the local optimum caused by limited charging experience.

### C. Simulation of EV Charging Control

We design a five-weekday uninterrupted trip to verify the performance of the charging control policy. In our simulation, EV drivers still follow the travel habits in Fig. 3. We selected one of the three types of EV to observe the charging control strategy of each type of EV. When EV is in driving mode, the consumed energy is related to driving habits, driving distance, and environmental uncertainties (i.e., road conditions and weather). To simplify, we assume that the power consumption rate of each EV is 5% per hour. The simulation results are shown in Fig. 5-7. The colored columns of different lengths in Fig. 5-7(a) represent charging/discharging rates in one hour at different locations. The positive rates indicate the EV is charging currently, while the negative one means the EV is discharging. The line reflects the electricity price trend, normalized and consistent throughout the RDN. The line in

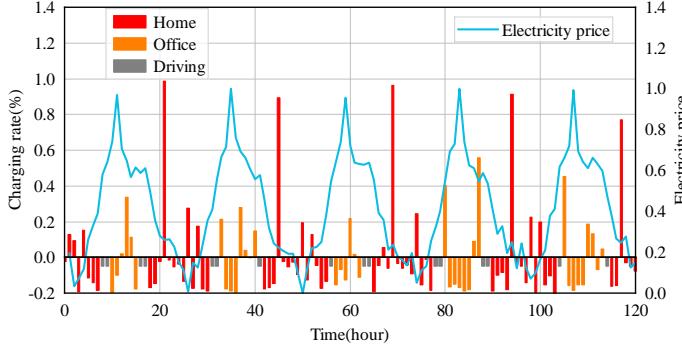
Fig. 5-7(b) reflects the SoC change during the charging control process.

As demonstrated in Fig. 5-7(a), three EVs commonly prefer charging at low electricity prices and discharging at high prices, regardless of their location, as a means to minimize costs. The proposed method can reduce costs by employing dynamic electricity prices within the charging control process. In other words, the EVs can effectively regulate the sequential load distribution of RDN and further stabilize grid load under the charging decisions. Besides, when other EVs work in charging mode on RDN in a timeslot, the current EV would avoid charging along with them to prevent exorbitant instantaneous power. For instance, since electricity prices reach the first valley at approximately the 21st hour, three EVs opt to charge and reduce costs at this time. Considering the load pressure on the relevant distribution network, the EVs are charged in the order of EV1, EV2, and EV3. This result confirms that the proposed approach recommends dispersed charging in time from a global perspective. Likewise, node injection can potentially elevate the voltage at grid nodes. To counteract this, our algorithm regulates the decentralized discharging of EVs. The above phenomenon occurs because the grid reward effectively considers the current grid load. Fig. 5-7(b) shows that the state of charge (SoC) remains no less than 0 throughout the simulation, indicating that the EV battery has ample power whenever needed for travel, thereby satisfying driver anxiety. The simulations confirm that the suggested algorithm enhances charging control policies, leading to lower charging expenses, reduced driver anxiety, and reduced RDN load. Moreover, this algorithm is suitable for all EVs.

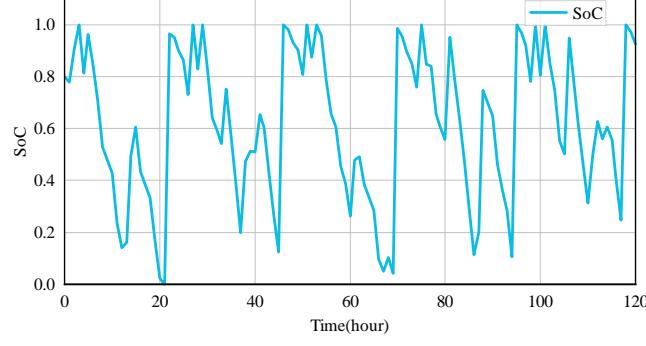
### D. Power Load Performance under the Grid Reward

The fluctuation in electricity prices exhibits a positive relationship with electricity demand. Under our policy, EVs prefer charging at times of low demand and discharging when demand is high, enabling them to play a novel role in electricity transactions. Under our assumption, the charging power of EVs is typically high, and simultaneous charging of all EVs could significantly augment the burden on the RDN, resulting in a rise in the load curve at the original valley and creating uncertainty in the power grid. Similarly, discharging also has the potential to introduce uncertainty, although the discharging power is typically lower than that of charging. Concentrated discharging can lead to an increase in node voltage, which further contributes to uncertainty. Such charging decisions would contravene the original purpose of V2G technologies. The simulation results in IV-C have proved that the three EVs would charge dispersively at the electricity peak. In this section, we design an ablation experiment to further assess the performance of the proposed approach from the perspective of the RDN.

