Coordinated Electric Vehicle Active and Reactive Power Control for Active Distribution Networks

Yi Wang, *Student Member, IEEE,* Dawei Qiu, *Member, IEEE,* Goran Strbac *Member, IEEE* and Zhiwei Gao, *Senior Member, IEEE,*

***Abstract*—The deployment of renewable energy in power systems may raise serious voltage instabilities. Electric vehicles (EVs), owing to their mobility and flexibility characteristics, can provide various ancillary services including active and reactive power. However, the distributed control of EVs under such scenarios is a complex decision-making problem with enormous dynamics and uncertainties. Most existing literature employs model-based approaches to formulate active and reactive power control problems, which require full models and are time-consuming. This paper proposes a multi-agent reinforcement learning algorithm featuring a deep deterministic policy gradient method and a parameter sharing framework to solve the EVs coordinated active and reactive power control problem towards both demand-side response and voltage regulations. The proposed algorithm can further enhance the learning stability and scalability with privacy perseverance via the location marginal prices. Simulation results based on a modified IEEE 15-bus network are developed to validate its effectiveness in providing system charging and voltage regulation services. The proposed LMP-PSDDPG algorithm is evaluated to achieve 38%, 16%, and 25% speedup, and 1.58, 0.69, and 0.27 times higher reward over the benchmarks DDPG, TD3 and LMP-DDPG, respectively.**

***Index Terms*—Electric vehicles, active distribution networks, active and reactive power control, location marginal prices, multi-agent reinforcement learning.**

I. Introduction

POWER systems are undergoing a significant transition from fossil fuel resources to the decarbonlization of *renewable energy resources* (RESs), promising to a low-carbon future [1]. However, besides the primary challenges posed by the intermittent nature of RESs, there are additional difficulties in voltage stability and reliability of power system operations due to the lack of voltage compensation devices [2]. To address above challenges, *electric vehicles* (EVs) have been deployed for various ancillary services including demand-side response and voltage regulations due to their significant advantages on mobility and flexibility [3]. As such, the research on the deployment of EVs towards both active and reactive power control for *active distribution networks* (ADNs) is a promising area. And, it is urgent to develop a smart and automatic control scheme for coordinated EVs active and reactive power control under such a highly complex environment with various uncertainties and dynamics.

Yi Wang, Dawei Qiu, and Goran Strbac are with the Department of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, U.K. (e-mail: yi.wang18@imperial.ac.uk, d.qiu15@imperial.ac.uk, g.strbac@imperial.ac.uk)

Zhiwei Gao is with the Faculty of Engineering and Environment, University of Northumbria, Newcastle upon Tyne, NE1 8ST, U.K. (e-mail:zhiwei.gao@northumbria.ac.uk).

A. Literature Review

The literature on EVs scheduling problems has been reviewed in [3], few of them however contribute to the coordination effect of EVs active and reactive power support. More specifically, the existing work focusing on model-based optimization approaches can be classified into two categories based on the deployed control mechanisms: centralized control [4]–[6] and decentralized control [7]–[10]. The centralized control methods are normally featured by relatively high solution qualities but long computing time and potential privacy issues, which have been applied to solve the coordinated active and reactive power control problems of EVs [4] as well as the voltage regulation and power loss reduction problems [5]. However, system uncertainties are not modeled in above two papers. To this end, a robust optimization method capturing uncertainties pertaining to system demand and EV patterns is proposed in [6] to control the active and reactive power of EVs towards voltage regulation. Compared to centralized control, decentralized control does not require central controllers, which can benefit privacy protection and ensure timely decision making [11]. More specifically, parallel consensus algorithm [7] and hierarchical coordination framework [8] are both used to solve the coordinated EVs active and reactive power control problems in a decentralized manner. *Vehicle-to-Grid* (V2G) reactive power dispatch is captured in [9] for the distributed EVs coordination. However, above papers [7]–[9] assume the fixed *location marginal prices* (LMPs), which cannot effectively capture the spatial benefits of EV fleets on voltage regulations. As a result, authors in [10] develop a decomposition method based on branch flow algorithm for self-dispatched EVs with the ADN towards optimal coordination adapting to spatiotemporally varying LMPs for real and reactive power, i.e., P-LMP and Q-LMP respectively.

Despite the aforementioned attempts [4]–[10] in solving the coordinated EV active and reactive power control problems, model-based optimization methods exhibit several limitations: 1) *distribution system operator* (DSO) is difficult to acquire explicitly the operation models and technical parameters of all self-controllable EVs in centralized control methods; 2) EVs are not willing to share local information with each other to reach a consensus in decentralized control methods; 3) solving a stochastic optimization problem is normally time consuming, moreover the uncertainties constructed from the pre-defined distributions may destroy their nature. In view of above drawbacks in model-based optimization approach, *reinforcement learning* (RL) [12] is a model-free approach that

studies the sequential and dynamic decision-making process of agents who can gradually learn the optimal control schemes by utilizing experiences acquired from its repeated interactions with the environment, without a *prior* knowledge. In addition, RL as an online learning method can make efficient use of increasing data, thereby capturing system uncertainties and adapting to various state conditions.

Despite the fact that RL lacks theoretical guarantees of convergence [12], previous works have successfully applied RL methods to various EV control problems [13], which can be classified into two categories based on the deployed EV numbers: single-agent RL (SARL) [14]–[23] and multi-agent RL (MARL) [24]–[29]. In SARL, deep Q-network (DQN) has been widely applied to solve various EV problems, e.g., real-time charging/discharging strategies [14], optimal EV charging navigation [15], and fast frequency regulation services [16]. However, DQN estimating the Q-values of finite actions is only possible for problems with simple discrete control actions. In order to handle the continuous action space, policy gradient theorem [12] is proposed to directly compute the values of control actions. Currently, deep deterministic policy gradient (DDPG) has been utilized to solve EV real-time voltage control problem [17], EV aggregator's bidding strategy problem [18], EV energy management strategy problem [19], EV optimal charging behaviors for satisfying the user's energy requirement [20], and pricing EV demand response with discrete charging rates [21]. Furthermore, in order to explore environment more widely and make training more stable, soft actor-critic (SAC) is also proposed to solve the EVs optimal charging control strategy [22], [23]. On the other hand, the research of MARL under EV concept is still limited. In [24], a multi-deep-Q-network is proposed to solve the optimal bidding strategy selection for EVs charging in an auction market. However, directly applying conventional DQN to each EV agent without extra information may suffer from instability issue, since all other agents' policies are implicitly formulated as part of the environment dynamics while their policies are continuously adjusted during the training process. To capture the other agents' control policy, a collective-policy model is proposed in [25] based on SAC that learns a fully decentralized policy for EVs charging strategy constrained by the transformer overload capacity. A hierarchical and hybrid MARL method based on proximal policy optimization (PPO) is proposed in [26] to optimize the multi-service provisions for EVs in a coupled power-transportation network. In order to handle the uncertainties of demand and price, a long-short term memory neural network is integrated with an actor-critic MARL method in [27] that learns the coordinated charging behaviors of EVs with privacy perseverance. Finally, a communication based MARL algorithm named CommNet is proposed in [28] to solve the energy management problem of distributed EV charging stations that allows information exchange with each other. However, this mechanism may destroy the privacy of local charging stations. As a result, a SAC-based MARL algorithm (MASAC) is developed in [29] to solve a strategic charging pricing scheme for EV operators. MASAC owing to its centralized training and decentralized execution framework [12] that takes all agents' information into the training process but requires only local information when executing actions, thereby protecting the private information in the execution process. Nevertheless, it is still problematic to apply MASAC to a large-scale system since the joint information increases proportionally with the agent size.

B. Contributions

To overcome these limitations, this paper proposes a novel MARL method to solve the coordination effect of EVs active and reactive power control problem in an ADN in responding to both P-LMPs and Q-LMPs, in which the specific contributions are outlined as below:

1) Formulating the active and reactive power control problem among multiple EVs in an ADN as a *Decentralized Partially Observable Markov Decision Process* (Dec-POMDP) [30]. In this case, EVs can maximize their utilization in distribution system operation to guarantee the active power supply and support the reactive power compensation.

2) Proposing a novel MARL method to efficiently solve the Dec-POMDP by constructing an actor-critic architecture-based DDPG method [31] to generate the continuous action spaces; abstracting a new Q-value function via P-LMPs and Q-LMPs to stabilize the training performance with enhanced scalability and protected privacy; and adopting a parameter sharing (PS) framework [32] to reach the distributed control with accelerated training speed.

3) Solving the MARL-based Dec-POMDP does not require the exact mathematical models and technical parameters of the distribution network system. Furthermore, the system uncertain and dynamic characteristics associated with the grid prices, demand, and PV generation can be learned during the training process without constructing any hypothetical probability distribution.

4) Training an on-line MARL-based control policy that can be directly deployed to the test dataset in milliseconds without any optimization. The trained policy is also generalized and automatic that can adapt to various uncertain conditions.

