EV Charging Strategy Considering Transformer Lifetime via Evolutionary Curriculum Learning-Based Multiagent Deep Reinforcement Learning

Li, Sichen; Hu, Weihao; Cao, Di; Zhang, Zhenyuan; Huang, Qi; Chen, Zhe; Blaabjerg, Frede

Published in:
I E E E Transactions on Smart Grid

DOI (link to publication from Publisher):
10.1109/TSG.2022.3167021

Publication date:
2022

Document Version
Accepted author manuscript, peer reviewed version

EV Charging Strategy Considering Transformer Lifetime Via Evolutionary Curriculum Learning-based Multi-agent Deep Reinforcement Learning

Sichen Li, Weihao Hu, *Senior Member, IEEE*, Di Cao, *Member, IEEE*, Zhenyuan Zhang, *Senior Member, IEEE*, Qi Huang, *Fellow, IEEE*, Zhe Chen, *Fellow, IEEE*, Frede Blaabjerg, *Fellow, IEEE*

***Abstract*—An accelerated loss of life (LOL) of distribution transformers has been observed in recent years owing to the increasing penetration of electric vehicles (EVs). This paper proposes an evolutionary curriculum learning (ECL)-based multi-agent deep reinforcement learning (MADRL) approach for optimizing transformer LOL while considering various charging demands of different EV owners. Specifically, the problem of charging multiple EVs is cast as a Markov game. It is solved by the proposed MADRL algorithm by modeling each EV controller as an agent with a specific objective. During the centralized training stage, a novel centralized ECL mechanism is adopted to enhance the coordination of multiple EVs. It enables the proposed approach to address the management of large-scale EV charging. When the training procedure is completed, the proposed approach is deployed in a decentralized manner. Herein, all the agents make decisions based solely on local information. The decentralized manner of execution helps preserve the privacy of EV owners, reduce the related communication cost, and avoid single-point failure. Comparative tests utilizing real-world data demonstrate that the proposed approach can achieve coordinated charging of a large number of EVs while satisfying the various charging demands of different EV owners.**

***Index Terms*—Evolutionary curriculum-based multi-agent deep reinforcement learning; transformer loss of life; management of EV charging.**

NOMENCLATURE

Abbreviations:

EV	Electric vehicle
MADRL	Multi-agent deep reinforcement learning
LOL	Loss of life
NLL	Normal insulation life of transformer
ECL	Evolutionary curriculum learning
MG	Markov games
MDP	Markov decision process
SOC	State-of-charge
HTS	High time sensitivity
MTS	Mid time sensitivity
MATD3	Multi-agent twin delayed deep deterministic policy gradient
PSN	Parameter space noise
ReLu	Rectified linear unit

Parameters:

Δt	Time interval
NLL	Normal insulation life
τ_{TO}	Oil time constant
τ_w	Winding time constant
$\Delta\theta_{\mathrm{TO,R}}$	Top-oil rise over ambient temperature at rated load
$\Delta\theta_{\mathrm{HS,R}}$	Winding hottest spot conductor rise over top-oil temperature at rated load
R	Ratio of rated load loss to no-load loss
K_{U} / K_i	Ratio of the ultimate/initial load L to the rated load
n / m	Empirically derived exponent used to calculate the variation of $\Delta\theta_{\mathrm{TO}}$ / $\Delta\theta_{\mathrm{HS}}$ with changes in load
t_{dep}^{j}	Departure time of j-th EV
P_{dsg}^{j} / P_{chg}^{j}	Maximum power of discharging and charging of j-th EV
$E_{\max}^{j}$	Battery capacity of j-th EV
$\mathrm{W}_{\mathrm{LOL}}$	Economic value of the transformer
W_{CP}	Measure in \$ to map the CP_t^j into money
Υ^j	Charging preference of the j-th EV owner
w	Weighting factor
c	Bound of action noise
τ	Soft replacement

Variables:

F_{AA}	Aging acceleration factor
F_{EQA}	Equivalent aging factor
θ_{HS}	Winding hottest-spot temperature
ε	Duration of load on the transformer
θ_{A}	Ambient temperature
$\Delta\theta_{\mathrm{TO}}$	Top-oil rise over ambient temperature
$\Delta\theta_{\mathrm{HS}}$	Winding hottest-spot rise over top-oil temperature
$\Delta\theta_{\mathrm{TO,U}}$ / $\Delta\theta_{\mathrm{TO,i}}$	Ultimate/initial top-oil rise over ambient temperature
$\Delta\theta_{\mathrm{HS,U}}$ / $\Delta\theta_{\mathrm{HS,i}}$	Ultimate/initial winding hottest-spot rise over top-oil temperature
L_t^{tot} / L_t^{bas} / L_t^{EV}	Total/basic/EV load

This work was supported by the National Key Research and Development Program of China under Grant 2018YFE0127600. (*Corresponding author: Di Cao*.)

Sichen Li, Weihao Hu, Di Cao and Zhenyuan Zhang are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: sichenli@std.uestc.edu.cn; whu@uestc.edu.cn; caodi@std.uestc.edu.cn; zhangzhenyuan@uestc.edu.cn).

