

Transfer Deep Reinforcement Learning-based Large-scale V2G Continuous Charging Coordination with Renewable Energy Sources

Yubao Zhang, Xin Chen* and Yuchen Zhang

Abstract—Due to the increasing popularity of electric vehicles (EVs) and the technological advancement of EV electronics, the vehicle-to-grid (V2G) technique and large-scale scheduling algorithms have been developed to achieve a high level of renewable energy and power grid stability. This paper proposes a deep reinforcement learning (DRL) method for the continuous charging/discharging coordination strategy in aggregating large-scale EVs in V2G mode with renewable energy sources (RES). The DRL coordination strategy can efficiently optimize the electric vehicle aggregator's (EVA's) real-time charging/discharging power with the state of charge (SOC) constraints of the EVA and the individual EV. Compared with uncontrolled charging, the load variance is reduced by 97.37% and the charging cost by 76.56%. The DRL coordination strategy further demonstrates outstanding transfer learning ability to microgrids with RES and large-scale EVA, as well as the complicated weekly scheduling. The DRL coordination strategy demonstrates flexible, adaptable, and scalable performance for the large-scale V2G under realistic operating conditions.

Index Terms—Deep reinforcement learning, proximal policy optimization, vehicle to grid, scheduling strategy, renewable energy sources.

I. INTRODUCTION

A. Background and Motivation

RECENT advances in battery storage technologies, that lowered the battery prices, together with unprecedented awareness towards CO_2 emissions, created momentum for electrification in the transportation sector [1]. Against this background, vehicle-to-grid (V2G) technology can promote the growth and development of electric transportation with a low carbon footprint and cost saving for consumers [2]. However, significant concerns and challenges for power systems operations are posed due to uncertainties in electric vehicle (EV) users' charging behavior.

This work was supported in part by the National Natural Science Foundation of China (Grant No.21773182 (B030103)) and the HPC Platform, Xi'an Jiaotong University. (Corresponding author: Xin Chen, e-mail: xin.chen.nj@xjtu.edu.cn)

Yubao Zhang is with Center of Nanomaterials for Renewable Energy, State Key Laboratory of Electrical Insulation and Power Equipment, School of Electrical Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China (e-mail: YubaoZhang@stu.xjtu.edu.cn).

Xin Chen is with Center of Nanomaterials for Renewable Energy, State Key Laboratory of Electrical Insulation and Power Equipment, School of Electrical Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China (e-mail: xin.chen.nj@xjtu.edu.cn).

Yuchen Zhang is with School of Electrical Information, Xi'an Jiaotong University, Xi'an, Shaanxi, China (e-mail: zyc20000201@stu.xjtu.edu.cn).

a) *EV/V2G Coordination Strategy*: EVs can effectively promote the efficient utilization of energy and decrease the power grid load fluctuation, which provides a broad prospect for the efficient, flexible and large-scale application of distributed renewable energy sources (RES). The orderly scheduling strategy of EVs/V2G has attracted the attention of researchers. Dabbaghjamanesh et al. [3] proposed a newly stochastic framework based on unscented transform to model uncertainties in EVs, e.g., the charging strategy and number of EVs to be charged. To model the competitive market for EV charging service providers, a versatile game-theoretic model was proposed by which each EV charging service provider aims to maximize individual profit through optimizing locations, numbers, the price of the service [4]. As for the charging scheduling strategy, in [5], a coordinated sectional droop charging control strategy was proposed for an electric vehicle aggregator (EVA) that participates in the frequency regulation of the microgrids with high penetration RES. In [6], a deep learning technique involving a multi-relation graph convolutional network was created to forecast the EV charging demand. In [7], an optimal dispatching strategy for V2G aggregator was proposed to satisfy the driving demand of EV owners and maximize the economic benefits of aggregator simultaneously.

In order to investigate the impact of EVs' charging/discharging behavior and demand side response resources on the economic operation of photovoltaic (PV) grid-connected microgrid system, a multi-objective model of microgrid economic scheduling was proposed [8]. To reduce EVA operation costs and maximize the travel utility for EV users participating in this service, Jin et al. developed an EVA optimization schedule model that combined a day-ahead optimization schedule and real-time optimization schedule [9]. The negative impact of large-scale PV and EV integration on distribution network is mutually reduced via V2G technology [10].

Swarm intelligent algorithms are widely used in EV scheduling problems, such as particle swarm optimization (PSO) [11], genetic algorithm [12] and so on. Because swarm intelligent algorithms are less dependent on mathematical model compared to classical analytical methods of optimization [13]. However, most swarm intelligence algorithms fall into local optimum easily, and convergence speed is very slow. Besides, when the optimization objective changes, the algorithm needs to be retrained and calibrated for new parameters.

Reinforcement learning (RL) introduces ambient intelli-

gence into the systems by providing a class of solution methods to the closed loop problem of processing the data to generate control decisions to react [14]. Specifically, the agents interact with the environment to learn optimal policies that map status or states to actions [15].

b) Deep Reinforcement Learning: Deep reinforcement learning (DRL) is characterized by a combination of RL and deep learning algorithms. Deep learning analyzes environmental information and extracts features from it; RL analyzes new environmental features based on these features, and have the ability to select actions in the new environment to achieve target rewards.

At present, many DRL algorithms have been proposed and applied to power system research, including the use of deep Q network (DQN) [16] and double DQN (to avoiding the problem of overestimation of the Q value in the DQN algorithm) [17] to optimize the voltage control of the distribution network; The dueling deep Q network [18] was used to solve the demand-side response problem [19], and it improved the DQN algorithm network structure [20], which can solve the problem of overestimation of the DQN value function, thereby enhancing the generalization ability of the model and improving the stability of the model. DQN can learn successful policies directly from sensory inputs using end-to-end DRL [21]. However, DQN can only handle discrete and low-dimensional action spaces [22]. Trust Region Policy Optimization (TRPO) is effective and robust for optimizing large nonlinear policies such as neural networks [23]. Despite its approximations that deviate from the theory, TRPO tends to give monotonic improvement, with little tuning of hyperparameters. Proximal policy optimization (PPO), has some of the benefits of TRPO, but is much simpler to implement, more general, and have better sample complexity (empirically). Besides overall strikes a favorable balance between sample complexity, simplicity, and wall-time [24]. Based on the fluctuations in the output of RES, the DRL optimization study of EV charging control was carried out [25], but the charging and discharging characteristics of the EV itself were not discussed in depth. In addition, for an individual EV, a DRL real-time scheduling method that considers the randomness of EV user behavior and the uncertainty of real-time electricity prices was proposed [26]. Yan et al. aimed to crack the individual EV charging scheduling problem based on DRL considering the driver's anxiety and the charging cost [27]. To address this challenge on the online scheduling of shared autonomous EV fleets, Qian et al. proposed a novel framework named the shadow-price DRL [28]. To dynamically adjust the EV price in a timely fashion to unlock the flexibility of EV customers, high-dimensional continuous charging cases were studied in [29], [30]. The DRL algorithms use reward and penalty functions to characterize user travel needs, and don't rely on physical models. It is noticed that traditional swarm intelligent algorithms still prevail in large-scale V2G charging/discharging coordination problems. However, the optimization process of these algorithms are costly, which can not meet the needs of large-scale V2G real-time scheduling.