II. Problem Formulation

A. Problem Setting

The problem is focusing on the coordination of a group of EVs in an ADN, depicted in Fig. 1, which can be divided into two layers: communication layer and power layer. In detail, the communication layer is operated by an EV aggregator who can communicate with individual EVs and exchange the limited information to help EVs make more informed decisions. Inside the power layer, distributed energy resources (DERs) including electric demand (ED), diesel generators (DGs), as well as the RES-based solar photovoltaics (PVs) and wind turbines (WTs) are appropriately deployed. EVs can move between different buses for two trips (work from home in the morning and back from office at night) per day and exhibit both G2V and V2G flexibility through charging stations. DSO is responsible for managing the controllable DERs and performing OPF to ensure the stable and secure operation.

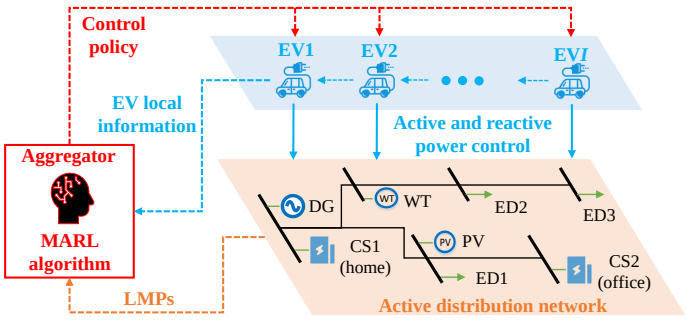

Fig. 1. Scheme of EV coordination in an ADN.

Unlike the centralized control that EVs are owned and managed by DSO, this paper assumes that all EVs are private-owned having their own traveling patterns and operation characteristics. Therefore, these EVs operate in a distributed manner without a central controller that can preserve their private information. However, we still introduce an aggregator who can provide proper information and incentives for EVs to reach a cooperative fashion. At each time step, after reading the local (active and reactive) power information of EDs and RESs as well as the battery information of energy content, each EV with a smart control algorithm can optimally manage its active and reactive power in the ADN. The objective of this problem is twofold: 1) EVs maximize the energy arbitrage by charging (discharging) active power at low (high) P-LMP periods; 2) EVs maximize the income by injecting or consuming reactive power at high Q-LMP periods.

B. Decentralized Partially Observable Markov Decision Process

We formulate the coordinated effect of EVs active and reactive power control problem as a finite Dec-POMDP with discrete time steps. The Dec-POMDP is defined by $\langle I, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$, which includes I agents (EVs), a set of global states $s \in \mathcal{S}$, a collection of local observations $\{o_i \in \mathcal{O}_{1:I}\}$, a collection of action sets $\{a_i \in \mathcal{A}_{1:I}\}$, a collection of reward functions $\{r_i \in \mathcal{R}_{1:I}\}$, and a state transition function $\mathcal{T}(s, o_{1:I}, a_{1:I}, \omega)$, where ω is the environment (ADN) stochasticity representing uncertain parameters. The time interval between two consecutive steps Δt = 1 hour. At time step t, each agent i chooses an action $a_{i,t}$ according to its control policy $\mu(o)$ based on its local observation $o_{i,t}$. The environment then moves into the next state s_{t+1} according to the state transition function $\mathcal{T}$. Each agent i obtains a reward $r_{i,t}$ and a new local observation $o_{i,t+1}$. Such process continues and then emits a trajectory of observations, actions, and rewards for each agent i: $\tau_i = o_{i,1}, a_{i,1}, r_{i,1}, o_{i,2}, ..., r_{i,T}$ over $\mathcal{O}_i \times \mathcal{A}_i \times \mathcal{O}_i \to \mathbb{R}$. Each agent i aims to maximize its cumulative discounted reward $R_i = \sum_{t=1}^{T} \gamma^t r_{i,t}$, where $\gamma \in [0, 1)$ is the discount factor and T = 24 hours is the daily horizon. The components of the Dec-POMDP are detailed as:

*1) **Observation**:* Each agent i at time step t has its local observation

$$o_{i,t} = [N^{ev}_{i,t}, P^{ed}_{i,t}, Q^{ed}_{i,t}, \tilde{P}^{res}_{i,t}, E^{ev}_{i,t}] \in \mathcal{O}_i, \tag{1}$$

comprising two parts: 1) the exogenous state unaffected by actions includes the bus location of charging station $N^{ev}_{i,t}$ where agent i is plugging in, the nodal active and reactive ED $P^{ed}_{i,t}, Q^{ed}_{i,t}$ as well as the nodal RES active power $\tilde{P}^{res}_{i,t}$; and 2) the endogenous state serving as the feedback signals of actions is modeled as the battery energy content $E^{ev}_{i,t}$.

*2) **Action**:* Each agent i at time step t controls its action

$$a_{i,t} = [a^p_{i,t}, a^q_{i,t}] \in \mathcal{A}_i, \tag{2}$$

comprising two parts: 1) the active power action $a^p_{i,t} \in [-1, 1]$ represents the magnitude of charging (positive) and discharging (negative) active power as a percentage of its active power capacity $P^{ev}_{i,t} \in [-\overline{P}^{ev}_i, \overline{P}^{ev}_i]$; and 2) the reactive power action $a^q_{i,t} \in [-1, 1]$ represents the magnitude of consuming (positive) and providing (negative) reactive power as a percentage of its reactive power limit $Q^{ev}_{i,t} \in [-\sqrt{(\overline{S}^{ev}_i)^2 - (P^{ev}_i)^2}), \sqrt{(\overline{S}^{ev}_i)^2 - (P^{ev}_i)^2})]$. The detailed EV charger model and its structure as well as the derivations of power limits are presented in Appendix section.

*3) **State Transition**:* The state transition is governed by $s_{t+1} = \mathcal{T}(s_t, o_{1:I,t}, a_{1:I,t}, \omega_t)$, influenced by the combination of environment current state s_t, all agents' local observations $o_{1:I,t}$ and actions $a_{1:I,t}$, as well as environment stochasticity $\omega_t = [\pi^P_t, \pi^Q_t, N^{ev}_{i,t}, P^{ed}_{d,t}, Q^{ed}_{d,t}, \tilde{P}^{res}_{g,t}]$. In this problem, ω_t is decoupled from the agents' actions and are characterized by inherent variability. For instance, the main grid active and reactive prices π^P_t, π^Q_t are determined by the market conditions, the parked location $N^{ev}_{i,t}$ is related to the EV users' residence, the active and reactive demand $P^{ed}_{d,t}, Q^{ed}_{d,t}$ are influenced by the energy usage behaviors, and the RES generation $\tilde{P}^{res}_{g,t}$ is affected by the solar radiation or wind speed. As a result, it is very difficult to explicitly model the distributions of these uncertain parameters. In machine learning area, RL remedies this problem in a data-driven approach that does not rely on accurate models of the underlying uncertainties but learns the dynamic characteristics directly from data sources [12].

By contrast, the state transition for endogenous state feature $E^{ev}_{i,t}$ can be directly determined by action $a^p_{i,t}$. More specifically, we reformulate the EV active power $P^{ev}_{i,t} = P^c_{i,t} + P^d_{i,t}$, where $P^c_{i,t}$ and $P^d_{i,t}$ represent the charging and discharging active power, respectively. Considering that EV cannot charge and discharge simultaneously as well as needs to ensure its battery energy capacity $\overline{E}^{ev}_i$, we have these two mutually exclusive quantities as

$$P^c_{i,t} = [\min(a^p_{i,t}\overline{P}^{ev}_i,\ (\overline{E}^{ev}_i - E^{ev}_{i,t})/(\eta^c_i \Delta t)]^+, \tag{3}$$

$$P^d_{i,t} = [\max(a^p_{i,t}\overline{P}^{ev}_i,\ -E^{ev}_{i,t}\eta^d_i/\Delta t]^-, \tag{4}$$

where operators $[\cdot]^{+/-} = \max / \min\{\cdot, 0\}$, η^c_i, η^d_i are charging and discharging efficiencies. Based on (3)-(4), we have the state transition $E^{ev}_{i,t}$ from time step t to $t+1$ expressed as

$$E^{ev}_{i,t+1} = \begin{cases} E^{ev}_{i,t} + (P^c_{i,t}\eta^c_i + P^d_{i,t}/\eta^d_i)\Delta t & \text{if } u^{ev}_{i,t} = 1 \\ E^{ev}_{i,t} - E^{tp}_i & \text{if } u^{ev}_{i,t} = 0 \end{cases}, \tag{5}$$

where binary $u^{ev}_{i,t} \in \{0, 1\}$ represents the status of agent i connected with the network at time step t, $u^{ev}_{i,t} = 1$ when

connected and $u_{i,t}^{ev} = 0$ when traveling. E_i^{tp} indicates the energy consumption of EV agent i in the traveling.