Qi Huang is with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China. He is also with the College of Energy, Chengdu University of Technology, Chengdu, China. (e-mail: hwong@uestc.edu.cn)

Zhe Chen and Frede Blaabjerg are with the Department of Energy Technology, Aalborg University, DK-9220 Aalborg, Denmark (e-mail: zch@et.aau.dk; fbl@et.aau.dk).

$CA_t^j / TS_t^j / CP_t^j / C_t^j$	Charging anxiety/time sensitivity/charging preference/charging cost of *j*-th EV at time *t*
t_{arr}^j	Arrival time of *j*-th EV
SOC_t^j	State-of-charge of *j*-th EV at time *t*
EP_t	Electricity price at time *t*
P_t^j	Power of *j*-th EV at time *t*
$\mathcal{O}_t^j$	Objective function of the *j*-th EV
$\mathfrak{S}$	Randomness from the environment.
$s_t^j / a_t^j / r_t^j$	State/action/reward of j-th agent at time *t*
X_t/A_t	State/action set at time *t*
ϕ	Parameters of critic network
θ	Parameters of actor network
ς	Noise of action
ζ	Variance of normal distribution

I. Introduction

Electric vehicles (EVs) are regarded as environmentally-friendly means of transportation because these can reduce air pollution and fossil fuel consumption. However, unregulated charging creates challenges for the operation of power grid [1], [2]. EVs are typically charged at home, whereby the major impact of EV charging would be on the distribution transformers [3]. Hence, the load of unregulated EV charging would mainly result in the overload of these transformers, and such an effect would intensify further as the penetration of EVs increases [4]. Specifically, the overload may cause an increase in the transformer hottest-spot temperature, resulting in material aging and accelerating the loss of transformer life [5], [6]. An approach to mitigate the negative impact is the reinforcement of transformer assets. However, it is ineffective to adapt to the development of EVs by enhancing transformer assets because the speed of transformer renewal is slower than the present increase in the popularity of EVs [3]. To mitigate the negative impact of EV charging on transformers, one feasible approach is to manage EV charging during idle time periods (i.e., the period when EV owners connect their EVs to the power grid). Incentivizing such demand-side management plans would help mitigate the impact on transformer life and reliability as well as prevent failures [4]. Accordingly, it is necessary to develop an EV charging strategy considering the transformer loss of life (LOL) and demand response arbitrage benefit.

Several studies have adopted centralization-based approaches for the management. In [3], a centralized approach was proposed to simultaneously optimize the transformer LOL and the satisfaction of EV owners. Ref. [4] first collected information of all the EVs and then, presented a fuzzy logic algorithm to schedule EV charging to optimize the transformer LOL. Residential distribution transformers need to supply power to a considerable number of EVs owning to the characteristics of the geographical clusters of EV owners [7] and their preference for charging at home [3]. In [8], it was reported that a transformer powers 30 EVs and that this number is likely to increase as EVs gain popularity [9]. The centralized approaches utilize a central controller to collect information from all the dispatch units and determine the optimal solution based on the global information. However, given the rapid growth of the EV market share, the centralized optimization approach is likely to encounter many limitations while addressing the management of large-scale EV charging. That is, these are vulnerable to single-point failures [10], may not be computationally efficient when confronted with large and complex EV-charging management [11], incur communication cost and their exposure to large-scale controls is more likely to weaken the privacy of EV owners than small-scale controls [12], etc.

The distribution-based approaches divide the entire optimization problem into several sub-problems that are solved in a distributed manner, potentially overcome the drawbacks of the centralized approach to a certain extent. Ref. [13] developed an algorithm to reduce the probability of transformer failure with minimum communication cost. However, owing to the absence of a coordinated mechanism among EVs, each EV is charged to maximize its own satisfaction. This approach cannot ensure that transformers would work in the desired state without human intervention in EV charging. Considering this, [8] proposed a reinforcement learning (RL)-based coordinated charging approach. It is aimed at preventing transformer overloading through coordination among EVs. Information is shared among EVs to enhance the coordination. However, such an operation would result in an increase in communication cost and breach of privacy. Ref. [14] presented a charging algorithm that manages PEV charging based on estimated transformer temperatures, and considers owner privacy. However, it is still unavoidable to incur communication costs. In addition, the above-mentioned studies omitted the various charging demands of different EV owners. This may not be realistic owing to many factors (commuting behavior, traffic conditions, etc.) that would result in unique charging-demands for each EV owner [15].

Considering these shortcomings, an actor-critic multi-agent deep reinforcement learning (MADRL) is considered in this study to manage the charging of multiple EVs. Herein, charging preference is utilized to characterize different EV owners having various charging demands. In addition, each EV controller is modeled as an intelligent agent to provide charging autonomy for enabling EV owners to select EV charging patterns to satisfy their unique demands. The actor-critic framework enables the critic network with rich additional information to induce the improvement of the actor network. This operation enables the actor network of each agent to achieve a coordinated relationship with others even when it is based only on local information at the execution stage [16]. Coordinated EV-charging management can be attained in a completely decentralized manner by leveraging this mechanism. This yields a novel guideline for studies related to EV charging considering privacy protection, minimization of communication costs associated with the deployment of communication devices, and prevention of single-point failure. However, current MADRL approaches may fail to optimize the large-scale coordinated EV-charging problem in a decentralized manner because the RL-based approach optimizes the action through trial and error. In the multi-agent scenario, the action space would increase exponentially as the number of agents increases. A larger action space would hinder agents from identifying good strategies through trial and error, particularly in the coordinated scenario [17]. These factors can result in limitations in the application of MADRL to coordinated EV-charging management. To overcome the limitations on the application of a fully decentralized control mechanism for large-scale EV-charging management, this paper proposes a novel decentralized control framework for minimizing the transformer LOL and the dissatisfaction of