We design a SAC-based algorithm (refer to [33]) without the grid reward from the global environment and a multi-agent SAC (SAC-MARL) algorithm with grid reward. The other settings keep consistent with the proposed approach. We conduct a 100-weekday charging control simulation for both



(a) Charging decision with different electricity prices

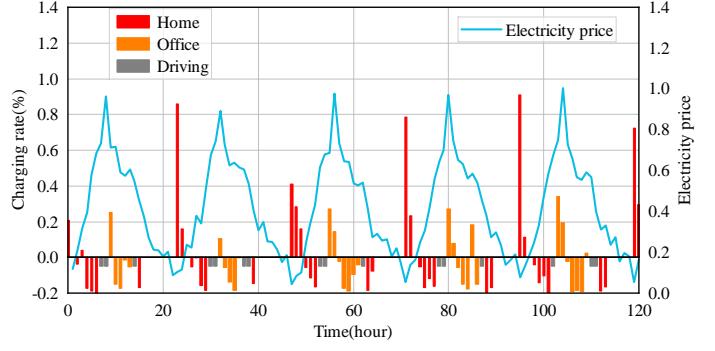


(b) EV SoC at a different time

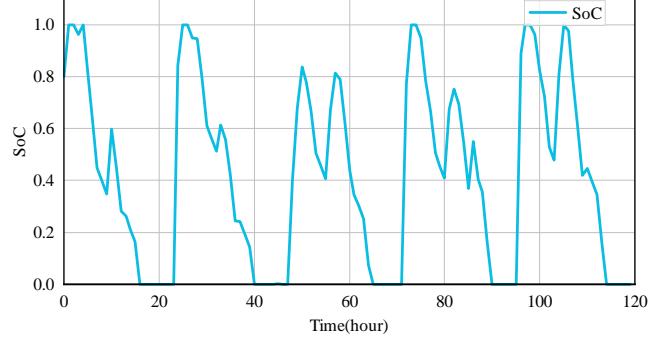
Fig. 5. Simulation of EV charging control for Kind-1 Drivers in five weekdays.

algorithms and get hourly charging/discharging action  $a_{i,t}$  for all EVs. To better demonstrate the charging load on the RDN, we calculate the power change  $g_t$  according to (15). The hourly power changes of the two algorithms are displayed in Fig. 8(a)-(b), with the colored bars representing the overall power the thirty-EV system is charging/discharging. Specifically, the charging/discharging power on the RDN in Fig. 8(a) fluctuates considerably with time under the control of SAC without grid reward. Conversely, the charging/discharging power on the RDN in Fig. 8(b) exhibits a comparatively smoother trend under the control of SAC-MARL, with a significant reduction in extreme power levels.

To quantify the power change resulting from charging/discharging on RDN during the simulation process, the standard deviation of power change ( $\sigma_g$ ) is compared. The variance of power changes in the distribution network quantifies their degree of dispersion, where smaller power changes result in reduced uncertainty factors in the power grid. Specifically, the standard deviations of the SAC-MARL and SAC are 0.0919 and 0.5163, respectively. The SAC-MARL has reduced the standard deviation by 82.2%, compared to the SAC algorithm without the grid reward. These results indicate that the algorithm with our proposed grid reward provides a more stable power curve. Hence, the algorithm with our proposed grid reward can regulate the charging behavior of each EV to apply a stabler charging load on the power grid.



(a) Charging decision with different electricity prices



(b) EV SoC at a different time

Fig. 6. Simulation of EV charging control for Kind-2 Drivers in five weekdays.

### E. Comparative Evaluation

In this section, we evaluate the effectiveness of the FL mechanism by comparing it with several benchmarks as follows:

1) **SAC-MADRL**: Multi-agent SAC takes the RDN as the global environment and considers the grid reward. The parameters setting and model structure are identical to that of FedSAC. SAC-MARL reduces the overhead of aggregation and broadcasting compared to FedSAC.

2) **DDPG-MADRL**: Deep deterministic policy gradient (DDPG) is an off-policy algorithm for continuous action spaces, including two actor and two critic networks. The multi-agent version introduces the global environment and considers the grid reward. The parameters setting and model structure are the same as that of FedSAC.