*4) **Reward**:* At the end of time step t, each agent i obtains its reward $r_{i,t}$. First of all, EVs connected with charging stations in the grid aims to maximize the revenue obtained from providing both active and reactive power services. This can be calculated based on the charging and discharging behaviors (i.e., actions) of EVs as well as the LMPs of charging stations EVs are connected with. Secondly, all EVs need to ensure the sufficient battery energy level for traveling purpose of two typical journeys per day. Since Dec-POMDP is a dynamic decision process, of which the physical constraint (sufficient traveling energy requirement) is a part of environment that can not obtained by the EV agents. A reward shaping mechanism that penalizes the constraint violation to such sufficient charging behavior is needed and can be assumed to be effective to address this challenge. As a result, given the above discussions, the reward function $r_{i,t}$ in equation (6) can be designed as two parts conditioned on two different situations: 1) the cost (revenue) of charging (discharging) active power $P_{i,t}^{ev}$ at P-LMP $\lambda_{i,t}^{P}$ together with the revenue of providing (consuming) reactive power $Q_{i,t}^{ev}$ at Q-LMP $\lambda_{i,t}^{Q}$ when EV is connected to the grid ($u_{i,t}^{ev} = 1$); 2) the penalty of insufficient charging upon departure, i.e., $E_{i,t}^{ev} \geq E_i^{tp}$ may not be satisfied when EV is traveling ($u_{i,t}^{ev} = 0$).

$$r_{i,t} = \begin{cases} -\lambda_{i,t}^{P} P_{i,t}^{ev} + \lambda_{i,t}^{Q} |Q_{i,t}^{ev}| & \text{if } u_{i,t}^{ev} = 1 \\ \kappa [E_{i,t}^{ev} - E_i^{tp}]^{-} & \text{if } u_{i,t}^{ev} = 0 \end{cases}, \tag{6}$$

where $\lambda_{i,t}^{P}, \lambda_{i,t}^{Q}$ are shadow prices (dual variables) of the active and reactive demand-supply equality constraints in the ADN. κ is a penalty factor to penalize the extent of constraint violation, which its sensitivity study will be performed in Section IV-C.

More specifically, we introduce a branch flow algorithm for the ADN operated by DSO [10], [33]. For each time step t, once EV i is parked to its specific charging station (i.e., home or office) at bus set B_{ev} and makes the real-time active and reactive power $P_{i,t}^{ev}, Q_{i,t}^{ev}$, the DSO can solve the following algorithm and calculates the nodal P-LMP and Q-LMP $\lambda_{b,t}^{P}, \lambda_{b,t}^{Q}$.

$$\Big\{ \min_{\Xi^{opf}} \sum_{g \in G} \big(c_g^P P_{g,t}^{dg} + c_g^Q |Q_{g,t}^{dg}| \big) + \sum_{b \in Gird} \big(\pi_t^P P_{b,t}^{gd} + \pi_t^Q Q_{b,t}^{gd} \big), \tag{7}$$

where

$$\begin{aligned} \Xi^{opf} = \{ & P_{b,t}^{gd}, Q_{b,t}^{gd}, P_{g,t}^{dg}, Q_{g,t}^{dg}, P_{g,t}^{res}, Q_{g,t}^{res}, \\ & P_{b,t}^{ex}, Q_{b,t}^{ex}, P_{bp,t}, Q_{bp,t}, V_{b,t}^2, \delta_{bp,t} \}, \end{aligned} \tag{8}$$

subject to

$$\begin{aligned} & \sum_{b \in B_{gd}} P_{b,t}^{gd} + \sum_{g \in B_{dg}} P_{g,t}^{dg} + \sum_{g \in B_{res}} P_{g,t}^{res} = \sum_{d \in B_{ed}} P_{d,t}^{ed} + \\ & \sum_{i \in B_{ev}} P_{i,t}^{ev} - \sum_{(p,b) \in L} P_{pb,t} + \sum_{(b,p) \in L} P_{bp,t} : \ \lambda_{b,t}^{P}, \ \forall b \in B, \end{aligned} \tag{9}$$

$$\begin{aligned} & \sum_{b \in B_{gd}} Q_{b,t}^{gd} + \sum_{g \in B_{dg}} Q_{g,t}^{dg} + \sum_{g \in B_{res}} Q_{g,t}^{res} = \sum_{d \in B_{ed}} Q_{d,t}^{ed} + \\ & \sum_{i \in B_{ev}} Q_{i,t}^{ev} - \sum_{(p,b) \in L} Q_{pb,t} + \sum_{(b,p) \in L} Q_{bp,t} : \ \lambda_{b,t}^{Q}, \ \forall b \in B, \end{aligned} \tag{10}$$

$$\begin{aligned} \nu_{b,t} - \nu_{p,t} = & 2(r_{bp} P_{bp,t} + x_{bp} Q_{bp,t}) \\ & -(r_{bp}^2 + x_{bp}^2) l_{bp,t}, \ \forall bp \in L, \end{aligned} \tag{11}$$

$$P_{bp,t}^2 + Q_{bp,t}^2 \leq l_{bp,t} \nu_{b,t}, \ \forall bp \in L, \tag{12}$$

$$\underline{\nu} \leq \nu_{b,t} \leq \overline{\nu}, \ \forall b \in B, \tag{13}$$

$$l_{bp,t} \leq \bar{l}_{bp}, \ \forall bp \in L, \tag{14}$$

$$\underline{P}_g^{dg} \leq P_{g,t}^{dg} \leq \overline{P}_g^{dg}, \ \forall g \in G, \tag{15}$$

$$(P_{g,t}^{dg})^2 + (Q_{g,t}^{dg})^2 \leq (\overline{S}_g^{dg})^2, \ \forall g \in G, \tag{16}$$

$$|Q_{g,t}^{dg}| \leq P_{g,t}^{dg} \tan(\cos^{-1} \delta_g^{dg}), \ \forall g \in G, \tag{17}$$

$$\underline{P}_b^{gd} \leq P_{b,t}^{gd} \leq \overline{P}_b^{gd}, \ \forall b \in Grid, \tag{18}$$

$$\underline{Q}_b^{gd} \leq Q_{b,t}^{gd} \leq \overline{Q}_b^{gd}, \ \forall b \in Grid, \tag{19}$$

$$(P_{b,t}^{gd})^2 + (Q_{b,t}^{gd})^2 \leq \overline{S}_b^{gd}, \ \forall b \in Grid, \tag{20}$$

$$0 \leq P_{g,t}^{res} \leq \tilde{P}_{g,t}^{res}, \ \forall g \in W, \tag{21}$$

$$(P_{g,t}^{res})^2 + (Q_{g,t}^{res})^2 \leq (\overline{S}_g^{res})^2, \ \forall g \in W, \tag{22}$$

$$|Q_{g,t}^{res}| \leq P_{g,t}^{res} \tan(\cos^{-1} \delta_g^{res}), \ \forall g \in W, \tag{23}$$

$$P_{g,t}^{res} = Q_{g,t}^{res} = 0, \text{if } t \notin T^{res}, \forall g \in W \Big\}, \ t \in T \tag{24}$$

where the objective function (7) towards cost minimization involves two parts: 1) the active power supply cost from DGs given the production cost c_g^P of DG unit g, the active power cost (revenue) by the positive (negative) exchange $P_{b,t}^{gd}$ between the main grid and the ADN at the grid active power price π_t^P; 2) the reactive power support cost from DGs given the reactive power support cost c_g^Q of DG unit g, the reactive power cost (revenue) by the positive (negative) exchange $Q_{b,t}^{gd}$ between the main grid and the ADN at the grid reactive power price π_t^Q.

The problem is then subject to the active and reactive power balances at bus b presented in (9)-(10), where their dual valuables $\lambda_{b,t}^{P}$ and $\lambda_{b,t}^{Q}$ constitute the system P-LMPs and Q-LMPs, respectively. The sets B_{gd}, B_{ed}, B_{dg}, B_{res} and B_{ev} correspond to the bus sets connected with the main grid, EDs, DGs, RESs and EVs located at bus b, respectively. Notably, the active and reactive power of EVs (i.e., $P_{i,t}$ and $Q_{i,t}$) are made by the EVs' actions defined in (2), that can be directly incorporated into (9) and (10) as parameters when solving the branch flow algorithm. Constraint (11) restricts the voltage magnitudes between the two buses, while constraint (12) describes the relationship between line flow, current and voltage. Constraints (13)-(14) are related to the operational constraints of voltage limit at bus b and ampacity limit for line $b-p$, while constraints (15)-(17) describe the feasible region of each DG, including the active and reactive power limits of DG unit g influenced by the rated power factor δ_g^{dg} [33], [34].