EV owners. It is based on a combination of the MADRL algorithm and the evolutionary curriculum learning (ECL) mechanism. The following are the main contributions of this study:

1) The proposed decentralized control framework can address scenarios wherein a large number of EVs need to be managed. This is achieved by integrating the ECL mechanism with the centralized training and decentralized execution framework. Specifically, the coordinated charging management of EVs is modeled as a Markov game (MG). The MG is solved using the proposed MADRL approach by modeling each EV controller as an intelligent agent. All the agents are trained in a centralized manner with curriculum learning and evolutionary selection process to develop a coordinated strategy. This is different from previous MADRL approaches [10], [16], and [18] that encounter substantial challenges in learning good strategies when the agent population is large.
2) The problem of EV-charging management is solved by the MADRL algorithm, which features centralized training and decentralized execution. The centralized training process and global critic enables the proposed approach to learn a coordinated control strategy. When the training process is completed, only local information would be required by each agent to make decisions. This is different from conventional distributed approaches, which achieve coordinated charging management through communication. This decentralized approach can help protect the privacy of EV owners, minimize communication cost, and prevent single-point failure.
3) The proposed approach can satisfy the various demands of different EV owners. This is achieved by modeling each EV controller as an agent and assigning various reward function to different agents. This differentiates the proposed approach from the centralized management approach that aims to optimize a summing objective function. It may be difficult to satisfy various demands of different EV owners by the simple optimization of a summing objective function.

The remainder of this paper is organized as follows. Section II introduces the mathematical formulations of the transformer and EV owners' optimization models. In Section III, the optimization problem is reformulated as an MG, and the proposed approach is described in detail. In Section IV, the simulations and comparative results are discussed in detail to demonstrate the effectiveness of the proposed approach. Finally, Section V concludes the paper.

II. System Model And Problem Formulation

This paper considers a distribution system that includes a distribution transformer and several residential loads. The residential loads can be divided into *1)* the basic load (the loads other than EV loads) and *2)* the loads of all the EVs. This study focuses on the co-optimization of transformer LOL and dissatisfaction of EV owners. There are two components need to be modeled: *1)* transformer LOL model and *2)* EV owners' optimization model.

A. Transformer Model

In this study, the IEEE standard C57.91 [19] is utilized to construct the transformer LOL model. It describes the impact of transformer loading on the corresponding LOL. Herein, transformer LOL is defined as

$$F_{\mathrm{AA}}=\exp\left(\frac{15000}{383}-\frac{15000}{\theta_{\mathrm{HS}}+273}\right),\ F_{\mathrm{EQA}}=\frac{\sum_{k=1}^{M}F_{\mathrm{AA},k}\Delta t}{\sum_{k=1}^{M}\Delta t} \quad (1)$$

$$\mathrm{LOL}=\frac{F_{\mathrm{EQA}}\times\mathcal{E}}{\mathrm{NLL}} \quad (2)$$

where F_{AA} denotes the aging acceleration factor, F_{EQA} is the equivalent aging factor, θ_{HS} is the winding hottest-spot temperature, M is the total number of time intervals, k denotes an index, $F_{\mathrm{AA},k}$ is the aging acceleration factor for the time interval Δt, $\mathcal{E}$ represents the duration of load L on the transformer, and NLL denotes the normal insulation life of transformer [19]. As the equations show, the transformer LOL defined in Eq. (2) is a function of θ_{HS}. θ_{HS} can be calculated using the following equation:

$$\theta_{\mathrm{HS}}=\theta_{\mathrm{A}}+\Delta\theta_{\mathrm{TO}}+\Delta\theta_{\mathrm{HS}} \quad (3)$$

where θ_{A} represents the ambient temperature, $\Delta\theta_{\mathrm{TO}}$ is the top-oil rise over ambient temperature, and $\Delta\theta_{\mathrm{HS}}$ indicates the winding hottest-spot rise over top-oil temperature. $\Delta\theta_{\mathrm{TO}}$ and $\Delta\theta_{\mathrm{HS}}$ are defined as

$$\Delta\theta_{\mathrm{TO}}=\left(\Delta\theta_{\mathrm{TO,U}}-\Delta\theta_{\mathrm{TO,i}}\right)\times\left(1-\mathrm{e}^{-\mathcal{E}/\tau_{\mathrm{TO}}}\right)+\Delta\theta_{\mathrm{TO,i}} \quad (4)$$

$$\Delta\theta_{\mathrm{HS}}=\left(\Delta\theta_{\mathrm{HS,U}}-\Delta\theta_{\mathrm{HS,i}}\right)\times\left(1-\mathrm{e}^{-\mathcal{E}/\tau_{w}}\right)+\Delta\theta_{\mathrm{HS,i}} \quad (5)$$