B. Scope and Contributions

This article proposes a novel DRL coordination strategy for peak cutting and valley filling to ensure the safe and stable operation of the power grid, accounting for the large-scale V2G continuous charging/discharging power. In order to schedule large-scale V2G efficiently, the DRL coordination strategy should take into consideration the EVA state of charge (SOC) constraints and EVA power allocation to individual EV. Then, the PPO algorithm is adopted to coordinate the large-scale V2G continuous charging/discharging problems. In contrast with traditional swarm intelligent algorithms and previous DRL algorithms, the proposed strategy optimized by the PPO algorithm sets up the difficulty in multi-dimensional state and continuous action spaces. Experiment results show that the proposed strategy is effective in suppressing grid load fluctuations and can be transferred to RES scenario, large-scale EVA scenario and weekly scenario. More specifically, our contributions of this article are outlined in the following.

First, the DRL formulation is proposed for the large-scale V2G continuous charging/discharging coordination to reduce the load variance in the consideration of the EVA state of charge (SOC) constraints and EVA power allocation to individual EV.

Second, The DRL coordination strategy can theoretically realize real-time precise continuous adjustment of charging/discharging power, because the PPO algorithm can solve the problems in multi-dimensional state and continuous action spaces, in contrast with discrete DRL algorithms such as DQN.

Finally, experiment results demonstrate that the proposed strategy achieves a significantly better peak shaving and valley filling performance in contrast with the PSO and DQN algorithms. Compared with uncontrolled charging, the proposed strategy reduces load variance by 97.37% and charging cost by 76.56%. Besides, the DRL coordination strategy further demonstrates outstanding transfer ability to microgrids with RES and different sizes of EVA, as well as the complicated long-time scale weekly scheduling.

II. COORDINATION PROBLEM FORMULATION

The DRL coordination test system is composed of wind turbines (WTs), PVs, and EVA, and these distributed energy installations are collectively managed by a central control entity with the PPO algorithm. It is reported that EVs will be parked at home for a majority of the time, and the batteries only require to be charged with enough energy before their departure time. Hence, EVs can collectively serve as large-capacity energy storage to mitigate the volatility and intermittency of PV and WT outputs, and facilitate the scheduling system to boost its total profit by moving electricity delivery through hours.

An EVA represents numerous EVs in the wholesale market and coordinates their operation according to the power load conditions and the EV operating characteristics such as traveling patterns and battery's/charger's operating parameters to optimize the corresponding objective functions. By only revealing the aggregate charging/discharging demand, the operators do not need to disclose all of their customers'

charging/discharging information to municipal operators and therefore protect the customer's privacy. In addition, with only the aggregate charging/discharging requirement information of EVA, the municipal operator can optimize the charging/discharging power trajectory for the EVA with considering the detailed charging/discharging needs of individual EV customers.

An EVA agent is a technical and commercial bridge between a system operator and EVs. Optimal participation of an EVA in day-ahead energy and reserve markets demands a DRL coordination strategy. From the system operator perspective, the aggregator is seen as a large source of generation or load, which provides ancillary services such as spinning and regulating reserve. With the DRL coordinated strategy at the EVA level, power grid can reduce the power load fluctuation.

The real-time SOC of the EVA is critical to the measurement of the real-time state of EVA and the EV charging/discharging management. The EVA SOC is enforced to meet the SOC travel demand of users. The EVA SOC at the k -th timeslot, SOC^k , is defined as,

$$SOC^k = \frac{P_{EVA}^k \Delta t + \sum_{n=1}^N SOC_n^k L_n^k Q_n}{\sum_{n=1}^N L_n^k Q_n}, \quad (1)$$

where

$$L_n^k = \begin{cases} 1, & \text{arrival} \\ 0, & \text{departure} \end{cases}. \quad (2)$$

SOC_n^k is the SOC of the n -th EV at the k -th timeslot, P_{EVA}^k is the charging/discharging power of EVA at the k -th timeslot, Q_n is the battery capacity of the n -th EV. Δt is the length of the timeslot, which is one hour. L_n^k is the scheduling state function of the n -th EV at the timeslot k . When $L_n^k = 1$, it means that the EV arrives home ready for the charging/discharging scheduling at the timeslot k . When $L_n^k = 0$, it means that the EV leaves home and can't be scheduled at the timeslot k .

The DRL coordination strategy optimizes the battery states of the examined EVA as an intermediary entity between the EVs and the power grid. The problem is formulated as follows:

$$\max_{P_{EVA}^k} \frac{\alpha}{\sigma^2} + \beta (P_{\max} - P_{\min}) + \chi |SOC_b^k - SOC^k|, \quad (3)$$

subject to:

$$SOC_{\min}^k \leq SOC^k \leq SOC_{\max}^k, \quad (4)$$

$$P_{EVA,dis}^{\max} \leq P_{EVA}^k \leq P_{EVA,ch}^{\max}, \quad (5)$$

$$P_{grid}^{\min} \leq P^k \leq P_{grid}^{\max}, \quad (6)$$

where,

$$P_{\max} = \max(P^{k-23}, \dots, P^k), \quad (7)$$

$$P_{\min} = \min(P^{k-23}, \dots, P^k), \quad (8)$$

$$P^k = P_{base}^k + P_{EVA}^k, \quad (9)$$

$$\sigma^2 = \frac{1}{K} \sum_{k=1}^K (P^k - \bar{P})^2, \quad (10)$$

$$\bar{P} = \frac{1}{K} \sum_{k=1}^K (P^k), \quad (11)$$

$$C = \sum_{k=1}^K \sum_{n=1}^N c^k p_n^k L_n^k \Delta t. \quad (12)$$

The objective function (3) maximizes the overall profit of the EVA and the power grid, which includes the following components: 1) the power grid load variance σ^2 ; 2) its difference between peak load P_{\max} and valley load P_{\min} ; and 3) the EVA SOC. And the solution variable is P_{EVA}^k . α is the reward coefficient of load variance, which is positive; β is the penalty coefficient of peak-valley difference of power grid, which is negative; χ is a positive constraint coefficient of EVA SOC, which ensures that the SOC does not exceed the boundary value; SOC_b^k is the SOC boundary value of EVA at the timeslot k , which includes SOC_{\min}^k and SOC_{\max}^k .

The EVA constraints are shown as (4)-(6). The SOC constraint is described in (4), where SOC_{\min}^k and SOC_{\max}^k are the minimum and maximum SOC of the EVA at the timeslot k respectively. The power of EVA constraint is described in (5), where $P_{EVA,dis}^{\max}$ and $P_{EVA,ch}^{\max}$ are the maximum discharging power and maximum charging power of the EVA respectively. The tie-line power constraint is described in (6), where P^k is the power load at the timeslot k calculated as (9), P_{grid}^{\min} and P_{grid}^{\max} are the minimum and maximum tie-line power respectively.