Constraints (18)-(20) correspond to the active/reactive power exchange limits between the ADN and main grid respectively. Furthermore, constraint (21) indicates the active power limit of RES g at time step t, which is related to solar irradiation or wind speed. Constraints (22)-(23) impose limits on the apparent power of RESs, reflecting their nameplate capacity $\overline{S}_g^{res}$ and rated power factor δ_g^{res}. Equation (24) imposes zero generation when $\tilde{P}_{g,t}^{res} = 0$, where T^{res} corresponds to the subset of time periods when RES generation is greater than 0.

III. Multi-Agent Reinforcement Learning Method

To efficiently solve the above Dec-POMDP, we propose a novel MARL method named LMP-PSDDPG with its general architecture being shown in Fig. 2. LMP-PSDDPG derives three concrete implementation details that are insightful and particularly critical to our proposed problem: 1) featuring an actor-critic architecture-based DDPG method [31] with its policy (actor) network outputting continuous actions and Q-value (critic) network correcting the weights of policy network; 2) incorporating LMPs into the Q-value function to stabilize the training performance and enhance the training scalability by capturing the system dynamics with privacy perseverance; 3) adopting a parameter sharing (PS) framework [32] to reach the distributed control manner without sharing local information.

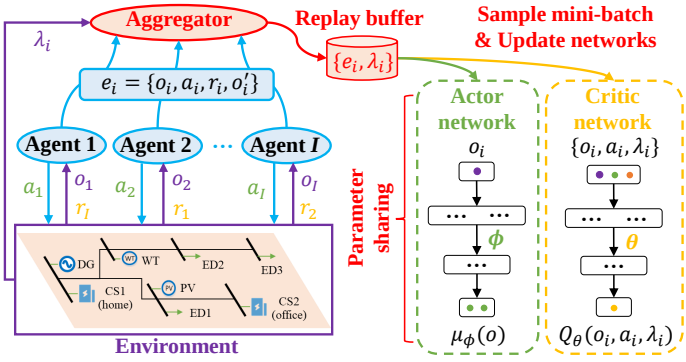

Fig. 2. Architecture of the proposed LMP-PSDDPG method.

A. Deep Deterministic Policy Gradient

The proposed method is constructed based on DDPG that contains two networks for different purposes. The actor network $\mu_{\phi_i}(o_i)$, parameterized by ϕ_i, takes as input the local observation o_i and outputs the continuous action a_i for each agent i. Here, the actor network is based on the deterministic policy gradient theorem [35] that specifies the current policy by deterministically mapping observation to a specific action. The critic network $Q_{\theta_i}(o_i, a_i)$, parameterized by θ_i, takes as input the concatenation of local observation o_i and executed action a_i of agent i, and outputs a scalar estimate of the Q-value to perform the policy evaluation task. More specifically, we update the weights of critic network with temporal difference (TD) learning [12] as

$$\mathcal{L}(\theta_i) = \mathbb{E}\big[(r_i + \gamma Q'_{\theta'_i}(o'_i, a'_i) - Q_{\theta_i}(o_i, a_i))^2\big], \tag{25}$$

where $Q'_{\theta'_i}(\cdot)$ is the target critic network whose parameters θ'_i are updated by having them softly tracking the online critic network θ_i, to give consistent target during TD learning. Furthermore, a'_i is the executed action according to the next observation o'_i of agent i using its target actor network $\mu'_{\phi'_i}(o'_i)$ whose parameters ϕ'_i are also updated by having them softly tracking the online actor network ϕ_i. Different from TD learning in critic network, the actor network is updated via the policy gradient theorem [12] as

$$\nabla_{\phi_i} J(\mu_{\phi_i}) = \nabla_{\phi_i} \mu_{\phi_i}(o_i) \nabla_{a_i} Q_{\theta_i}(o_i, a_i)|_{a_i = \mu_{\phi_i}(o_i)}. \tag{26}$$

B. Capturing System Dynamics via Location Marginal Prices

Directly applying DDPG to the multi-agent setup may become problematic, since independently learning algorithm treating other agents as part of the environment appears non-stationary from the view of any agent. As a result, a centralized Q-value $Q_{\theta_i}(o_{1:I}, a_{1:I})$ with access to all agents' local observations and actions is introduced in MADDPG [36] and is widely used in the existing MARL algorithms that can stabilize the training performance. However, it is difficult to directly acquire other agents' local observations and actions in our proposed problem, since EVs with privacy concern are not willing to exchange their traveling patterns and charging activities with each other. To this end, this paper assumes the aggregator as a trusted third party that can make use of the LMPs and incorporate them into the centralized Q-value function to epitomize the key information of the system dynamics. In this context, the centralized Q-value function for each agent i can be approximated as

$$Q_{\theta_i}(o_{1:I}, a_{1:I}) \approx Q_{\theta_i}(o_i, a_i, \lambda_i), \tag{27}$$

where $\lambda_i = [\lambda_i^P, \lambda_i^Q]$ denote the P-LMP and Q-LMP for agent i. It can be observed that λ_i is an embedded function that captures the system dynamics through the branch flow algorithm. First, the observations of nodal demand and PV generation as well as the actions of EVs active and reactive power are integrated into the active and reactive power balances (9) and (10). Second, λ_i is a systematic index that not only represents the local demand-supply status but also is influenced by the other agents' information and activities. Third, λ_i is critical for agents to adjust their control behaviors. In detail, the higher value of λ_i^P encourages EV agent i to reduce/increase its active charging/discharging behavior in G2V/V2G status, and vice versa. Similarly, the higher value of λ_i^Q encourages EV agent i to inject or consume reactive power into the grid for voltage regulations. As a result, incorporating λ_i into the centralized Q-value function, agent i can make informed decisions on the basis of the impact of other agents' actions in the ADN, albeit not knowing their specific information, thereby protecting the EV privacy and also improving the scalability.

C. Parameter Sharing Framework

Since we consider a Dec-POMDP of I agents with the same observation, action and reward function, their policies can be trained with enhanced efficiency by using a PS framework [32]. PS allows all agents to share the parameters of a single control policy. This enables the shared policy to be trained with the sample experiences gathered by all agents,

while still allowing different behaviors among different agents, since each agent receives different local observations. In order to realize this framework, we assume that the experiences acquired from the environment of all local EV agents are transmitted to the central aggregator for updating the shared policy $\mu_\phi = \mu_{\phi_i}, \forall i \in \mathcal{I}$ parameterized by ϕ. This policy μ_ϕ is then broadcast to all local EV agents to compute actions executed to the environment. Similarly, the critic network used to estimate Q-value function can be also trained in a PS framework $Q_\theta = Q_{\theta_i}, \forall i \in \mathcal{I}$ parameterized by θ.

LMP-PSDDPG is an off-policy MARL method that requires the past experiences to update the networks. Thus, an experience replay buffer $\mathcal{D}$ is employed. The buffer is a cache storing the past experiences of all agents acquired from the environment. In detail, an experience is a transition tuple that contains $e_{i,t} = (o_{i,t}, a_{i,t}, r_{i,t}, o_{i,t+1})$ used to update policy and $\lambda_{i,t}$ used to abstract the Q-value function. On every iteration of training process, we sample uniformly a minibatch of J mixed experiences from the shared replay buffer $\{(e_j, \lambda_j)\}_{j=1}^{J} \sim \mathcal{D}$ to compute the mean-squared TD error of online critic network

$$\mathcal{L}(\theta) = \frac{1}{J}\sum_{j=1}^{J}\Big[\big(y_j - Q_\theta(o_j, a_j, \lambda_j)\big)^2\Big], \tag{28}$$

where the target Q-value

$$y_j = r_j + \gamma Q'_{\theta'}\big(o_{j+1}, \mu'_{\phi'}(o_{j+1}), \lambda_{j+1}\big), \tag{29}$$

here $Q'_{\theta'}(\cdot)$ and $\mu'_{\phi'}(\cdot)$ are respectively the target critic and actor networks, softly updated with their online networks. λ_{j+1} is calculated from the branch flow algorithm given the target actions $\mu'_{\phi'}(o_{j+1})$ conditioned on next observations o_{j+1}.