3) **TD3-MADRL**: Twin delayed deep deterministic policy gradient (TD3) is an extension of the DDPG. TD3 employs the clipped double-Q learning to prevent overestimation and stabilize learning. The multi-agent version introduces the global environment and considers the grid reward. The parameters setting keeps consistent with FedSAC.

We compare the above algorithms from two dimensions: the convergence and generalization. The convergence of agents based on various algorithms is demonstrated with average sum reward curves during training. The smooth reward curves for the thirty agents are shown in Fig. 9. In addition, we choose testing rewards and standard deviation of power change during

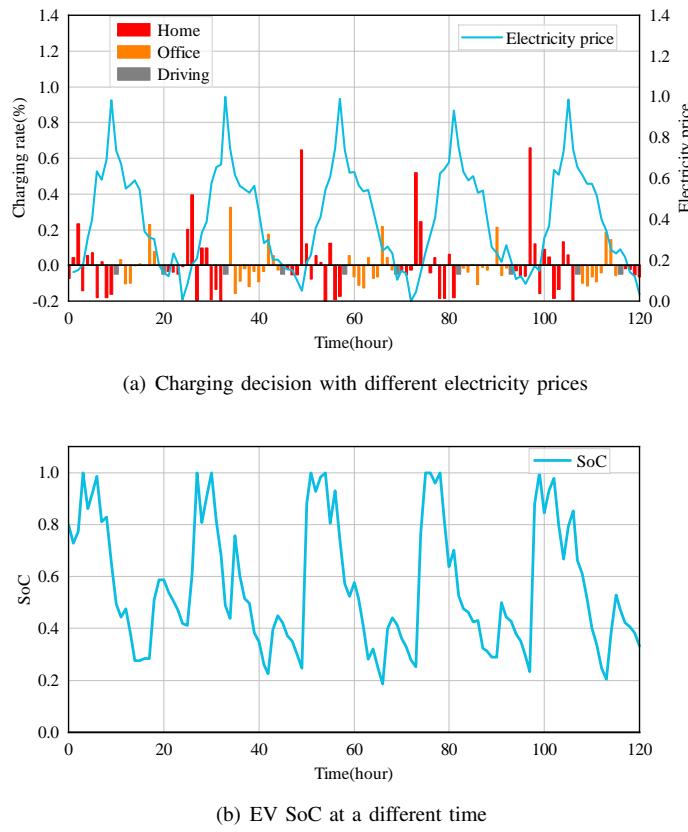


Fig. 7. Simulation of EV charging control for Kind-3 Drivers in five weekdays.

the simulation as metrics to compare the generalization of thirty EVs. The test reward evaluates the performance of the simulated EV charging control strategies, with a higher value indicating better performance. The meanings of the variance of power grid changes have been elaborated in Section IV-D. The simulation settings are consistent with those described in the preceding section. The testing rewards and average/standard deviation of power change are listed in Table I. The power changes of different algorithms on RDN are shown in Fig. 8.

Fig. 9 illustrates the convergence process of thirty EVs. All the algorithms achieve a well-converged state under the same number of training episodes. Our proposed FedSAC algorithm, which applies FL mechanisms, achieves the highest sum reward and is the best-performing algorithm. While the SAC-MARL algorithm also performs well, its limitation of isolated data makes its performance slightly inferior to that of FedSAC. However, the reward curve of FedSAC exhibits more volatility than SAC-MARL, demonstrating that FedSAC is slightly inferior in terms of stability. This is attributed to the non-uniformity and heterogeneity of data in distributed environments. TD3-MARL and DDPG-MADRL, which lack exploration of the environment, are far inferior to the SAC-based algorithms. Due to the optimization of the Q-function by the TD3 algorithm, its performance is still significantly better than that of DDPG-MARL.

Table I compares the generalization of different algorithms, and Fig. 8 compares the charging/discharging impact on RDN.