P_{\max} and P_{\min} are the maximum and minimum value of power load respectively after scheduling, the calculation is shown in (7) and (8). The load variance σ^2 is an indicator reflecting the stability of power grid load. The smaller its value is, the more stable the grid load is. The mathematical model of the load variance is shown as (10), K is the number of timeslots. \bar{P} is the average of the total power load during the day, which is shown as (11), P_{base}^k is the baseload of at the timeslot k . C is the EV's charging/discharging cost or profit, the positive value is cost and the negative value is the profit. The mathematical model of the cost C is shown as (12). p_n^k is the charging/discharging power of the n -th EV at the timeslot k . c^k is the time of use (TOU) tariff at the timeslot k .

Based on the EVA charging/discharging scheduling, individual EV SOC constraints need to be satisfied. The consensus algorithm has the advantages of convenience, speed and practicality in solving multi-agent co-optimization problems, especially power dynamic allocation problems [31]. Therefore, to obtain the individual EV power, an EVA power allocation algorithm based on SOC buffer consistency is proposed. The SOC buffer factor κ can be defined as:

$$\kappa = \begin{cases} \frac{(SOC_n^k - SOC_n^{k-1})}{SOC_{n,max}^k - SOC_n^{k-1}} & P_{EVA}^k > 0 \\ \frac{(SOC_n^{k-1} - SOC_n^k)}{SOC_n^{k-1} - SOC_{n,min}^k} & P_{EVA}^k < 0 \end{cases}, \quad (13)$$

$$p_n^{k*} = (SOC_n^k - SOC_n^{k-1}) * Q_n / \Delta t, \quad (14)$$

$$P_{EVA}^k = \sum_{n=1}^N p_n^{k*} L_n^k, \quad (15)$$

subject to

$$SOC_{n,min}^k \leq SOC_n^k \leq SOC_{n,max}^k. \quad (16)$$

Where $SOC_{n,min}^k$ and $SOC_{n,max}^k$ are the minimum and maximum SOC for the n -th EV at the timeslot k respectively. In particular, when $SOC_{n,max}^k = SOC_n^{k-1}$ and $P_{EVA}^k > 0$, $p_n^{k*} = 0$; when $SOC_{n,min}^k = SOC_n^{k-1}$ and $P_{EVA}^k < 0$, $p_n^{k*} = 0$.

After calculating p_n^{k*} , a safety check correction is needed to obtain the true charging/discharging power p_n^k . During scheduling, the upper and lower limits of the output power of each EV are used to correct the EV SOC, because the maximum charging and discharging power should not be exceeded. An individual EV should meet the power constraint, the charging/discharging power p_n^k is corrected as below.

$$p_n^k = \begin{cases} p_n^{k*} & p_{dis}^{max} \leq p_n^{k*} \leq p_{ch}^{max} \\ p_{ch}^{max} & P_{EVA}^k > 0, p_n^{k*} > p_{ch}^{max} \\ p_{dis}^{max} & P_{EVA}^k < 0, p_n^{k*} < p_{dis}^{max} \end{cases}. \quad (17)$$

Where p_{ch}^{max} and p_{dis}^{max} are the maximum charging and discharging power of the EV, respectively. It's worth mentioning that the SOC buffer factor κ during the allocation of each timeslot, and its value is jointly determined by each EV SOC, the EVA power and the EVA SOC. To allocate the EVA power to the individual EV correctly, all the EVs should have the same SOC buffer factor κ in a timeslot. Therefore the power allocation, P_{EVA}^k , is carried out when all the EVs have the same buffer factor κ . During scheduling, EVs with larger SOCs take a larger share of power when $P_{EVA}^k > 0$; EVs with smaller SOCs take a larger share of power when $P_{EVA}^k < 0$. The SOC of each EV can effectively converge and all reach the desired SOC before their departure time, solving the problem of the redistribution caused by the initial SOC difference.

The V2G technology can improve the penetration of intermittent renewable energy in microgrids. The microgrid includes a variety of RES (such as wind power, PV power, etc.). With RES, the goal in (18) of the DRL coordination strategy is to increase the consumption of renewable energy as much as possible while meeting the SOC driving demands and reducing the load variance.

$$\max_{P_{EVA}^k} \frac{\alpha}{\sigma^2} + \beta (P_{max} - P_{min}) + \chi |SOC_b^k - SOC^k| + \frac{\psi}{f_1}, \quad (18)$$

where

$$f_1 = \frac{1}{K} \sum_{k=1}^K P_{load}^k, \quad (19)$$

$$P_{load}^k = -P_{pv}^k - P_{wt}^k + P_{EVA}^k, \quad (20)$$

$$P^k = P_{base}^k - P_{pv}^k - P_{wt}^k + P_{EVA}^k, \quad (21)$$

$$P_{pv}^k = P_N \frac{I_r^k}{I_N} [1 + \alpha_T (T^k - T_N)], \quad (22)$$

$$P_{wt}^k = \begin{cases} 0 & 0 \leq v^k \leq v_{ci} \text{ or } v^k \geq v_{co} \\ \frac{v^k - v_{ci}}{v_r - v_{ci}} P_{w-rate} & v_{ci} < v^k \leq v_r \\ P_{w-rate} & v_r < v^k < v_{co} \end{cases}. \quad (23)$$

where ψ is the reward coefficient of the mean net load of the micro-grid f_1 , which is negative and ensures the absorption of maximum renewable energy generation. The goal with RES is minimizing the f_1 as shown in (19). In (20), P_{load}^k is net load of microgrid at the timeslot k , which is negative. Because the renewable energy generation is negative, the smaller the P_{load}^k , the more renewable energy is consumed.

P_{pv}^k is PV power at the timeslot k , and the PV output model is as (22). Where P_N is the PV output in standard state, I_r^k is the actual solar radiation intensity at time k , I_N is the solar radiation intensity under ideal conditions, α_T is the power temperature coefficient of PV panels, T^k is the actual temperature of PV panels, $T_N = 25^\circ C$.

P_{wt}^k is wind power at the timeslot k . Relationships between the output wind power P_{wt}^k and WT wind speed v^k can be represented by piecewise functions, as (23). Where P_{w-rate} is the rated output power of WT, v_{ci} is the cut-in wind speed, v_r is the rated wind speed, v_{co} is the cut-out wind speed.

III. PROPOSED METHODOLOGY

In the DRL algorithms, the agent interacts with an environment through a sequence of observations, actions and rewards. The goal of the agent is to select actions in a fashion that maximizes cumulative future reward [21]. DRL can be defined as a Markov decision process which includes: 1) a state space \mathcal{S} ; 2) an action space \mathcal{A} ; 3) a transition dynamics distribution with conditional transition probability $p(s_{k+1} | s_k, a_k)$, satisfying the Markov property, i.e., $p(s_{k+1} | s_k, a_k) = p(s_{k+1} | s_1, a_1, \dots, s_k, a_k)$; and 4) a reward $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$.

PPO, as a class of continuous DRL algorithms, is applied to the DRL coordination strategy for the large-scale EVA that can effectively jointly optimize the grid load. The DRL coordination strategy is presented as Fig. 1.