On the other hand, the online actor network employs the policy gradient theorem, which can be expressed as

$$\nabla_\phi J(\mu_\phi) = \frac{1}{J}\sum_{j=1}^{J}\big[\nabla_\phi \mu_\phi(o_j)\nabla_{a_j} Q_\theta(o_j, a_j, \lambda_j)|_{a_j=\mu_\phi(o_j)}\big]. \tag{30}$$

The following updates are then applied to the weights of the online and target networks, where $\alpha^\theta, \alpha^\phi$ are the learning rates of gradient descent algorithm for online critic and actor networks, and τ is the soft update rate for target networks.

$$\theta \leftarrow \theta - \alpha^\theta \nabla_\theta \mathcal{L}(\theta) \quad \text{and} \quad \theta' \leftarrow \tau\theta + (1-\tau)\theta', \tag{31}$$

$$\phi \leftarrow \phi + \alpha^\phi \nabla_\phi J(\mu) \quad \text{and} \quad \phi' \leftarrow \tau\phi + (1-\tau)\phi'. \tag{32}$$

Moreover, in order to help the agents explore the environment and acquire more valuable experiences, we add a random Gaussian noise $\mathcal{N}(0, \sigma_{i,t}^2)$ to the online actor network (policy) $\mu_\phi(o_{i,t})$, constructing an exploration policy

$$\hat{\mu}(o_{i,t}) = \mu_\phi(o_{i,t}) + \mathcal{N}(0, \sigma_{i,t}^2). \tag{33}$$

Finally, the pseudo-code of the proposed LMP-PSDDPG is presented in Algorithm 1:

Algorithm 1 LMP-PSDDPG for I agents

```
Initialize parameters θ, φ for online shared networks and copy them to their respective target network weights θ', φ'
Initialize a shared replay buffer D for all agents
for episode (i.e., day) = 1 to M do
    Initialize a random process N(0, σ²_{i,t}) for action exploration
    Initialize global state s_0 and local observation o_{i,0}
    for time step (i.e., hour) t = 1 to T do
        For each agent i, selects action a_{i,t} = μ̂(o_{i,t}) according to current observation o_{i,t} using (33)
        Execute all agents' actions a_t = [a_{1,t}, ..., a_{I,t}] to ADN
        DSO solves branch flow algorithm (7)-(24) and obtains LMPs λ_{b,t} = [λ^P_{b,t}, λ^Q_{b,t}] (dual variables of (9)-(10))
        For each agent i, observes current reward r_{i,t} and next observation o_{i,t}, then transits local experience e_{i,t} = (o_{i,t}, a_{i,t}, r_{i,t}, o_{i,t+1}) to EV aggregator
        EV aggregator concatenates LMPs λ_{i,t} and local experience e_{i,t} together and store them to buffer D
        Update state s_t ← s_{t+1} and observation o_{i,t} ← o_{i,t+1}
        Sample a minibatch (e_j, λ_j)_{j=1}^J from D
        Update online critic and critic networks in (28) and (30)
        Update weights of online and target networks in (31)-(32)
    end for
end for
```

IV. Case Studies

A. Experimental Setup

Case studies are carried out on an IEEE 15-bus distribution network modified from [37], which includes 8 EDs, 1 DG, 2 PVs, 2 WTs and 2 charging stations (home and office) for EVs. More specifically, its network topology is provided in Fig. 3. We implement all the experiments on a real-world open-source dataset recorded from Open Energy Data Initiative (OEDI) [38]. We collect the corresponding electric loads and PV & wind power generations of residential and commercial areas with hourly resolution for over a yearly horizon our experiments. The electricity prices are collected from Nord-Pool group [39]. To evaluate the proposed MARL method, we split these time-series data into the training (first 11 months) and test (last 1 month) sets. The apparent power capacities of 2 PVs and 1 WT are set as 400 kVA and 50 kVA, respectively. A power factor of 0.95 (0.85) is assumed for residential (commercial) nodes [10]. Nodal voltage level is limited between 0.95 and 1.05 p.u. [10]. The technical parameters of 2 (identical) DGs and 2 real-world EV models (Tesla Model-S and Nio ES 8) are provided in Table I and II, respectively. Here, the initial battery energy levels and the departure time of two trips are both constructed from the truncated normal distributions representing the EVs traveling uncertainties. In addition, the reactive power cost is assumed equal to 10% the value of active cost for both grid prices and DG generation costs [10].

TABLE I
Technical Parameters of Identical DG

$\underline{P}^{dg}$ (kW)	$\overline{P}^{dg}$ (kW)	$\overline{S}^{dg}$ (kVA)	c^P (£/kWh)
0	400	400	0.055

We compare the proposed LMP-PSDDPG with three benchmark MARL methods: 1) **DDPG** [31]: each agent adopts a DDPG method independently, the critic network does not

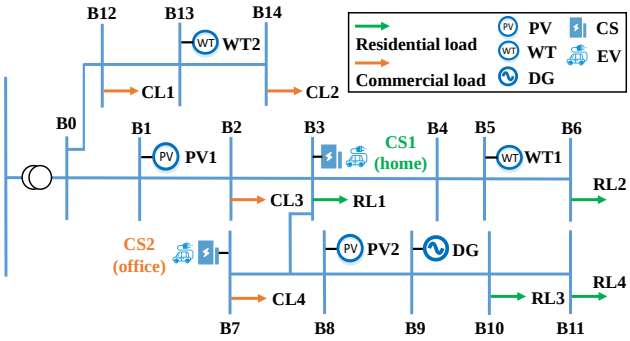

Fig. 3. Network structure of the modified 15-bus power system.

TABLE II
TECHNICAL PARAMETERS OF 2 REAL-WORLD EV MODELS

Parameters	Tesla Model S	Nio ES 8
$\overline{S}^{ev}$ (kVA)	16.7	17.6
$\overline{P}^{ev}$ (kW)	16.7	17.6
$\overline{E}^{ev}$ (kWh)	100	100
η^c/η^d (%)	90	90
$E^{ev}_{t=0}$ (kWh)	$\mathcal{TN}(30, 10^2, 10, 50)$	$\mathcal{TN}(30, 10^2, 10, 50)$
Departure time from home (h)	$\mathcal{TN}(7, 2^2, 5, 9)$	$\mathcal{TN}(7, 2^2, 5, 9)$
Departure time from office (h)	$\mathcal{TN}(18, 2^2, 16, 20)$	$\mathcal{TN}(18, 2^2, 16, 20)$
E^{tp} (kWh)	16.7	17.6

incorporate P-LMPs and Q-LMPs into the Q-value function, there is thus no PS framework, each agent trains its own policy based on its individual experiences; 2) **TD3** [40]: based on DDPG, another critic network is applied to obtain the target Q-value (taking the minimum value between these two estimates), with the aim of overcoming the Q-value overestimation issue in DDPG and stabilize the training performance. 3) **LMP-DDPG**: based on DDPG, each agent incorporate P-LMPs and Q-LMPs into its individual Q-value function without the shared experience replay buffer and policy. For each MARL method, we run 1,000 episodes with the same 10 random seeds for environment and weights initialization.

All MARL methods use the parameters from original DDPG paper [31]. Adam optimizer is used for both actor and critic networks with learning rates $\alpha^{\phi} = 10^{-4}$ and $\alpha^{\theta} = 10^{-3}$, respectively. The soft update rate $\tau = 10^{-2}$. A discount factor of $\gamma = 0.99$ is used for the critic network. The minibatch size $J = 64$ and the replay buffer size $|\mathcal{D}| = 10^6$. For both actor and critic networks we use Multilayer Perceptrons with two hidden layers with 400 and 300 units, respectively.

B. Training and Test Performance

This section lies in comparing the training and test performance of four examined MARL methods in terms of policy quality and convergence speed for the proposed problem. Fig. 4 illustrates the convergence curve of episodic reward of 3 utilized EVs for different MARL methods, where the solid lines and the shaded areas respectively depict the moving average over 50 episodes (smoothing learning curves) and the oscillations of the original reward, dots on the lines indicate the numbers of episodes to reach convergence. Finally, Tables III and IV present the number of episodes to reach convergence during the training process and the averaged cumulative rewards over 31 test days for 3 EVs, respectively.

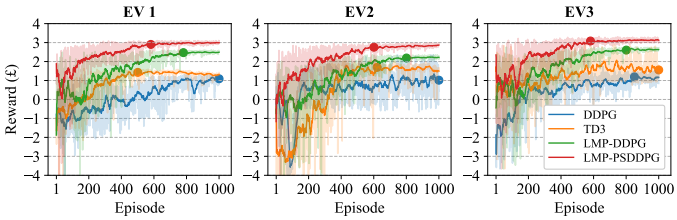

Fig. 4. Episodic reward of 3 EVs for different MARL methods.

TABLE III
CONVERGED EPISODES OF 3 EVS FOR DIFFERENT MARL METHODS

Method	EV1	EV2	EV3	In Average
DDPG	1000*	1000*	850	950
TD3	500	620	1000+	707
LMP-DDPG	780	800	800	793
LMP-PSDDPG	580	600	600	593

* EV1 and EV2 fail to reach convergence.
+ EV3 fails to reach convergence.