We compare the standard deviation of power change and their decline ratio compared to SAC defined in Section IV-E. The bold values in the table represent the best test results among all algorithms. As shown in Table I(a), the average test reward of thirty EV agents under FedSAC is -3720, which surpasses other comparative algorithms. Furthermore, the  $\sigma_g$  of thirty EV agents trained by FedSAC is 8.63 in Table. I(b), indicating the best performance among all algorithms. Therefore, the charging/discharging load on RDN (i.e., Fig. 8(c)) controlled by FedSAC is more stable than that of other algorithms. On the other hand, the DDPG-MADRL performs the worst on both average reward and standard deviation of power change, which is consistent with the training analysis. In addition, the average reward and  $\sigma_g$  of SAC-MADRL are -4412 and 9.19, respectively, which are the closest to the optimal values. This phenomenon clarifies that the FL mechanism can improve the generalization of SAC. Since the FedSAC outperforms comparative algorithms in test rewards and standard deviation, FedSAC has the best generalization effect.

The results above clarify that the introduction of the FL mechanism dramatically improves EV agents' convergence effect and generalization ability. In summary, the comparisons demonstrate that the proposed FedSAC outperforms existing algorithms in EV charging control.

TABLE I  
PERFORMANCE OF DIFFERENT ALGORITHMS DURING THE SIMULATION

(a) Cumulative average test rewards of different algorithms

	sum	reward	power	reward	anxiety	reward	grid	reward
DDPG-MARL	-5855		-211		-4136		-1508	
TD3-MARL	-4763		396		-3698		-1461	
SAC-MARL	-4412		519		-3790		-1141	
FedSAC	<b>-3720</b>		550		-3227		-1043	

(b) Standard deviation of power change and their decline ratio

	FedSAC	SAC-MADRL	TD3-MADRL	DDPG-MADRL	SAC
$\sigma_g(1e-2)$	<b>8.63</b>	9.19	12.29	12.41	51.63
decline	<b>83.28%</b>	82.2%	76.2%	75.96%	0%

## V. CONCLUSION

This paper proposes a federated deep reinforcement learning algorithm for the multi-EV charging control on an RDN. A smart OPF-based RDN is built first, which includes numerous isolated EVs connected to it, and an agent is used to manage EV charging control. A mathematical model is proposed to denote the RDN idle degree. The EV charging control problem is formulated as an MDP considering V2G profits, RDN idle degree, and driver's anxiety. Firstly, the individual optimal charging control strategy of each EV is learned by SAC. Then the FL mechanism is introduced to generate a global model by aggregating the local parameters from each agent. The driver's privacy is guaranteed during training. The case studies based on real-world data demonstrate that the proposed FedSAC can learn a stable cooperative charging control strategy and encourages decentralized EV charging, significantly reducing