At timeslot k , we observe the system state s_k which includes the information about the SOC of the EVA, the load variance and the 24-hour load values. Based on this information, the agent will pick the charging/discharging action a_k . This action represents the amount of energy that the EVA will be charged or discharged during this timeslot. After executing this action, we can observe the new system state s_{k+1} and choose the new charging/discharging action a_{k+1} for timeslot $k+1$.

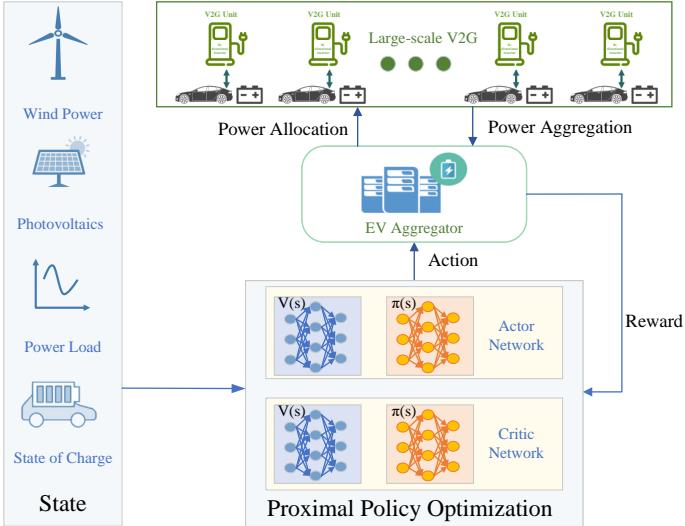


Fig. 1. Workflow of proposed DRL coordination strategy by the PPO scheduling.

A. Proximal Policy Optimization (PPO)

An actor-critic, model-free algorithm was presented based on the deterministic policy gradient that can operate over continuous action spaces [22]. In both actor network and critic network, there is a value network $V(s)$ and a policy network $\pi(s)$. PPO has some benefits of TRPO, but is much simpler to implement. PPO algorithm that uses fixed-length trajectory segments is shown below. Each iteration, each of Y (parallel) actors collect K time steps of data. Then we construct the surrogate loss on these YK time steps of data, and optimize it with mini batch stochastic gradient descent (or usually for better performance [32]), for epochs. The main objective of PPO is the following, which is (approximately) maximized each iteration:

$$L_k^{CVB}(\theta) = \hat{\mathbb{E}}_k [L_k^C(\theta) - c_1 L_k^{VF}(\theta) + c_2 B[\pi_\theta](s_k)]. \quad (24)$$

Where

$$L_k^C(\theta) = \hat{\mathbb{E}}_k [\min(L_k^{CPI}, G_k^{CLIP})], \quad (25)$$

$$L_k^{CPI} = r_k(\theta) \hat{A}_k, \quad (26)$$

$$G_k^{CLIP} = \text{clip}(r_k(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_k, \quad (27)$$

$$L_k^{VF} = (V_\theta(s_k) - V_k^{\text{targ}})^2, \quad (28)$$

$$\hat{A}_k = \delta_k + (\gamma \lambda) \delta_{k+1} + \dots + (\gamma \lambda)^{K-k+1} \delta_{T-1}, \quad (29)$$

$$\delta_k = r_k + \gamma V(s_{k+1}) - V(s_k), \quad (30)$$

$$r_k(\theta) = \frac{\pi_\theta(a_k | s_k)}{\pi_{\theta_{\text{old}}}(a_k | s_k)}, \text{ so } r(\theta_{\text{old}}) = 1. \quad (31)$$

Where c_1, c_2 are coefficients, and B denotes an entropy bonus, and L_k^{VF} is a squared-error loss, epsilon is a hyperparameter, say, $\epsilon = 0.2$. The truncated version of generalized advantage estimation is \hat{A}_k , k specifies the time index in $[0, K]$, within a given length- K trajectory segment, in which γ is the discount factor determining the agent's horizon. $r_k(\theta)$ is the probability ratio, the behavior policy $\pi_\theta(a_k | s_k)$ denotes making an observation s_k and taking an action a_k at each timeslot k , θ is the parameter of actor neural network.

B. DRL Formulation of Examined Problem

In this section, the large-scale V2G continuous charging/discharging coordination problem formulation by the PPO scheduling is presented as Fig. 1. When the EVs park at home, the DRL coordination strategy achieves peak shaving, valley filling and cost reduction by controlling the charging/discharging time and power of EVA. Besides, the proposed strategy is also used to reduce the volatility of renewable energy generation.

We detail the DRL formulation of the examined large-scale V2G continuous charging/discharging coordination problem, the key elements of which are outlined in the following.

a) Agent: The examined EVA constitutes the agent, which gradually learns how to improve its retail charging/discharging decisions by utilizing experiences from its repeated interactions with the environment.

b) Environment: The environment consists of the EVs, power grid and RES (wind power and PV power), with all of which the EVA interacts.

c) State: This initial state space $\mathcal{S}_i = (P^{k-23}, \dots, P^k, SOC^k, \sigma_k^2)$ encapsulates three types of information:

- (1) (P^{k-23}, \dots, P^k) denotes the past 24-hour power grid load values at the timeslot k ;
- (2) SOC^k represents the EVA SOC at the timeslot k ;
- (3) σ_k^2 indicates the load variance at the timeslot k .

In order to prevent gradient explosion of the neural network, the state values are normalized. The state at timeslot k is defined as a vector \mathcal{S} .

$$S(j) = \frac{\mathcal{S}_i(j) - \text{Mean}(\mathcal{S}_i)}{\text{Std}(\mathcal{S}_i)}. \quad (32)$$

Where $S(j)$ represents the j -th state value, which is normally distributed between 0 and 1. $\text{Mean}(\mathcal{S}_i)$ is the average of the set of initial state values, $\text{Std}(\mathcal{S}_i)$ is the standard deviation of the set of initial state values.

d) Action: Given the state \mathcal{S} , a_k is the action of an agent at the timeslot k .

$$a_k = P_{EVA}^k. \quad (33)$$

The action a_k represents the charging/discharging power of the EVA at the timeslot k . Let a_k be positive when the EVA is charging and negative when discharging. Besides, we assume that the V2G equipment provides continuous charging/discharging power.

e) Reward: Reward is the optimization goal for the agent and is the function of the load variance. To avoid sparse rewards during the agent learning process, the functions below highlight the rewards and penalties for every action undertaken by the agent while scheduling EVA charging/discharging.

$$r_g = \frac{\alpha}{\sigma^2} + \beta (P_{\max} - P_{\min}) + \chi |SOC_b^k - SOC^k|, \quad (34)$$

$$r_r = \frac{\alpha}{\sigma^2} + \beta (P_{\max} - P_{\min}) + \chi |SOC_b^k - SOC^k| + \frac{\psi}{f_1}. \quad (35)$$

Where r_g is the reward of the agent in the baseload scenario, r_r is the reward of the agent in the RES scenario.