The first observation in Fig. 4 is that DDPG (blue) exhibits the lowest reward levels and the highest oscillations for all 3 EVs, where EV1 and EV2 even fail to reach convergence within 1,000 episodes, EV3 converges within 850 episodes. This is because the independent learning algorithm that focuses on local information while ignoring the system dynamics, rendering the environment non-stationary and consequently an unstable learning behavior. Although TD3 (orange) is able to improve the policy quality and stabilize the training performance given by its double critic networks, EV3 still fails to reach convergence within 1,000 episodes. In addition, the rewards of all 3 EVs are relatively low. In this context, LMP-DDPG (green) integrating with LMPs (capturing system dynamics) can effectively mitigate such non-stationary issue and exhibits the superior performance in both policy quality and stability compared to DDPG and TD3. However, LMP-DDPG relies on individual replay buffer to update the policy that may not fully explore the environment, thereby resulting in the sub-optimal policy and the slow convergence speed. To this end, our proposed LMP-PSDDPG (red) owing to its PS framework learns a shared policy from a shared replay buffer. As a result, the experiences acquired from the environment by all agents can be used to update one common policy. In this case, more valuable experiences are possible to be used to improve the policy quality and the shared policy is more frequent to be updated to accelerate the convergence speed.

TABLE IV
CUMULATIVE REWARDS OF 3 EVS OVER 31 TEST DAYS FOR DIFFERENT MARL METHODS

Method	DDPG	TD3	LMP-DDPG	LMP-PSDDPG
Test reward (£)	108	165	219	279

In test process, we first freeze and load the learned weight parameters of actor networks in all four MARL methods, and then apply them to determine the active and reactive power

schedules for each EV across the 31 test days based on its local observations. In other words, during test process, the agents do not communicate with the aggregator, the decision-making process is performed in a fully distributed and privacy-preserving fashion via the deployed actor networks. The numerical results in Table IV show that the test rewards of all 3 EVs are the highest in LMP-PSDDPG, followed by LMP-DDPG, TD3, and the lowest in DDPG, which exhibits the same trend in the training performance.

In relative terms, the proposed LMP-PSDDPG achieves 38%, 16%, and 25% speedup (Table III), and 1.58, 0.69, and 0.27 times higher test reward (Tables IV) over the benchmarks DDPG, TD3, and LMP-DDPG, respectively.

C. Sensitivity on Penalty Factor in Reward Function

Since MARL is a model-free control algorithm that cannot handle the physical constraint (e.g., traveling energy requirement $E_{i,t}^{ev} \geq E_i^{tp}$ when EV i is departing) in conventional optimization approach, penalizing such constraint violation thus is an effective way to address such issue. To this end, selecting a suitable penalty factor κ (in £/kWh) becomes important, since a relatively small value of κ is not efficient to penalize the constrain violation; while a relatively large value of κ may destroy the policy quality of original MARL method. This section aims to investigate the impact of penalty factor κ in the reward function (6) by doing the sensitivity analysis on a set of values κ = [0, 0.1, 0.5, 1.0, 5.0]. In detail, Fig. 5 shows the converged reward (blue) and constraint violation (orange) of 3 EVs (in total) for different values of κ.

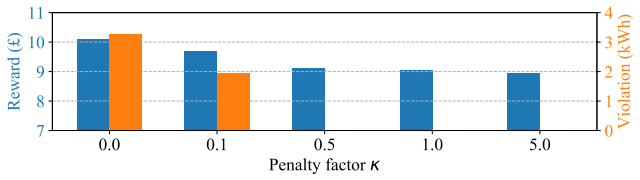

Fig. 5. Converged reward and constraint violation of 3 EVs (in total) for different penalty factors.

It can be seen that the constraint violation and reward both decrease with the increase of penalty factor κ. When $\kappa \geq 0.5$, the constraint violation will be significantly reduced to zero, which means the constraint of traveling energy requirement is completely guaranteed. It can be further observed that reward is the highest when $\kappa = 0$, this is because EVs (saving cost) never charge for the traveling purpose, but only for the power supply. More specifically, the reward decreases significantly as the penalty factor κ increases, until it decreases slightly when $\kappa \geq 0.5$. The continued downward trend of reward indicates that the control policy is not being optimal, although the constraint is fully satisfied. To this end, we would like to select $\kappa = 0.5$ as the suitable value for our experiment.

D. Active and Reactive Power Analysis

This section lies in analyzing the learned policy of LMP-PSDDPG for 3 EVs' active and reactive power schedules as well as their daily SoC dynamics for one test day, as depicted in Fig. 6. Additionally, the corresponding results of P-LMPs and Q-LMPs, voltage profiles at bus 3 (home) and bus 7 (office), as well as the demand-supply balances of system active and reactive power are illustrated in Fig. 7.

Regarding the active power control, 3 EVs mainly choose to charge power in the morning (at home) and mid-day (at office), as depicted in Fig. 6(a-c). The charging behaviors in these two periods can increase their SoC (Fig. 6(g-i)) to ensure sufficient energy for the two trips of going to work and returning home, respectively. From the economic perspective, EVs are trying to maximize the energy arbitrage by firstly charging power in the low P-LMP periods of morning and mid-day, and then discharging power in the high P-LMP periods of evening, as depicted in Fig. 6(a-c). It is noted that P-LMPs in the mid-day periods of hours 11-14 are zero (Fig. 7(a)), since the PV generation is extremely large during this period. In this case, the system demand in active power is fully supplied by 100% PV resources (Fig. 7(c)) with zero P-LMPs.

On the other hand, Q-LMPs follow the similar trends to the P-LMPs, high in the morning and evening and zero in the mid-day, as shown in Fig. 7(a). This is driven by the reactive power injection from PV resources into the grid, as shown in Fig. 7(d). Observing such Q-LMP trends, all 3 EVs aiming to obtain higher revenue choose to provide reactive power support for the system in the morning and evening as well as several hours in the afternoon before returning home (Fig. 6(d-f)) when Q-LMPs are relatively high.

Go further, let us analyze the voltage profiles of buses 3 (home) and 7 (office). As shown in Fig. 7(b), the voltages at buses 3 and 7 exhibit the complementary profiles in the morning and evening. More specifically, the voltages are much higher at bus 3 than these at bus 7 for these two periods, since EVs parking at bus 3 can provide sufficient reactive power for the grid in the morning and evening (Fig. 6(d-f)). However, the reactive power in the mid-day is fully supported by renewable energy resources (Fig. 7(d)), the voltages thus in the mid-day are much flatter and exhibit similar levels at both buses 3 and 7. The interesting results show that the voltages at bus 3 (blue) dramatically decrease from hours 8 to 9 when EVs leave home, and further decrease during hours 16-18 when there are almost no PV resources, but start increasing after hour 18 when EVs park at home to support reactive power. In contrast, the voltages at bus 7 (orange) dramatically increase after hour 7 when PV resources are abundant, and further increase from hours 14 to 17 when EVs park at office to inject a certain level of reactive power, but start dramatically decreasing after hour 17 when both EVs (leaving office) and PV (the sun going down) cannot provide reactive power support.

Thus, it can be found that the proposed LMP-PSDDPG is able to learn an effective active and reactive power scheduling policy for all 3 EVs to coordinately integrate with the DSO.

E. Scalability Analysis and Evaluation of EV Benefits

To further analyze the scalability of the proposed LMP-PSDDPG and quantify the benefits of EVs on system active and reactive power control, a detailed comparison for different numbers of EVs (3, 30 and 100) is developed in this section,

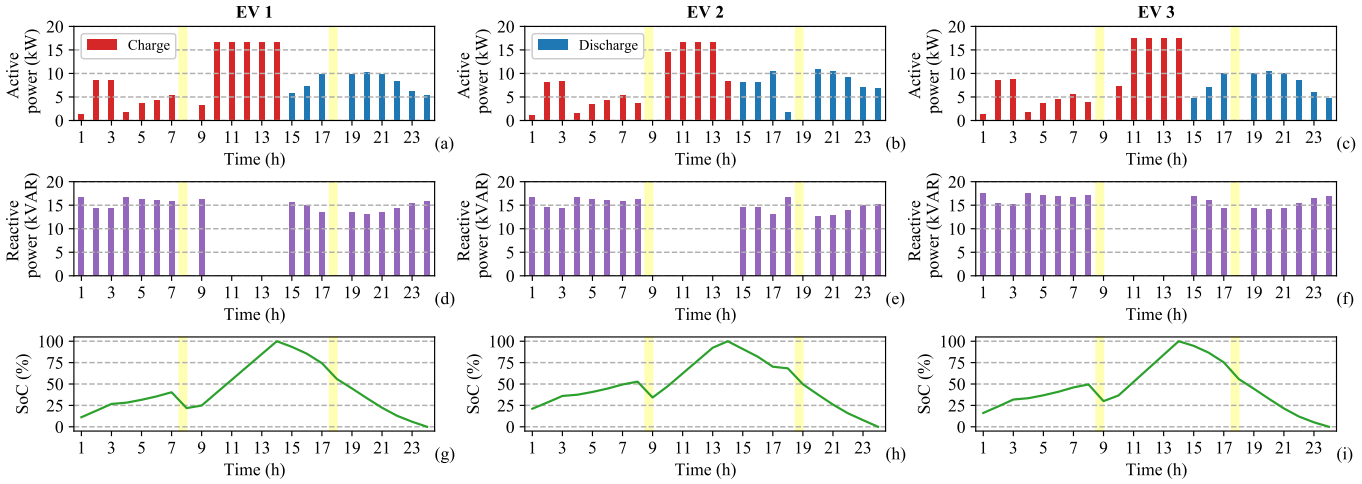

Fig. 6. Active (a-c) and reactive (d-f) power schedules as well as SoC dynamics (g-i) of 3 EVs, yellow areas indicate the mobility of EVs on the travelling.