where $\Delta\theta_{\mathrm{TO,U}}$ and $\Delta\theta_{\mathrm{TO,i}}$ denote the ultimate and initial top-oil rise over ambient temperature, respectively; $\Delta\theta_{\mathrm{HS,U}}$ and $\Delta\theta_{\mathrm{HS,i}}$ are the ultimate and initial winding hottest-spot rise over top-oil temperature, respectively; and τ_{TO} and τ_{w} represent the oil time and winding time constants, respectively. The two constants are explained in detail in [19]. $\Delta\theta_{\mathrm{TO,U}}$, $\Delta\theta_{\mathrm{TO,i}}$, $\Delta\theta_{\mathrm{HS,U}}$, and $\Delta\theta_{\mathrm{HS,i}}$ are defined by the following equations:

$$\Delta\theta_{\mathrm{TO,U}}=\Delta\theta_{\mathrm{TO,R}}\times\left[\frac{K_{\mathrm{U}}^{2}R+1}{R+1}\right]^{n},\ \Delta\theta_{\mathrm{TO,i}}=\Delta\theta_{\mathrm{TO,R}}\times\left[\frac{K_{i}^{2}R+1}{R+1}\right]^{n} \quad (6)$$

$$\Delta\theta_{\mathrm{HS,U}}=\Delta\theta_{\mathrm{HS,R}}\times K_{\mathrm{U}}^{2m} \quad (7a)$$

$$\Delta\theta_{\mathrm{HS,i}}=\Delta\theta_{\mathrm{HS,R}}\times K_{i}^{2m} \quad (7b)$$

where $\Delta\theta_{\mathrm{TO,R}}$ is the top-oil rise over ambient temperature at rated load, $\Delta\theta_{\mathrm{HS,R}}$ denotes the winding hottest-spot conductor rise over top-oil temperature at rated load, R is the ratio of rated load loss to no-load loss, K_{U} is the ratio of the ultimate load L to the rated load, K_i is the ratio of the initial load L to the rated load, and n / m is an exponent used to calculate the variation in $\Delta\theta_{\mathrm{TO}}$ / $\Delta\theta_{\mathrm{HS}}$ with that in the load. n and m are determined empirically by the type of transformer. A detailed explanation is presented in [19]. Here, K_{U} is influenced by the total load $\mathbf{L}_t^{tot}$ of the transformer [19]. $\mathbf{L}_t^{tot}$ is defined as

$$\mathbf{L}_t^{tot}=\mathbf{L}_t^{bas}+\mathbf{L}_t^{EV} \quad (8)$$

where $\mathbf{L}_t^{bas}$ represents the basic load and $\mathbf{L}_t^{EV}=\sum_{j=1}^{N}P_t^{j}\Delta t$ denotes the load of N EVs. As indicated by [13], $\mathbf{L}_t^{bas}$ needs to be forecasted according to historical data. The forecasted value of $\mathbf{L}_t^{bas}$ is expressed as $\hat{\mathbf{L}}_t^{bas}$.

B. EV Owners' Optimization Model

The EV owners' objective is to minimize their dissatisfaction. This paper considers two types of

dissatisfaction of EV owners, *1) charging preference* and *2) charging cost*.

1) Charging preference: It is used to characterize different charging patterns of EV owners. We consider the combination of charging anxiety and time sensitivity to represent charging preference to mitigate the influence of early stopping of charging caused by uncertain events during the EV charging period that cannot satisfy the energy demands of EV owners.

Charging anxiety indicates EV owner's concern that the EV does not have adequate energy to reach its destination. The physical significance of charging anxiety is similar to that of range anxiety mentioned in [20]. However, charging anxiety would more realistically reflect an individual's psychological state. In a realistic scenario, the main disadvantage of EVs compared with vehicles that use conventional fuels is the longer energy-filling time. During the charging period, the EV owners may experience anxiety regarding whether the battery energy would attain the level (after charging) that is required to satisfy their travel demands. Unlike the range anxiety noted in [20] that only provides the dissatisfaction feedback to the controller at departure time, charging anxiety is associated with the length of the charging time and sends dissatisfaction feedback at each time slot for the duration of charging. The charging anxiety CA of the *j*-th EV at time *t* is defined as

$$\mathrm{CA}_t^j = \frac{E_{\max}^j \left(1 - SOC_t^j\right)}{t_{dep}^j - t} \tag{9}$$

where $E_{\max}^j$ is the battery capacity of the *j*-th EV, t_{dep}^j represents its departure time, and SOC_t^j denotes its state-of-charge (SOC). The charging anxiety values under the time-varying albeit fixed SOC scenario are shown in Fig. 1 to explain CA_t^j better. Specifically, the figure shows the charging period from the arrival time t_{arr} to 1 h prior to the departure time ($t_{dep}-1$). In Fig. 1, $E_{\max}^j$ = 24 kWh, and SOC_t^j = 0.1 during the charging period. It is evident from the figure that because the charging time would decrease as the EV owner approaches t_{dep}^j (which would reduce the probability that the battery would be charged to the desired energy in a limited time), the charging anxiety varies more with the reduction in remaining charging time than it did earlier.

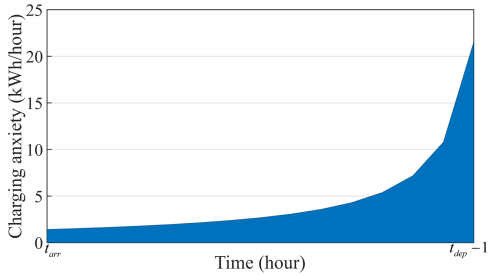


Fig. 1. The charging anxiety varying with charging time.