IV. EXPERIMENT RESULTS

A. Test System and Implementation

The following experiment results examine the EVA scheduling problem in the context of a single day with hourly resolution. The EVA serves 509 EVs for eliminating EV charging load peaks. The total load, the baseload and uncontrolled EV load in the examined day are shown as Figure 2, the total load profile represents the uncontrolled EV load plus the baseload profile. The objective function is set as (3). With RES, the total load of the microgrid results from the balance of EV, baseload, PV and wind power.

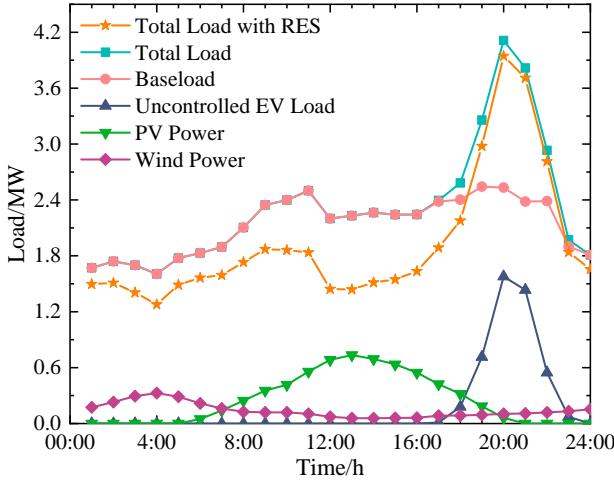


Fig. 2. Daily load profiles with and without RES in a microgrid.

The assumed values of the remaining technical parameters of the EV are $p_{ch}^{max} = 6kW$, $p_{dis}^{max} = -6kW$, $Q_n = 24kWh$ for every EV. The detailed EV load model is shown in [33]. The arrival time is sampled from $N(18, 1^2)$ and is bounded between 15 and 21. For the departure time, its distribution $N(8, 1^2)$ is bounded between 6 and 10. The SOC of n -th EV SOC_n bounded between 0.2 and 0.8 is sampled from $N(0.5, 0.1^2)$. It should be noted that in the mentioned test systems in this article, the EVA is scheduled from timeslot 15 to 10. It is worth noting that the DRL coordination strategy does not rely on any knowledge of the distributions of these random variables. Thus, the proposed strategy can be transferred to different modeling mechanisms.

B. Training Process of PPO

The PPO algorithm requires two networks, namely a critic network for evaluating a state and an actor network mapping a state into a probability distribution over the action space. Both networks are implemented with fully connected layers with the same input size. Both the critic and actor network have one output neuron and employ the rectified linearity units (ReLU) for all hidden layers. The Adam optimizer is employed for learning the neural network weights with a learning rate $lr^a = 10^{-6}$ and $lr^c = 2 * 10^{-6}$ for the actor and critic, respectively. For the critic, we use a discount factor of $\gamma = 0.95$. We use a hyperparameter $\epsilon = 0.2$ in order to avoid an excessively large policy update. We use $us = 10$ as the critic and actor

target network updating step. The output layer of the actor is a sigmoid layer to bound the continuous actions. We train with a mini batch size of $BS = 32$ and for $M = 3 * 10^5$ episodes, with $L = 20$ hours per episode. Finally, we use $\alpha = 10$, $\beta = -5$, $\chi = 10$ for the penalty weighting constant in (34).

During the training process, the reward and the load variance over $3 * 10^5$ episodes are calculated and are illustrated in Fig. 3.

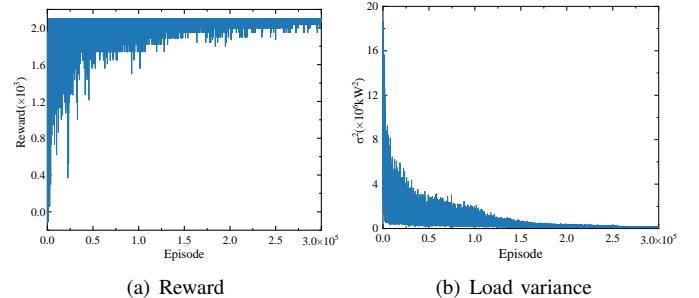


Fig. 3. Convergence process during the training.

As demonstrated in Fig. 3(a), the charging/discharging schedule is randomly selected during the initial learning stages since the EVA is gathering more experiences by randomly exploring different, not necessarily profitable, actions. However, as the learning process progresses and more experiences are collected, the reward keeps increasing, and eventually converges around 2100 with small oscillations. This result demonstrates that the proposed strategy succeeds in learning a policy to maximize reward.

In Fig. 3(b), the load variance converges around $10663.4kW^2$ at episode $3 * 10^5$. Compared with the load variance of uncontrolled charging $419660.9kW^2$, the proposed strategy can reduce the load variance by 97.46%. After the training, the model optimized by the PPO algorithm is saved for transfer learning in different scheduling scenarios.

C. PPO Performance

The performance of the proposed strategy is evaluated with the load variance reflecting the capability of the peak shaving and valley filling during scheduling operation. In a test day, the load variance is calculated as (10). The load variance at each timeslot is calculated using the load profile of the previous 24 hours. In Fig. 4(a), the optimal load profile is the total load profile with EVA scheduled according to the DRL coordination strategy and the blue bar represents the number of scheduled EVs in a day. The optimal load profile is flatter than the total load profile with uncontrolled EVA. The DRL coordination strategy by the PPO scheduling achieves peak shaving and valley filling by allocating the EVA charging demand to the load valley.

At the arrival time of timeslot 15, the coordination scheduling starts to connect EVA to the power grid and the arrival load variance is $90269.2kW^2$. At the departure time of timeslot 10, the departure load variance becomes $11012.5kW^2$. During the scheduling period, the PPO agent gives charging/discharging actions according to the states of the load and the EVA SOC.

EVA discharges at the peak timeslot 16-21 and is charged at the remaining time in the scheduling period. The load variance and charging/discharging power of EVA are shown in Fig. 4(b).

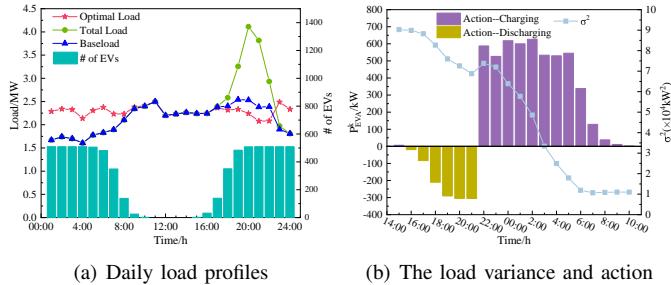


Fig. 4. Daily load profiles, load variance and action in the scheduling period.

V. TRANSFER LEARNING COORDINATION

The transfer learning ability of the DRL coordination strategy is demonstrated and evaluated with the three test systems, the RES test system, the large-scale EVA test system and the weekly test system.

A. Test Systems for Transfer

In the RES test system, the RES are connected to the power grid as demonstrated in Figure 2, and we use $\alpha = 10$, $\beta = -5$, $\chi = 10$, $\psi = -1$ for the penalty weighting constants in (35). The data of RES is shown in [34]. In the large-scale EVA test system, EVAs with 5090, 20360, 35630, and 50900 EVs are considered where the baseload, the PV power and wind power combined is 100 times higher than the load value in the microgrid in Figure 2. In the weekly test system, the scheduling time is set as one week to evaluate the proposed strategy in long-time scale effect. In the weekly test system [35], uniformly distributed noises at $\pm 10\%$ are introduced for the baseload and the RES.