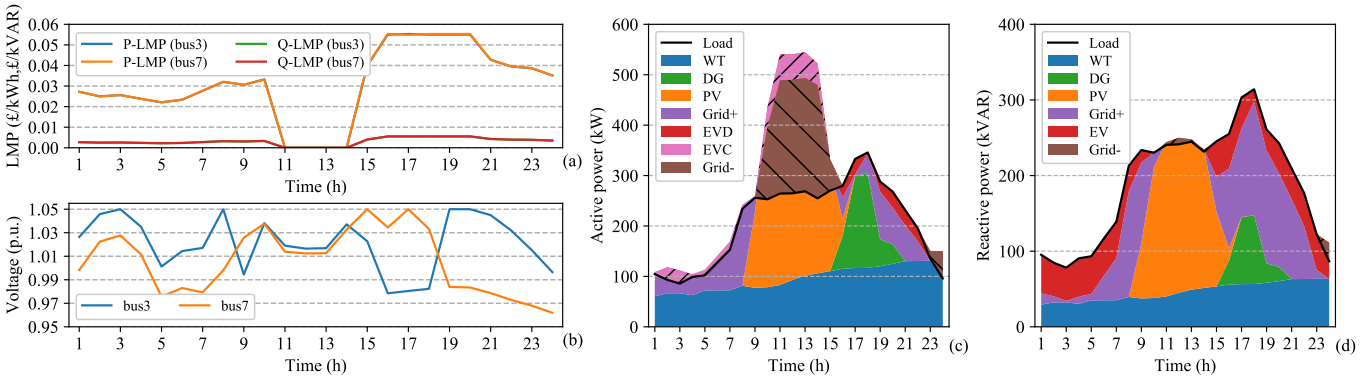

Fig. 7. P-LMPs and Q-LMPs at buses 3 and 7 (a), voltage profiles at buses 3 and 7 (b), active (c) and reactive (d) power supply for the ADN.

where the aggregated active (a) and reactive (b) power schedules of all EVs, the net demand and generation profiles of the ADN (c), as well as the voltage profiles at home bus 3 (d) and office bus 7 (e) for different EV numbers are illustrated in Fig. 8. It is noted that the proposed LMP-PSDDPG is constructed on a PS framework, which means the trained control policy (i.e., actor network) is common and can used by all EVs once the policy is well trained. As a result, we do not need to retrain LMP-PSDDPG for 30 and 100 EV cases, while directly using the policy trained in Section IV-B.

TABLE V
BUSINESS CASES OF EVS, DSO, AND SOCIAL WELFARE (SUM OF EVS NET PROFIT AND DSO NEGATIVE COST) FOR DIFFERENT EV NUMBERS

Number of EVs	Active revenue (£)	Reactive revenue (£)	EV net profit (£)	DSO cost (£)	Social welfare (£)
0	-	-	-	130.93	-130.93
3	6.24	3.18	9.42	105.66	-96.24
30	40.71	16.90	57.61	36.11	21.50
100	57.27	20.23	77.50	13.16	64.34

Firstly, it can be observed from Fig. 8(a) that both active charging and discharging power are enhanced when EV numbers are increased to 30 and 100, which results in the significant increase of off-peak demand in the morning and the higher effect of absorbing PV resources in the mid-day as well as the significant reduction of peak demand in the evening, as shown in Fig. 8(c). The interesting results show that DSO starts exporting power to the main grid in the evening under the 30 and 100 EVs scenarios (orange and green line in Fig. 8(c)), since the EV flexibility exceeds the system peak demand and EVs can sell their surplus back to the grid at high P-LMP periods to obtain the extra discharging revenue (Table V). Besides the economic benefits of EVs themselves, DSO also benefits from the large-scale EVs deployment in reducing RES curtailment in the mid-day, and can acquire more economic benefits at 36.11 £ and 13.16 £ total cost (Table V), respectively.

Secondly, it can be observed from Fig. 8(b) that the reactive power support is also enhanced when EV numbers are increased to 30 and 100. However, the flexibility of EVs' reactive power is not significantly utilized as the case in active power, since voltages are already at relatively high levels and even reach to 1.05 p.u. for many periods in the morning and evening when EVs park at bus 3, as shown in Fig. 8(d). On the other hand, the voltages are also increased at bus 7 in the morning and evening under 30 and 100 EVs scenario, which further demonstrates the benefit of EV flexibility in improving system overall voltage quality.

Finally, it can observed from Table V that the business

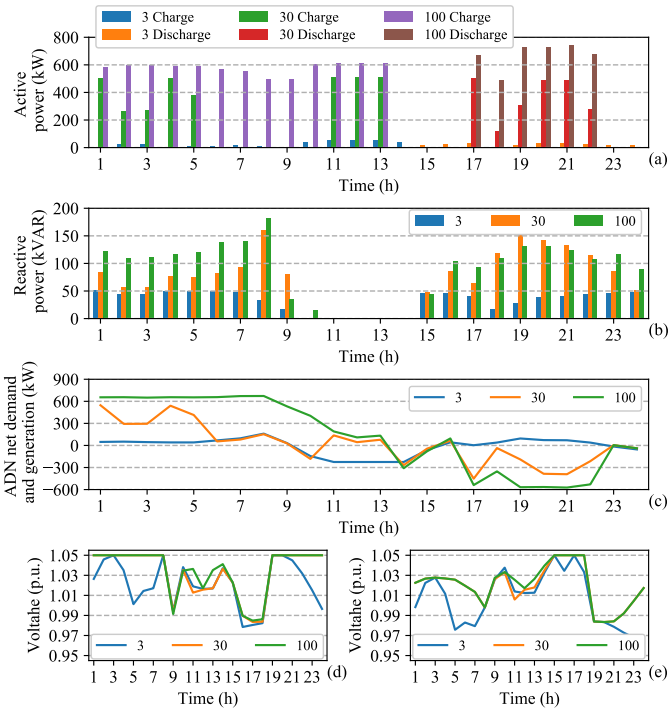

Fig. 8. EVs' active (a) and reactive (b) power, system demand/generation profiles (c), voltage levels at bus 3 (d) and bus 7 (e) for different EV numbers.

cases are all increased under the 30 and 100 EVs scenarios. However, these quantities do not expect to raise proportionally as the EV numbers from 3 to 30 and 100, since the DSO does not require much flexibility to support its active and reactive power requirements. Nevertheless, the DSO still benefit a lot from the large-scale deployment of EVs in reducing its operation cost. Such effect together with EVs' net profit also significantly increase the social welfare, resulting in more efficient network operation. It can be concluded that our proposed LMP-PSDDPG is able to learn an effective active and reactive control policy to various EV numbers, thereby evaluating its scalability performance.

V. Conclusions

This paper has proposed a novel MARL method named LMP-PSDDPG to address the coordinated active and reactive power scheduling problem of multiple self-dispatched EVs towards both demand-side response and voltage regulations. The proposed MARL method employs a PS framework incorporating LMPs to reach a distributed control matter, enhance the training scalability and preserve the EVs' privacy. Uncertainties associated with RESs, demand and EV traveling patterns are encompassed in a real-world open-source dataset through the training procedure of MARL. Experiment results based on a modified IEEE 15-bus network demonstrate the effectiveness of EVs in providing demand-side response, regulating voltage profiles, and improving social welfare, while the superior performance of the proposed LMP-PSDDPG method in optimality, stability and scalability with respect to the state-of-the-art MARL methods has been testified.

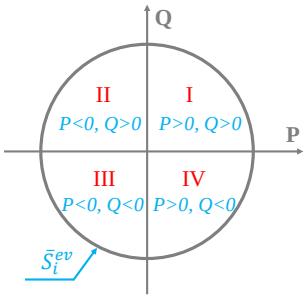

Fig. 9. Capability curve of an EV charger.

Appendix A
EV Capability Curve

The chargers used for EVs in charging stations include two bidirectional converters, i.e., AC/DC and DC/DC converters. In general, the full-bridge converter and the half-bridge bidirectional (buck/boost) converter are used to AC/DC and DC/DC converters, respectively [6]. These converters facilitate active and reactive power control of EVs that could be used in the ADN operation, as depicted in Fig. 9. The capability curve of an EV charger includes four modes: I) charging and inductive mode, II) discharging and inductive mode, III) discharging and capacitive mode, and IV) charging and capacitive mode [8].