Inspired by the concept mentioned in [21], time sensitivity is introduced here to reflect the differences among EV owners in terms of the preference to charge their batteries within the charging time. Time sensitivity is the anxiety of having to stop the charging prematurely owing to certain uncertain events during the charging time that would result in an unsatisfied charging demand. Specifically, EV owners with high time sensitivity (HTS) tend to charge at early charging periods. This implies that HTS imparts EV owners with the tendency to access more battery-energy earlier (compared with EV owners displaying mid time sensitivity (MTS)) before they are likely to encounter uncertain events. The influence of time sensitivity for the *j*-th EV at time *t* is defined as

$$\mathrm{TS}_t^j = \begin{cases} B_t^j & \text{for MTS} \\ \log_{Z^j}\left[B_t^j\left(Z^j - 1\right) + 1\right] & \text{for HTS} \end{cases}$$

$$B_t^j = \min\left(\frac{\max\left(t + 1 - t_{arr}^j,\ 0\right)}{t_{dep}^j - t_{arr}^j},\ 1\right) \tag{10}$$

where Z is the time sensitivity ratio for the EV owners with HTS. The variation in TS_t^j with time is shown in Fig. 2. It is evident from the figure that at time *t*, the MTS has the smallest influence among these seven lines. This implies that the EV owners with MTS have lower anxiety, which results in a lower willingness to charge earlier than the EV owners with HTS. Moreover, the larger Z for HTS results in a larger influence. This implies that the larger Z is, the higher the willingness to satisfy the energy demands as early as feasible before approaching the departure time.

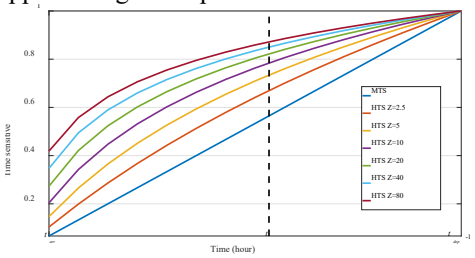


Fig. 2. The time sensitivity varying with charging time.

Combining the charging anxiety CA and time sensitivity TS, the charging preference CP of the *j*-th EV at time *t* can be defined as

$$\mathrm{CP}_t^j = \mathrm{CA}_t^j \times \mathrm{TS}_t^j,\ t \in \left[t_{arr}, t_{dep}-1\right] \tag{11}$$

Fig. 3 shows the variation in CP_t^j with time. Here, the blue and orange areas denote HTS with Z = 80 and MTS, respectively, for an equal CA. As the figure shows, TS provides diversity for CP. This implies that the EV owners with HTS have a higher CP to charge in early charging periods than EV owners with MTS.

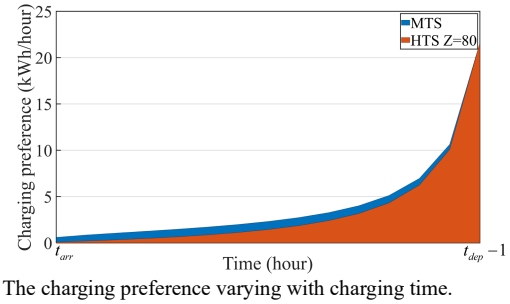


Fig. 3. The charging preference varying with charging time.

2) Charging cost: The utilization of real-time signals such as the time-of-use price can substantially reduce the charging cost. The charging cost can be expressed as

$$\mathrm{C}_t^j = EP_t \cdot P_t^j \cdot \Delta t$$
$$s.t.\ P_{dsg}^j \le P_t^j \le P_{chg}^j \tag{12}$$

where C_t^j denotes the charging cost of the *j*-th EV at time *t*. C_t^j is positive when the EV draws energy from the grid to charge, and negative when the EV earns by discharging the

energy to the grid. EP_t is the electricity price at time *t*. It needs to be forecasted based on historical EP data. The forecasted value of EP_t is denoted by $\widetilde{EP}_t$. P_t^j represents the power of the *j*-th EV at time *t*. P_{dsg}^j and P_{chg}^j are the maximum discharging and charging power, respectively, of the *j*-th EV. For this study, the selling price of electricity is equal to its purchasing price as recommended by [20].

III. Proposed Framework for EV-Charging Management

Real-time EV-charging management can be an optimization problem that minimizes the following equation considering the above-mentioned models. The objective function of the *j*-th EV $\mathcal{O}_t^j$ at time *t* is defined as

$$\mathcal{O}_t^j : \min \frac{\mathrm{W_{LOL}}\left(\mathrm{LOL}_t^{tot} - \mathrm{LOL}_t^{bas}\right)}{N} + \mathrm{W_{CP}}\mathrm{CP}_t^j + \mathrm{C}_t^j \quad (13)$$
$$s.t.\ P_{dsg}^j \le P_t^j \le P_{chg}^j$$

where N indicates the number of EVs. LOL_t^{tot} and LOL_t^{bas} are the LOLs determined by $\mathbf{L}_t^{tot}$ and $\mathbf{L}_t^{bas}$, respectively. $\mathrm{W_{LOL}}$ indicates the economic value of the transformer, and $\mathrm{LOL}_t^{tot} - \mathrm{LOL}_t^{bas}$ denotes the LOL under the load of the N EVs. $\mathrm{W_{CP}}$ is used to measure in USD to map the CP_t^j to currency.