B. EVA Scheduling with RES

The DRL agent gives charging/discharging actions according to the load values when RES are connected in the RES test system. And the load variance decreases from $91688.4kW^2$ to $54127.0kW^2$. We can observe that the proposed strategy reduces the load variance by 88.89% in comparison with the uncontrolled charging. With the PPO scheduling, the optimal load profiles with RES are shown in the figure 5(a). Under the circumstance that the load fluctuation is intensified due to the RES, the proposed strategy still can effectively reduce the peak load and stabilize the power load with RES.

C. Large-scale EVA Scheduling

The load profiles in the four large-scale EVA test systems with different numbers of EVs are demonstrated in Fig. 5(b). It can be seen that as EVA scales up, the effect of the peak shaving and valley filling become more prominent. Besides it, Fig. 5(b) shows that the PPO scheduling can be transferred to

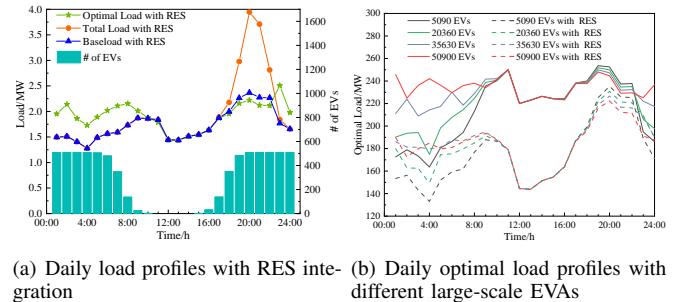


Fig. 5. Transfer learning daily load profiles with RES and large-scale EVA.

TABLE I
COMPARISON OF OPTIMAL LOAD VARIANCES FOR THE BASELOAD AND RES WITH DIFFERENT EVA SIZES.

EVA (# EV)	Baseload(MW^2)	RES(MW^2)
5090	712.35(34.61%) ^a	792.26(33.59%)
20360	321.75 (81.82%)	569.44 (73.34%)
35630	137.27(95.11%)	485.58(85.86%)
50900	70.97(98.31%)	421.16 (91.72%)

^aThe percentage of the load variance reduction with respect to the uncontrolled charging are given in the bracket for the corresponding baseload and RES scenarios.

the RES scenario directly with good performance as well. The optimal load variances for the four different EVAs are listed in Table I. As shown in Table I, the optimal load variance decreases with the increasing EVA size. Clearly, the larger EVA, the better DRL coordination strategy works in terms of the load variance performance. Besides it, the load variance for comparison is the departure time load variance. The scheduling optimization effect in the RES scenario is slightly worse than the baseload scenario, because the DRL coordination strategy doesn't work at noon when PV generates power.

D. Weekly EVA Scheduling

The baseload profile, total load profile and optimal total profile in the weekly test systems with and without RES are shown in Fig. 6. The DRL coordination strategy shows the strong advantages over the uncontrolled charging. The uncontrolled charging of EVs causes the large increasing of the existing peak loads and thus significantly causes the increasing of the load variance. Additionally, daytime load valleys cannot be scheduled, because the valleys are caused by RES and all EVs are not under scheduling. The green line represents the load profile of charging/discharging behavior generated by the PPO scheduling with perfect foresight. In the baseload scenario, the proposed strategy optimized by the PPO algorithm reduces the load variance by 95.04% compared with the uncontrolled charging. Fig. 6 also shows the charging/discharging schedule with RES. The EVs are consistently charged at night during the load valleys and discharged during the load peaks, which promote the consumption and reduce the volatility of RES. In summary, the proposed strategy produces no not-enough-energy cases, no power-limit-exceeded steps, and reduces the load variance by 83.08% compared with the uncontrolled charging with RES.

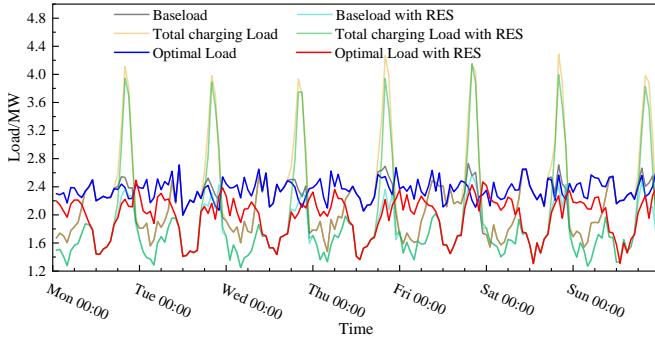


Fig. 6. Charging/Discharging behavior in the weekly test system and optimal load profiles optimized by the PPO algorithm.

E. Transfer Performance Comparison

a) *RES Test*: The PSO algorithm [36], [37] and DQN algorithm [21], [35] are widely used for the EV scheduling. The comparison between the baseload and RES scenarios is given in Table II.

TABLE II
COMPARISON OF OPTIMAL LOAD VARIANCES.

	PPO _T ^b	PPO _R ^c	PSO	DQN	Uncon ^d
BL ^a	11012.5 (97.37%)	10663.4 (97.46%)	30910.0 (92.63%)	33617.4 (91.99%)	419660.9 (0)
RES	54127.0 (88.89%)	53440.4 (89.03%)	71800.0 (85.27%)	64279.5 (86.81%)	487292.5 (0)

^aThe BL stands for Baseload, ^b the PPO_T stands for transfer PPO scheduling, ^c the PPO_R stands for retrained PPO scheduling, ^d the Uncon stands for uncontrolled charging.

The PPO algorithm is directly transferred for the EVA scheduling. For comparison, PPO is retrained as well for the EVA scheduling. In the baseload scenario, the optimal load variances scheduled according to the retrained PPO and directly transfer PPO scheduling are 10663.4kW² and 11012.5kW², respectively. Compared with the load variance of uncontrolled charging 419660.9kW², the the load variance can be reduced by 97.46% and 97.37%, respectively. Besides it, the optimal load variances according to the PSO and DQN scheduling can be reduced by 92.63% and 91.99%, respectively.

In the RES scenario, the retrained PPO and directly transfer PPO scheduling gives the load variances 53440.4kW² and 54127.0kW², respectively. The PSO and DQN scheduling can reduce the load variance by 89.03% and 88.89%, respectively in reference with the uncontrolled charging 487292.5kW². The PSO and DQN scheduling can reduce the load variance by 88.89% and 86.81%, respectively.