As illustrated in Fig. 9 together with the model presented in [10], [17], EV scheduling is constrained by its active power limit $-\overline{P}_i^{ev} \leq P_{i,t}^{ev} \leq \overline{P}_i^{ev}$ and apparent power limit $(P_{i,t}^{ev})^2+(Q_{i,t}^{ev})^2 \leq (\overline{S}_i^{ev})^2$. Since RL method cannot deal with the above two coupled operation constraints simultaneously, we use two-step derivations to sequentially calculate the lower and upper bounds of active and reactive power $P_{i,t}^{ev}, Q_{i,t}^{ev}$ given the actions $a_{i,t}^p, a_{i,t}^q$ as well as the active power capacity $\overline{P}_i^{ev}$ and apparent power capacity $\overline{S}_i^{ev}$.

References

[1] G. Strbac, *et al.*, "Cost-effective decarbonization in a decentralized market: The benefits of using flexible technologies and resources," *IEEE Power Energy M.*, vol. 17, no. 2, pp. 25–36, Mar.-Apr. 2019.

[2] T. Xu and W. Wu, "Accelerated admm-based fully distributed inverter-based volt/var control strategy for active distribution networks," *IEEE Trans. Industr. Inform.*, vol. 16, no. 12, pp. 7532–7543, Dec. 2020.

[3] M. H. K. Tushar, A. W. Zeineddine, and C. Assi, "Demand-side management by regulating charging and discharging of the ev, ess, and utilizing renewable energy," *IEEE Trans. Industr. Inform.*, vol. 14, no. 1, pp. 117–126, Jan. 2018.

[4] P. Andrianesis and M. Caramanis, "Distribution network marginal costs: enhanced ac opf including transformer degradation," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 3910–3920, Sep. 2020.

[5] S. Pirouzi, M. A. Latify, and G. R. Yousefi, "Conjugate active and reactive power management in a smart distribution network through electric vehicles: A mixed integer-linear programming model," *Sustain. Energy, Grids Netw.*, vol. 22, p. 100344, Jun. 2020.

[6] S. Pirouzi, J. Aghaei, M. A. Latify, G. R. Yousefi, and G. Mokryani, "A robust optimization approach for active and reactive power management in smart distribution networks using electric vehicles," *IEEE Syst. J.*, vol. 12, no. 3, pp. 2699–2710, Sep. 2018.

[7] Q. Hu, S. Bu, and V. Terzija, "A distributed p and q provision based voltage regulation scheme by incentivized ev fleet charging for resistive distribution networks," *IEEE Trans. Transp. Electrification*, vol. 7, no. 4, pp. 2376–2389, Dec. 2021.

[8] J. Wang, G. R. Bharati, S. Paudyal, O. Ceylan, B. P. Bhattarai, and K. S. Myers, "Coordinated electric vehicle charging with reactive power support to distribution grids," *IEEE Trans. Industr. Inform.*, vol. 15, no. 1, pp. 54–63, Jan. 2019.

[9] W. Zhang, O. Gandhi, H. Quan, C. D. Rodríguez-Gallegos, and D. Srinivasan, "A multi-agent based integrated volt-var optimization engine for fast vehicle-to-grid reactive power dispatch and electric vehicle coordination," *Appl. Energy*, vol. 229, pp. 96–110, Nov. 2018.

[10] P. Andrianesis, M. Caramanis, and N. Li, "Optimal distributed energy resource coordination: A decomposition method based on distribution locational marginal costs," *IEEE Trans. Smart Grid*, vol. 13, no. 2, pp. 1200–1212, Mar. 2022.

[11] A. Zahedmanesh, K. M. Muttaqi, and D. Sutanto, "Active and reactive power control of pev fast charging stations using a consecutive horizon-based energy management process," *IEEE Trans. Industr. Inform.*, vol. 17, no. 10, pp. 6742–6753, Oct. 2021.

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[13] H. M. Abdullah, A. Gastli, and L. Ben-Brahim, "Reinforcement learning based ev charging management systems–a review," *IEEE Access*, vol. 9, pp. 41 506–41 531, Mar. 2021.

[14] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time ev charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246–5257, Sep. 2019.

[15] 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 Trans. Smart Grid*, vol. 11, no. 2, pp. 1714–1723, Mar. 2019.

[16] X. Wang, J. Wang, and J. Liu, "Vehicle to grid frequency regulation capacity optimal scheduling for battery swapping station using deep q-network," *IEEE Trans. Industr. Inform.*, vol. 17, no. 2, pp. 1342–1351, Feb. 2021.

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

[18] Y. Tao, J. Qiu, and S. Lai, "Deep reinforcement learning based bidding strategy for evas in local energy market considering information asymmetry," *IEEE Trans. Industr. Inform.*, vol. 18, no. 6, pp. 3831–3842, Jun. 2022.

[19] Y. Tao, J. Qiu, S. Lai, X. Zhang, Y. Wang, and G. Wang, "A human-machine reinforcement learning method for cooperative energy management," *IEEE Trans. Industr. Inform.*, vol. 18, no. 5, pp. 2974–2985, May 2022.

[20] F. Zhang, Q. Yang, and D. An, "Cddpg: A deep-reinforcement-learning-based approach for electric vehicle charging control," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3075–3087, Mar. 2020.

[21] D. Qiu, Y. Ye, D. Papadaskalopoulos, and G. Strbac, "A deep reinforcement learning method for pricing electric vehicles with discrete charging levels," *IEEE Trans. Ind. Appl.*, vol. 56, no. 5, pp. 5901–5912, Apr. 2020.

[22] J. Jin and Y. Xu, "Optimal policy characterization enhanced actor-critic approach for electric vehicle charging scheduling in a power distribution network," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1416–1428, Mar. 2021.

[28] M. Shin, D.-H. Choi, and J. Kim, "Cooperative management for pv/ess-enabled electric vehicle charging stations: A multiagent deep reinforcement learning approach," *IEEE Trans. Industr. Inform.*, vol. 16, no. 5, pp. 3493–3503, May. 2019.

[23] 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 Trans. Smart Grid*, vol. 12, no. 6, pp. 5124–5134, Nov. 2021.

[24] Y. Zhang, Z. Zhang, Q. Yang, D. An, D. Li, and C. Li, "Ev charging bidding by multi-dqn reinforcement learning in electricity auction market," *Neurocomputing*, vol. 397, pp. 404–414, Jul. 2020.

[25] L. Yan, X. Chen, Y. Chen, and J. Wen, "A cooperative charging control strategy for electric vehicles based on multi-agent deep reinforcement learning," *IEEE Trans. Industr. Inform.*, 2022.

[26] D. Qiu, Y. Wang, M. Sun, and G. Strbac, "Multi-service provision for electric vehicles in power-transportation networks towards a low-carbon transition: A hierarchical and hybrid multi-agent reinforcement learning approach," *Appl. Energy*, vol. 313, p. 118790, May 2022.

[27] S. Li, W. Hu, D. Cao, Z. Zhang, Q. Huang, Z. Chen, and F. Blaabjerg, "A multi-agent deep reinforcement learning-based approach for the optimization of transformer life using coordinated electric vehicles," *IEEE Trans. Industr. Inform.*, 2022.

[29] Y. Lu, Y. Liang, Z. Ding, Q. Wu, T. Ding, and W.-J. Lee, "Deep reinforcement learning based charging pricing for autonomous mobility-on-demand system," *IEEE Trans. Smart Grid*, vol. 13, no. 2, pp. 1412–1426, Mar. 2022.

[30] F. A. Oliehoek and C. Amato, *A concise introduction to decentralized POMDPs*. Springer, 2016.

[31] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, May. 2016, pp. 1–14.

[32] J. K. Terry, N. Grammel, A. Hari, and L. Santos, "Parameter sharing is surprisingly useful for multi-agent deep reinforcement learning," *arXiv e-prints*, pp. arXiv–2005, 2020.

[33] T. Ding, Y. Lin, Z. Bie, and C. Chen, "A resilient microgrid formation strategy for load restoration considering master-slave distributed generators and topology reconfiguration," *Appl. Energy*, vol. 199, pp. 205–216, Aug. 2017.

[34] L. Bai, J. Wang, C. Wang, C. Chen, and F. Li, "Distribution locational marginal pricing (dlmp) for congestion management and voltage support," *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 4061–4073, Jul. 2017.

[35] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Machine Learn. (ICML)*, Beijing, China, Jun. 2014, pp. 387–395.

[36] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, pp. 6379–6390, 2017.

[37] A. Papavasiliou, "Analysis of distribution locational marginal prices," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4872–4882, Sep. 2017.

[38] Office of Energy Efficiency & Renewable Energy, "Commercial and Residential Hourly Load Profiles for all TMY3 Locations in the United States." [Online]. Available: https://openei.org/datasets/dataset/

[39] Nord Pool, "Historical market data," 2021. [Online]. Available: https://www.nordpoolgroup.com/historical-market-data/

[40] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. 35th Int. Conf. Mach. Learn.* PMLR, 2018, pp. 1587–1596.