In this section, the optimization problem is formulated as an MG, which is then solved by the proposed approach. It should be noted that MG differs from the Markov decision process (MDP). The MDP can only describe single-agent environments, whereas MG is a game theory [22] extension to MDP-like environments. It mainly studies the problem of coordination and competition in multi-agent systems.

A. Problem Reformulation

The MG is mainly composed of four components:

- *State:* The state of the *j*-th EV $s_t^j \in S^j$ at time *t* is defined as

$$s_t^j = (\Upsilon^j,\ \tilde{\theta}_{\mathrm{HS},t}^{bas},\ \mathbf{L}_{t-1}^{EV},\ \widetilde{EP}_t,\ t,\ SOC_t^j,\ t_{dep}^j,\ SOC_{t_{dep}^j}^j) \quad (14)$$

where Υ^j denotes the charging preference of the *j*-th EV owner. $\tilde{\theta}_{\mathrm{HS},t}^{bas}$ is the winding hottest-spot temperature at time *t* and is influenced by the $\tilde{\mathbf{L}}_t^{bas}$ [19]. Note that $\tilde{\mathbf{L}}_t^{bas}$ is a forecasted value, and $\tilde{\theta}_{\mathrm{HS},t}^{bas}$ is not the true value. The true value is denoted as $\theta_{\mathrm{HS},t}^{bas}$ and is influenced by $\mathbf{L}_t^{bas}$.

- *Action:* The action in the MG represents the charging/discharging power of EVs. The action of the *j*-th EV at time *t* is defined as $a_t^j = P_t^j$. That the action a_t^j is active implies charging, and a negative value denotes discharging.

- *Reward Function:* In this study, we consider the co-optimization of transformer LOL and the dissatisfaction of all the EVs as follows:

$$r_t^j = -\left(w \cdot \mathcal{O}_t^j + (1-w) \cdot \sum_{i=1}^{N} \mathcal{O}_t^i\right) \quad (15)$$

Here, $w \in [0,1]$ denotes the weighting factor that indicates the relative importance between the individual objective $\mathcal{O}_t^j$ and overall objective (the transformer LOL and the dissatisfaction of all the EV owners, i.e., $\sum_{i=1}^{N} \mathcal{O}_t^i$). For example, only individual objective is considered when the weighting factor w is set to one. When w is set to zero, the overall objective reward accounts for the largest proportion in the reward function.

- *Transition Function:* Following the action a_t, the state transforms from s_t^j to s_{t+1}^j at time-step $t+1$ with the transition function $s_{t+1}^j = \mathcal{T}\left(s_t^j, a_t^j, \mathfrak{S}\right)$. Here, $\mathfrak{S}$ indicates the randomness from the environment.

a_t^j mainly alters the deterministic elements in s_t^j, i.e., $SOC_{t+1}^j = SOC_t^j + a_t^j \Delta t / E_{\max}^j$. For the $\widetilde{EP}_t$ in s_t^j, the state transition is influenced by $\mathfrak{S}$. The transition of $\theta_{\mathrm{HS},t}$ is affected by both a_t^j and $\mathfrak{S}$.

B. Proposed Control Model Approach

The proposed control approach is a combination of MADRL and curriculum learning, i.e., the multi-agent twin-delayed deep deterministic policy gradient (MATD3) algorithm and ECL training mechanism.

1) Brief introduction of MATD3: The MATD3 is the multi-agent variant [23] of the TD3 [24] algorithm. It is used to solve the MG by modeling each EV controller as a TD3 agent within the centralized training framework. Considering N agents with parameterized continuous policies, the gradient for the *j*-th agent can be expressed as ($j \in [1, N]$)

$$\nabla_{\theta_j} J(\theta_j) = \mathbb{E}_{\mathrm{X}_t,\ \mathrm{A}_t^{\setminus j} \sim \mathcal{D}}[\nabla_{\theta_j} \mu_{\theta_j}^j(s_t^j) \nabla_{a_t^j} Q_{\phi_{j,1}}^j(\mathrm{X}_t, \mathrm{A}_t)|_{a_t^j = \mu_{\theta_j}^j(s_t^j)}] \quad (16a).$$

Here, $\mathrm{X}_t = [s_t^1, ..., s_t^N]$ is the state set of time *t*, $\mathrm{A}_t^{\setminus j} = [a_t^1, ..., a_t^{j-1}, a_t^{j+1}, ..., a_t^N]$ is the action set of other agents of time *t*, $\mathcal{D}$ denotes the replay buffer (contains experiences of the N agents), $\mu_{\theta_j}^j$ represents the policy network of the *j*-th agent with parameter θ_j, $\mathrm{A}_t = [a_t^1, ..., a_t^N]$ represents the action set of time *t*, and $Q_{\phi_{j,1}}^j(\mathrm{X}_t, \mathrm{A}_t)$ is the centralized action-value function with parameters $\phi_{j,1}$. In TD3, the action-value function pair $Q_{\phi_{j,i=1,2}}^j$ is used to eliminate overestimation. $Q_{\phi_{j,1}}^j(\mathrm{X}_t, \mathrm{A}_t)$ and $Q_{\phi_{j,2}}^j(\mathrm{X}_t, \mathrm{A}_t)$ are updated as