For the comparison of the transfer learning performance, the difference of the peak shaving and valley filling effect between the retrained PPO and transfer PPO scheduling is not significant. The retraining process takes several hours and cannot be used for real-time scheduling of large-scale V2G. The above results and analysis can also prove the transfer learning effectiveness of our algorithm. What's more, after the training process, the DRL coordination strategy by the PPO

scheduling can also be transferred to other scenarios. The iterative problem caused by excessive number of EVs is avoided. It takes about 1.26 ms to generate one schedule, which can meet the needs of real-time large-scale V2G scheduling. As the PSO algorithm usually suffers from the curse of dimensionality due to the huge search space [38], it cannot be transferred to the RES scenario, and the PSO scheduling results here are those obtained by retraining.

b) *Large-scale EVA Test*: The transfer ability of the proposed strategy has yielded good results in the PPO scheduling period of large-scale V2G. The optimal load variances according to the DQN scheduling are shown in Table III. As the number of EVs increases, the load variance decreases. Besides, the effect of peak shaving and valley filling is very obvious in both scenarios. As shown in Table I and Table III, the PPO

TABLE III
OPTIMAL LOAD VARIANCES ACCORDING THE DQN SCHEDULING.

EVA(# EVs)	BaseloadMW ²)	RES(MW ²)
5090	794.02(27.11%)	869.92(27.08%)
20360	519.28(70.67%)	814.31(61.88%)
35630	436.36 (84.45%)	695.37(79.75%)
50900	322.49(92.31%)	652.72 (87.17%)

scheduling is much better than the DQN scheduling in terms of peak shaving and valley filling effect with different large-scale EVAs. The superior performance of the PPO scheduling is driven by its ability to capture a continuous action space in contrast with the naive discretization approach adopted by DQN. The charging/discharging power of the EVA is determined by the number of EVs and their SOCs. When the scale of EVA changes in V2G, DQN needs to increase the discrete action dimension. DQN doesn't have good transfer ablity so that the DQN algorithm needs to be retrained and cannot be directly transferred to different scenarios. The PPO scheduling can adjust the charging/discharging power continuously, which makes the scheduling more accurate and efficient. The PPO scheduling performs significantly better in the peak shaving and valley filling effect than the existing DQN and PSO scheduling, and can meet the demand of real-time scheduling of large-scale V2G in microgrids with RES.

c) *Weekly Test*: The performance according to the PPO and DQN scheduling in a typical week are shown in Table IV. Compared with the uncontrolled charging in a typical week, the DQN scheduling can reduce the load variance by 88.90%. The PPO scheduling can reduce the load variance by 95.04% much better than the DQN scheduling. In the microgrid with RES, the DQN scheduling [35] can reduce the load variance by 81.85%, the PPO scheduling can reduce the load variance by 85.96% better than DQN.

VI. DISCUSSION

A. Bounded SOC Constraints

SOC is the key parameter to properly control the EV and to secure the power responses due to changes in operating conditions. The SOC constraints will ensure the DRL coordination strategy satisfy the driving demand of the EV users. The

TABLE IV
COMPARISON OF OPTIMAL LOAD VARIANCES WITHOUT AND WITH RES.

	PPO	DQN	Uncontrolled
Baseload	20805.63 (95.04%)	46551.97 (88.90%)	419264.32 (0)
RES	67073.08 (85.96%)	86752.84 (81.85%)	477868.74 (0)

SOC constraints can reduce the long-term capacity fade rate and achieve higher number of equivalent full cycles or higher amount of cumulative discharge capacity over the battery's useful life.

a) *Baseload*: In Fig. 7(a), the red star represents the EVA SOC SOC^k in the scheduling period. The scheduling period is from timeslot 15 to 10. At departure time, the SOC of EVA is 0.83, which meets the departure SOC needs (Between 0.8 and 0.9). When the EVA scheduling is completed, the individual EV scheduling is required. The power allocation is elaborated in (13)-(17). Therefore, the SOC of the $n - th$ EV all the time should satisfy (16). And the distribution of individual EV SOC in the PPO scheduling period is shown in Fig. 7(a). The EV SOCs in the scheduling time from timeslot 15 to 19 are more dispersed than in the remaining time, because EV SOCs differ a lot when EVs just arrive home for the charging/discharging scheduling. The scheduling is based on the EVA SOC. At the departure time, the SOC of individual EV is 0.83. In the scheduling period, the SOC of individual EV is always between 0.2 and 0.9, satisfying the SOC constraint of the individual EV. The bounded SOC constraints prevent battery overcharge and overdischarge, protect the battery from rapid degradation, and extend the service life of the battery.

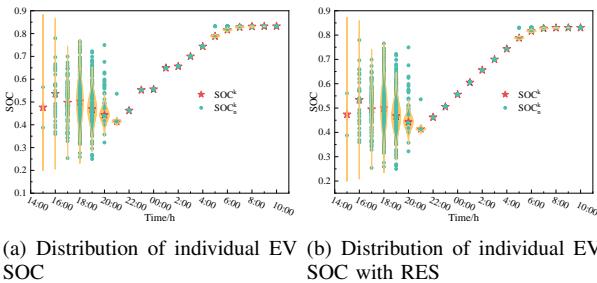


Fig. 7. Distribution of individual EV SOC with and without RES.

b) *Transfer Learning Scheduling with RES*: The DRL coordination strategy has the transfer ability and can be used to schedule large-scale V2G continuous charging/discharging in microgrids with RES. In order to verify that the SOC of EVA in the transfer learning strategy does not cause to cross the boundary, the distribution of each EV SOC with RES is elaborated in Fig. 7(b). At the departure time, the EVA SOC and the individual EV SOC are both 0.87, which also satisfies the SOC constraints.

B. Cost Analysis

In addition to the peak shaving and valley filling in terms of load variance reduction, the DRL coordination strategy

also demonstrates the cost reduction benefit. Users can obtain the economic benefit from the discharging in the peak load shaving, because the TOU tariffs are set differently for the peak and valley hours of the grid load. For example, the TOU tariff is set at 0.8 CNY/kWh for shoulder hours, 0.4 CNY/kWh for off-peak hours, and 1.2 CNY/kWh for peak hours. The DRL coordination strategy reduces the EV charging cost by 76.56% in comparison to the uncontrolled charging with the charging cost of 9074.4 CNY. For the RES integration, the charging cost is reduced by 81.41%.

VII. CONCLUSION

This paper applies the PPO algorithm in the DRL domain to large-scale V2G continuous charging/discharging coordination in microgrids with RES. The constrained EVA SOC and EVA power allocation are presented to schedule large-scale V2G efficiently. Then, taking the minimum load variance as the main objective, the constraint conditions such as RES and SOC of EVA and individual EV are considered. Finally, the PPO algorithm was adapted to optimize the EVA real-time charging/discharging power in continuous actions. Compared with swarm intelligence algorithms (e.g., PSO) and discrete DRL algorithms (e.g., DQN), the DRL coordination strategy optimized by the PPO is superior. Compared with uncontrolled charging, the DRL coordination strategy can reduce the load variance by 97.37% and the charging cost by 76.56% in the baseload test system, and reduce the load variance by 89.03% and the charging cost by 81.41% with RES. In a typical week, the proposed strategy can reduce the load variance by 95.04% in the baseload scenario and reduce the load variance by 83.08% with RES. Besides, the continuous DRL-based large-scale V2G coordinated charging/discharging strategy optimized by the PPO algorithm is independent of prediction information. Therefore, it can meet real-time optimization and EV SOC driving demands. In addition, large-scale V2G coordination has important reference significance for the orderly scheduling of energy storage systems in microgrids.