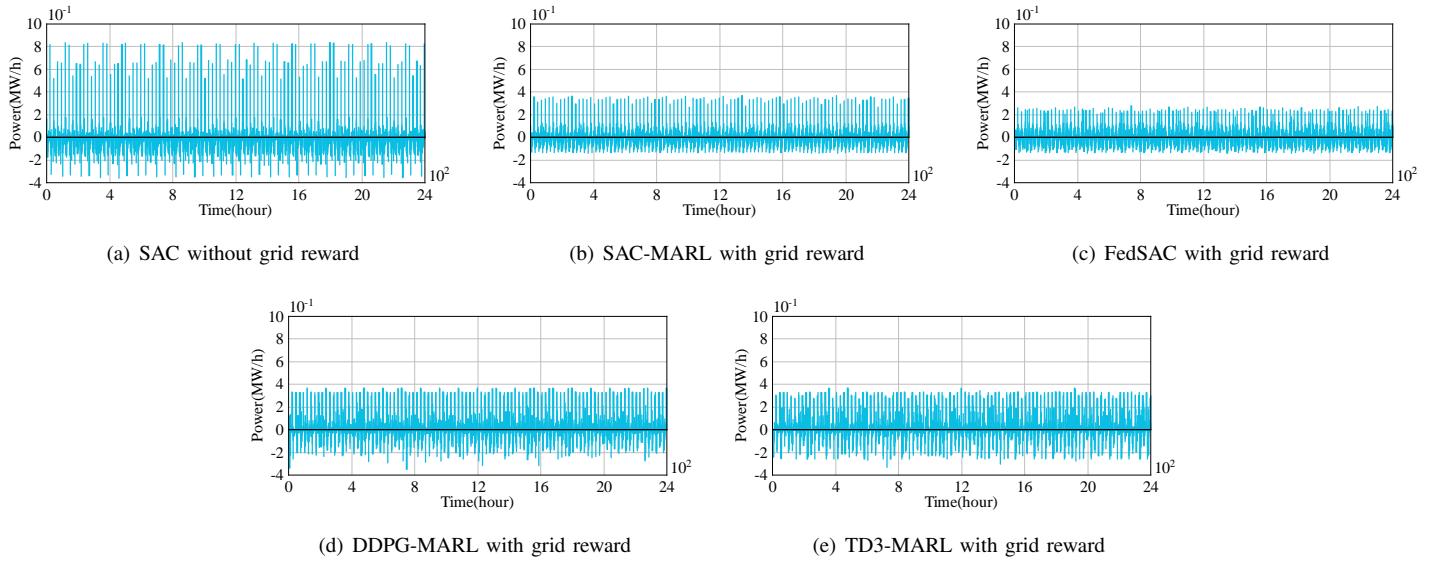


Fig. 8. Charging/Discharging power of thirty EVs on RDN.

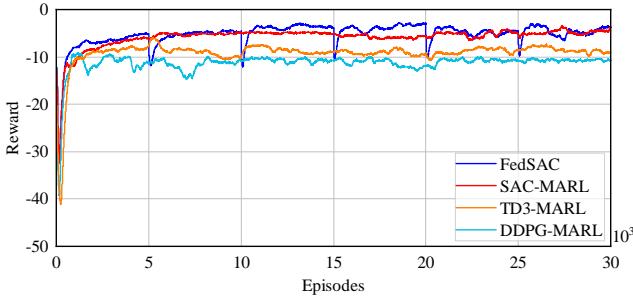


Fig. 9. Comparisons of training process.

the load fluctuations on RDN. Besides, the comparisons clarify that the proposed algorithm performs better on convergence effect and generalization ability.

In this paper, the proposed FedSAC has slight shortcomings in terms of stability. Some potential methods such as weighted FedAVG and local training optimization can be tried to address such issues. Besides, the mobility of EVs is not considered because their stochastic nature makes it challenging to model. The future work will investigate the modeling approach to EV mobility and consider how to introduce it into EV charging control.

## REFERENCES

- [1] X. Zhang, F. Gao, X. Gong, Z. Wang, and Y. Liu, "Comparison of climate change impact between power system of electric vehicles and internal combustion engine vehicles," in *Advances in Energy and Environmental Materials: Proceedings of Chinese Materials Conference 2017 18th*. Springer, 2018, pp. 739–747.
- [2] International Energy Agency (IEA), "Global ev outlook 2021: Accelerating the transition to electric vehicles," <https://www.iea.org/reports/global-ev-outlook-2021>, 2021, accessed: February 16, 2023.
- [3] M. R. Khalid, M. S. Alam, A. Sarwar, and M. J. Asghar, "A comprehensive review on electric vehicles charging infrastructures and their impacts on power-quality of the utility grid," *ETransportation*, vol. 1, p. 100006, 2019.
- [4] M. Nour, J. P. Chaves-Ávila, G. Magdy, and Á. Sánchez-Miralles, "Review of positive and negative impacts of electric vehicles charging on electric power systems," *Energies*, vol. 13, no. 18, p. 4675, 2020.
- [5] Y. Jiang, Q. Ye, B. Sun, Y. Wu, and D. H. Tsang, "Data-driven coordinated charging for electric vehicles with continuous charging rates: A deep policy gradient approach," *IEEE Internet of Things Journal*, vol. 9, no. 14, pp. 12 395–12 412, 2021.
- [6] J. A. P. Lopes, F. J. Soares, and P. M. R. Almeida, "Integration of electric vehicles in the electric power system," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 168–183, 2011.