$$Loss = (Q_{\phi_{j,i=1,2}}^j(\mathrm{X}_t, \mathrm{A}_t) - y)^2$$
$$y = r_t^j + \gamma \min\left[Q_{\phi'_{j,1}}^j(\mathrm{X}_{t+1}, \mathrm{A}_{t+1}),\ Q_{\phi'_{j,2}}^j(\mathrm{X}_{t+1}, \mathrm{A}_{t+1})\right] \quad (16b)$$

where $\mathrm{A}_{t+1} = [\mu_{\theta'_1}^1(s_{t+1}^1) + \varsigma^1, ..., \mu_{\theta'_N}^N(s_{t+1}^N) + \varsigma^N]$. ς denotes the noise and can be defined as

$$\varsigma \sim clip(\mathcal{N}(0, \zeta), -c, c) \quad (17)$$

where γ is the reward discount factor, $\mathcal{N}$ represents the normal distribution, ζ is the variance, and c denotes the bound of noise.

$\phi'_{j,i=1,2}$ and θ'_j can be updated as

$$\begin{aligned} \phi'_{j,i=1,2} &\leftarrow \tau\phi_{j,i=1,2} + (1-\tau)\phi'_{j,i=1,2} \\ \theta'_j &\leftarrow \tau\theta_j + (1-\tau)\theta'_j \end{aligned} \quad (18)$$

where soft replacement $\tau \ll 1$.

2) Introduction of ECL mechanism: ECL is an RL-based training mechanism that can be integrated conveniently with any RL-based algorithm without being limited to particular algorithms. Inspired by [25], it can be concluded that "survival of the fittest" provides a thought for the application of curriculum learning in RL, enabling for a better integration of the two. In light of this, ECL consists of two

critical components: curriculum learning and evolutionary calculation. The original core concept of curriculum learning is to start small: learning is to perform the more convenient task first and then, gradually increase the task difficulty throughout the training process [26]. A notable challenge that MADRL presently encounters is that these may fail to optimize the large-scale EV-charging problem because the action space would increase exponentially with the increase in the number of agents. Curriculum learning is adopted in the proposed approach to overcome the challenge. A highly intuitive approach in accordance with the concept of starting small is to 1) decompose large-scale EV-charging into several stages in which the number of EVs per stage increases successively and 2) begin training at the stage with the minimum number of EVs and terminate it at the final stage with the desired number of EVs. This enables the agents to leverage the experiences from the previous stages to adapt progressively to the present stage. A direct approach to increase the number of EVs to the next stage is cloning. Given N trained EVs with the parameter $\boldsymbol{\omega}_i = [\omega^1, \omega^2, ..., \omega^N]$ at the i-th stage, the initial parameters of the next stage with xN EVs are obtained through cloning at the $(i + 1)$-th stage. This can be expressed mathematically as $\boldsymbol{\omega}_{i+1} = [\underbrace{\boldsymbol{\omega}_i, \boldsymbol{\omega}_i, ..., \boldsymbol{\omega}_i}_{x}]$. However, it is challenging to ensure the overall performance by using this approach to introduce new EVs to the present stage by directly cloning existing EVs from the previous stage. This is mainly a result of the fact that successfully trained agents from the previous stage may not adapt effectively to the present stage with increased number of agents in the environment. This occurs because policy parameters applicable to the previous stage need not be the best initialization for present stage owing the increase in the number of EVs. To ensure ideal performance of the final stage with the desired number of EVs, we need to promote agents with better adaptability at each stage of training. Considering this, the evolutionary calculation is integrated with curriculum learning to enhance the capability of the proposed approach for scaling adaption during the curriculum learning-based training process. Therefore, the proposed training mechanism is called evolutionary curriculum learning (ECL). The ECL mechanism deployed in this model has several steps:

(i) Population initialization: Suppose there are I populations each having n agents with Υ roles. That is, n EV controllers manage EV charging for EV owners with Υ types of charging preferences. Note that the I populations are individual and parallel to each other. The parallel trainings are applied on each population, which implies that I parallel trainings are performed simultaneously.

(ii) Population evaluation: Each population is evaluated after the I trainings are completed. The purpose of evaluation is to continuously select and train the agents with good performance and discard those with low performance. As mentioned above, each population has n agents (n EV controllers) with Υ roles (Υ types of charging preferences of EV owners). $\sigma_i^{(r)}$ denotes the set of agents functioning in role r contained in the i-th population. Here, $i \in [1, ..., I]$, and $r \in [1, ..., \Upsilon]$. An evaluation is performed in a new population consisting of all roles in which the populations are random selected to form the new one population $\left[\sigma_{i_1}^{(1)}, \sigma_{i_2}^{(2)}, ..., \sigma_{i_\Upsilon}^{(\Upsilon)}\right]$. Following this rule, we can determine that I populations can constitute I^Υ new populations. The I^Υ new populations run individually, and in parallel to interact with the environment. Meanwhile, the environment would create a large number of scenarios with randomness to evaluate the adaptability of all the populations according to the environmental feedback rewards. After the rewards of I^Υ new populations are collected completely, these would be averaged and sorted among the same role.