Future work aims at incorporating the realistic variability of the state features (namely state of health and state of power) in the learning procedure of the DRL coordination strategy through the employment of data from large EV trials. This will help us test and enhance the generalization performance of the proposed strategy to render it robust against such variability, which constitutes a major current research challenge in the area of electric mobility.

REFERENCES

- [1] M. B. Tookanlou, S. A. P. Kani, and M. Marzband, "An optimal day-ahead scheduling framework for e-mobility ecosystem operation with drivers' preferences," *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 5245–5257, 2021.
- [2] J. Singh and R. Tiwari, "Cost benefit analysis for v2g implementation of electric vehicles in distribution system," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, pp. 5963–5973, 2020.
- [3] M. Dabbaghjananesh, A. Kavousi-Fard, and J. Zhang, "Stochastic modeling and integration of plug-in hybrid electric vehicles in reconfigurable microgrids with deep learning-based forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4394–4403, 2020.
- [4] E. Zavvos, "Planning and analysing competing en-route charging stations for electric vehicles: A game-theoretic approach," Ph.D. dissertation, University of Southampton, 2019.

- [5] X. Zhu, M. Xia, and H.-D. Chiang, "Coordinated sectional droop charging control for ev aggregator enhancing frequency stability of microgrid with high penetration of renewable energy sources," *Applied Energy*, vol. 210, pp. 936–943, 2018.
- [6] C. Li, Z. Dong, G. Chen, B. Zhou, J. Zhang, and X. Yu, "Data-driven planning of electric vehicle charging infrastructure: A case study of sydney, australia," *IEEE Transactions on Smart Grid*, 2021.
- [7] C. Peng, J. Zou, L. Lian, and L. Li, "An optimal dispatching strategy for v2g aggregator participating in supplementary frequency regulation considering ev driving demand and aggregator's benefits," *Applied Energy*, vol. 190, pp. 591–599, 2017.
- [8] H. Hou, M. Xue, Y. Xu, Z. Xiao, X. Deng, T. Xu, P. Liu, and R. Cui, "Multi-objective economic dispatch of a microgrid considering electric vehicle and transferable load," *Applied Energy*, vol. 262, p. 114489, 2020.
- [9] Y. Jin, B. Yu, M. Seo, and S. Han, "Optimal aggregation design for massive v2g participation in energy market," *IEEE Access*, vol. 8, pp. 211 794–211 808, 2020.
- [10] G. R. C. Mouli, M. Kefayati, R. Baldick, and P. Bauer, "Integrated pv charging of ev fleet based on energy prices, v2g, and offer of reserves," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1313–1325, 2017.
- [11] M. O. Badawy and Y. Sozer, "Power flow management of a grid tied pv-battery system for electric vehicles charging," *IEEE Transactions on Industry Applications*, vol. 53, no. 2, pp. 1347–1357, 2016.
- [12] H. Mao, J. Shi, Y. Zhou, and G. Zhang, "The electric vehicle routing problem with time windows and multiple recharging options," *IEEE Access*, vol. 8, pp. 114 864–114 875, 2020.
- [13] H. Li, L. Wang, D. Lin, and X. Zhang, "A nash game model of multi-agent participation in renewable energy consumption and the solving method via transfer reinforcement learning," *Proceedings of the CSEE*, vol. 39, no. 14, pp. 4135–4150, 2019.
- [14] L. Lei, Y. Tan, K. Zheng, S. Liu, K. Zhang, and X. Shen, "Deep reinforcement learning for autonomous internet of things: Model, applications and challenges," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1722–1760, 2020.
- [15] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," 2011.
- [16] Q. Huang, R. Huang, W. Hao, J. Tan, R. Fan, and Z. Huang, "Adaptive power system emergency control using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1171–1182, 2019.
- [17] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2313–2323, 2019.
- [18] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *International conference on machine learning*. PMLR, 2016, pp. 1995–2003.
- [19] B. Wang, Y. Li, W. Ming, and S. Wang, "Deep reinforcement learning method for demand response management of interruptible load," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3146–3155, 2020.
- [20] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.
- [21] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [22] 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," *arXiv preprint arXiv:1509.02971*, 2015.
- [23] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.
- [24] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [25] L. Chen, F. Yang, S. Wu, and Q. Xing, "Electric vehicle charging navigation strategy based on data driven and deep reinforcement learning," in *Proceedings of the 5th International Conference on Control Engineering and Artificial Intelligence*, 2021, pp. 16–23.
- [26] 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.
- [27] 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.
- [28] T. Qian, C. Shao, X. Wang, Q. Zhou, and M. Shahidehpour, "Shadow-price drl: A framework for online scheduling of shared autonomous evs fleets," *IEEE Transactions on Smart Grid*, 2022.
- [29] Z. Zhao and C. K. Lee, "Dynamic pricing for ev charging stations: A deep reinforcement learning approach," *IEEE Transactions on Transportation Electrification*, vol. 8, no. 2, pp. 2456–2468, 2021.
- [30] D. Qiu, Y. Ye, D. Papadaskalopoulos, and G. Strbac, "A deep reinforcement learning method for pricing electric vehicles with discrete charging levels," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, pp. 5901–5912, 2020.
- [31] X. Zhang, T. Yu, B. Yang, and L. Li, "Virtual generation tribe based robust collaborative consensus algorithm for dynamic generation command dispatch optimization of smart grid," *Energy*, vol. 101, pp. 34–51, 2016.
- [32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [33] Y. Zhang, H. Hou, J. Huang, Q. Zhang, A. Tang, and S. Zhu, "An optimal subsidy scheduling strategy for electric vehicles in multi-energy systems," *Energy Reports*, vol. 7, pp. 44–49, 2021.
- [34] H. Hou, M. Xue, Y. Xu, J. Tang, G. Zhu, P. Liu, and T. Xu, "Multiobjective joint economic dispatching of a microgrid with multiple distributed generation," *Energies*, vol. 11, no. 12, p. 3264, 2018.
- [35] F. Tuchnitz, N. Eboll, J. Schlund, and M. Pruckner, "Development and evaluation of a smart charging strategy for an electric vehicle fleet based on reinforcement learning," *Applied Energy*, vol. 285, p. 116382, 2021.
- [36] S. M. Arif, A. Hussain, T. T. Lie, S. M. Ahsan, and H. A. Khan, "Analytical hybrid particle swarm optimization algorithm for optimal siting and sizing of distributed generation in smart grid," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1221–1230, 2020.
- [37] M. A. Gormez, M. E. Haque, and Y. Sozer, "Cost optimization of an opportunity charging bus network," *IEEE Transactions on Industry Applications*, vol. 57, no. 3, pp. 2850–2858, 2021.
- [38] R. Jafari, M. M. Javidi, and M. Kuchaki Rafsanjani, "Using deep reinforcement learning approach for solving the multiple sequence alignment problem," *SN Applied Sciences*, vol. 1, no. 6, pp. 1–12, 2019.