- [7] W. Kempton and J. Tomic, "Vehicle-to-grid power implementation: From stabilizing the grid to supporting large-scale renewable energy," *Journal of power sources*, vol. 144, no. 1, pp. 280–294, 2005.
- [8] B. K. Sovacool, J. Kester, L. Noel, and G. Z. de Rubens, "Actors, business models, and innovation activity systems for vehicle-to-grid (v2g) technology: A comprehensive review," *Renewable and Sustainable Energy Reviews*, vol. 131, p. 109963, 2020.
- [9] J. Lin, K.-C. Leung, and V. O. Li, "Optimal scheduling with vehicle-to-grid regulation service," *IEEE Internet of Things Journal*, vol. 1, no. 6, pp. 556–569, 2014.
- [10] Q. Sun, H. Lv, S. Gao, K. Wei, and M. Mauersberger, "Optimized control of reversible vsc with stability mechanism study in smes based v2g system," *IEEE Transactions on Applied Superconductivity*, vol. 29, no. 2, pp. 1–6, 2019.
- [11] X. Chen, H. Wang, F. Wu, Y. Wu, M. C. González, and J. Zhang, "Multimicrogrid load balancing through ev charging networks," *IEEE Internet of Things Journal*, vol. 9, no. 7, pp. 5019–5026, 2021.
- [12] X. Tang, S. Bi, and Y.-J. A. Zhang, "Distributed routing and charging scheduling optimization for internet of electric vehicles," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 136–148, 2018.
- [13] B. Wang, Y. Wang, H. Nazaripouya, C. Qiu, C.-C. Chu, and R. Gadhi, "Predictive scheduling framework for electric vehicles with uncertainties of user behaviors," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 52–63, 2016.
- [14] M. Muratori, "Impact of uncoordinated plug-in electric vehicle charging on residential power demand," *Nature Energy*, vol. 3, no. 3, pp. 193–201, 2018.
- [15] J. Jin and Y. Xu, "Shortest-path-based deep reinforcement learning for ev charging routing under stochastic traffic condition and electricity prices," *IEEE Internet of Things Journal*, vol. 9, no. 22, pp. 22 571–22 581, 2022.
- [16] M. Moschella, P. Ferraro, E. Crisostomi, and R. Shorten, "Decentralized assignment of electric vehicles at charging stations based on personalized cost functions and distributed ledger technologies," *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 11 112–11 122, 2021.
- [17] Y. Cao, H. Wang, D. Li, and G. Zhang, "Smart online charging algorithm for electric vehicles via customized actor-critic learning," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 684–694, 2021.

- [18] D. Wu, H. Zeng, C. Lu, and B. Boulet, "Two-stage energy management for office buildings with workplace ev charging and renewable energy," *IEEE Transactions on Transportation Electrification*, vol. 3, no. 1, pp. 225–237, 2017.
- [19] E. Sortomme and M. A. El-Sharkawi, "Optimal combined bidding of vehicle-to-grid ancillary services," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 70–79, 2011.
- [20] A. T. Al-Awami and E. Sortomme, "Coordinating vehicle-to-grid services with energy trading," *IEEE Transactions on smart grid*, vol. 3, no. 1, pp. 453–462, 2011.
- [21] S. Han, S. Han, and K. Sezaki, "Development of an optimal vehicle-to-grid aggregator for frequency regulation," *IEEE Transactions on smart grid*, vol. 1, no. 1, pp. 65–72, 2010.
- [22] Q. Huang, Q.-S. Jia, Z. Qiu, X. Guan, and G. Deconinck, "Matching ev charging load with uncertain wind: A simulation-based policy improvement approach," *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1425–1433, 2015.
- [23] H.-M. Chung, S. Maharjan, Y. Zhang, and F. Eliassen, "Intelligent charging management of electric vehicles considering dynamic user behavior and renewable energy: A stochastic game approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 12, pp. 7760–7771, 2020.
- [24] A. Alsabagh, B. Wu, and C. Ma, "Distributed electric vehicles charging management considering time anxiety and customer behaviors," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2422–2431, 2020.
- [25] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.
- [26] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [27] Y. Li, "Deep reinforcement learning: Opportunities and challenges," *arXiv preprint arXiv:2202.11296*, 2022.
- [28] J. Wang, W. Xu, Y. Gu, W. Song, and T. C. Green, "Multi-agent reinforcement learning for active voltage control on power distribution networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3271–3284, 2021.
- [29] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time ev charging scheduling based on deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5246–5257, 2018.
- [30] F. Zhang, Q. Yang, and D. An, "Cddpg: A deep-reinforcement-learning-based approach for electric vehicle charging control," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3075–3087, 2020.
- [31] T. Qian, C. Shao, X. Wang, and M. Shahidehpour, "Deep reinforcement learning for ev charging navigation by coordinating smart grid and intelligent transportation system," *IEEE transactions on smart grid*, vol. 11, no. 2, pp. 1714–1723, 2019.
- [32] H. Li, Z. Wan, and H. He, "Constrained ev charging scheduling based on safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2427–2439, 2019.
- [33] L. Yan, X. Chen, J. Zhou, Y. Chen, and J. Wen, "Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5124–5134, 2021.
- [34] T. Mai, H. Yao, Z. Xiong, S. Guo, and D. T. Niyato, "Multi-agent actor-critic reinforcement learning based in-network load balance," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.
- [35] F. L. Da Silva, C. E. Nishida, D. M. Roijers, and A. H. R. Costa, "Coordination of electric vehicle charging through multiagent reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2347–2356, 2019.
- [36] J. Jin and Y. Xu, "Optimal policy characterization enhanced actor-critic approach for electric vehicle charging scheduling in a power distribution network," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1416–1428, 2020.
- [37] L. Yan, X. Chen, Y. Chen, and J. Wen, "A cooperative charging control strategy for electric vehicles based on multiagent deep reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 8765–8775, 2022.
- [38] Y. Shi, S. Xia, Y. Zhou, Y. Mao, C. Jiang, and M. Tao, "Vertical federated learning over cloud-ran: Convergence analysis and system optimization," *arXiv preprint arXiv:2305.06279*, 2023.
- [39] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6g: Vision, enabling technologies, and applications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 5–36, 2021.
- [40] H. Yang, K.-Y. Lam, L. Xiao, Z. Xiong, H. Hu, D. Niyato, and H. Vincent Poor, "Lead federated neuromorphic learning for wireless edge artificial intelligence," *Nature communications*, vol. 13, no. 1, p. 4269, 2022.
- [41] S. Frank, I. Steponavice, and S. Rebennack, "Optimal power flow: A bibliographic survey i: Formulations and deterministic methods," *Energy systems*, vol. 3, pp. 221–258, 2012.
- [42] M. Chakravorty and D. Das, "Voltage stability analysis of radial distribution networks," *International Journal of Electrical Power & Energy Systems*, vol. 23, no. 2, pp. 129–135, 2001.
- [43] Y. Zhan, P. Li, Z. Qu, D. Zeng, and S. Guo, "A learning-based incentive mechanism for federated learning," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6360–6368, 2020.
- [44] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
- [45] Z. Wang, J. Qiu, Y. Zhou, Y. Shi, L. Fu, W. Chen, and K. B. Letaief, "Federated learning via intelligent reflecting surface," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 808–822, 2021.
- [46] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.
- [47] J. Qi, Q. Zhou, L. Lei, and K. Zheng, "Federated reinforcement learning: Techniques, applications, and open challenges," *arXiv preprint arXiv:2108.11887*, 2021.
- [48] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge ai: Algorithms and systems," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2167–2191, 2020.
- [49] Y. Shi, Y. Zhou, D. Wen, Y. Wu, C. Jiang, and K. B. Letaief, "Task-oriented communications for 6g: Vision, principles, and technologies," *arXiv preprint arXiv:2303.10920*, 2023.
- [50] H. Yang, J. Zhao, Z. Xiong, K.-Y. Lam, S. Sun, and L. Xiao, "Privacy-preserving federated learning for uav-enabled networks: Learning-based joint scheduling and resource management," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 3144–3159, 2021.
- [51] H. Fan, C. Duan, C.-K. Zhang, L. Jiang, C. Mao, and D. Wang, "ADMM-based multiperiod optimal power flow considering plug-in electric vehicles charging," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 3886–3897, 2017.
- [52] M. Ehsani, Y. Gao, S. Longo, and K. M. Ebrahimi, *Modern electric, hybrid electric, and fuel cell vehicles*. CRC press, 2018.
- [53] M. Xu, J. Peng, B. Gupta, J. Kang, Z. Xiong, Z. Li, and A. A. Abd El-Latif, "Multiagent federated reinforcement learning for secure incentive mechanism in intelligent cyber-physical systems," *IEEE Internet of Things Journal*, vol. 9, no. 22, pp. 22095–22108, 2021.
- [54] I. N. England. Real-time maps and charts. Accessed: Feb. 19, 2023. [Online]. Available: <https://www.iso-ne.com/isoexpress/>
- [55] F. H. Administration. National household travel survey. Accessed: Feb. 19, 2023. [Online]. Available: <https://nhts.ornl.gov/>
- [56] S. M. Myint and S. W. Naing, "Network reconfiguration for loss reduction and voltage stability improvement of 74-bus radial distribution system using particle swarm optimization algorithm," *Int. J. Electr. Electron. Data Commun.*, vol. 3, no. 6, 2015.
- [57] R. A. Jabr, A. H. Coonick, and B. J. Cory, "A primal-dual interior point method for optimal power flow dispatching," *IEEE Transactions on Power Systems*, vol. 17, no. 3, pp. 654–662, 2002.