(iii) Population selection: We assume that the final evaluation stage has I populations and that there are I rankings in each role. For each role, the agent sets with top-k reward would survive, and the other agent sets (total is I - k) would be discarded.

(iv) Crossover among populations: After the selection of each role is complete, the surviving agent sets are mixed with those who acting the same role as themselves, and then match with those who acting the different role to form a new population [25]. Essentially, given the top-k agent sets playing the r-th role $\left[\sigma_1^{(r)}, ..., \sigma_k^{(r)}\right]$, there are $0.5 \cdot \left(k^2 + k\right)$ mix combinations under the same role $\underbrace{\left[\left(\sigma_1^{(r)}\sigma_1^{(r)}\right), ..., \left(\sigma_k^{(r)}\sigma_k^{(r)}\right)\right]}_{0.5\left(k^2+k\right)}$. After the mix, the match is applied among the different roles. For example, assuming there are Υ=2 in a population, the new population would consist of $\sigma_{i_1}^{(1)}$ and $\sigma_{i_2}^{(2)}$ extracted from the i_1-th and i_2-th populations, respectively. Here, $i_1, i_2 \in [1, ..., k]$. Thus, the Υ roles would have $\left(0.5 \cdot \left(k^2 + k\right)\right)^\Upsilon$ match combinations where Υ=2, and each match combination is the new population. Then, each new population would be mutated individually in the next step.

(v) Population mutation: Mutation is an important step in evolutionary calculation. It maintains the diversity of evolution by transforming the chromosome loci at the parameter level, thereby preventing the premature phenomenon of local optimum. This is similar to the use of the exploratory mechanism in RL. The introduction of the exploratory mechanism in RL can impel policy to explore in a decision space. Furthermore, an effective exploratory mechanism can accelerate convergence while preventing falling into local optimum. Considering this, the exploratory mechanism is adopted in the proposed approach as the key mutated approach. However, the conventional exploratory mechanism generally adds noise directly to the action [27], [28]. This could potentially impair the dependencies between state and action and results in low performance in complex environments [29]. To overcome the ambiguity caused by conventional exploration mechanisms and better match the concept of mutation at the parameter level in evolutionary calculations, the parameter space noise (PSN) [30] is used during the mutation step. Variations in the parameter level can induce a consistent, complex, and state-dependent transformation to policy over multiple time-steps. This ensures that the PSN can provide natural and efficient exploration for effectively directing the improvement of control performance in the parameter space [30], [31].

Finally, steps ***(ii)***–***(v)*** need to be repeated until the number of agents equals the desired number. The detailed steps of the proposed approach are shown in **Algorithm 1**.

Algorithm 1 Training of proposed approach

1: Initialize the x populations. Each population has n agents with Υ roles
2: Execute parallel MATD3 training on the x populations
3: **while** n less than desired number of agents:
4: Evaluate all the populations according to the reward
5: Select the top-k populations to execute the crossover step according to the evaluation results
6: Execute the mix step to combine the agent sets with identical role
7: Execute the match step among the different roles to construct a new population.
8: Execute parallel MATD3 training on new populations
9: **end**
10: Evaluate all the populations according to the reward
11: Return the best population according to the evaluation
12: Train the best population until reward convergence

3) Architectures of MATD3 networks deployed in ECL mechanism:

Critic architecture (include Q_ϕ and $Q_{\phi'}$): The ECL mechanism is applied in the training of MATD3. In this study, the entire training process is decomposed into multiple stages wherein the number of agents in the environment increases with each successive stage. In the training process, the agents first learn to play in a simpler environment with fewer agents and then, progressively adapt to the more complex environment with more agents. They repeat the steps until the desired number is attained in the environment. It should be noted that the fixed neural network cannot process the changing state dimensions. In this study, the training process will have different number of agents in the environment at different stages due to the presence of the ECL training mechanism. Specifically, the critic network of each agent receives information from all the agents within the environment in a centralized manner during training, the input dimensions of the critic network of each agent will change when the number of agents increases. Inspired by [25], the self-attention network [32] is adopted in the proposed approach to solve this challenge. Specifically, the self-attention network is deployed in an action-value function approximated neural network (shortened to action-value network). Its architecture is shown in Fig. 4. Note that each agent has an individual action-value network that has architecture identical to that of the other agents. Specifically, the j-th agent action-value network can be represented as follows (see Fig. 4; $j \in [1, N]$):

$$\mathrm{e}_j = f_j\left(s^j, a^j\right),\ \mathrm{e}_1 = g_j\left(s^1, a^1\right),\ ...,\ \mathrm{e}_N = g_j\left(s^N, a^N\right)$$

where e_j , e_1, and e_N represent the encoders of the $\left(s^j, a^j\right)$, $\left(s^1, a^1\right)$, and $\left(s^N, a^N\right)$ pairs, respectively; f_j denotes the light-blue feed forward neural network of the j-th agent with an input layer, a hidden layer of rectified linear units function (ReLu) neurons, and a ReLu output layer; and g_j represents the light green feed forward neural network, which has an architecture identical to that of f_j .

After all the encoders $[\mathrm{e}_1, \mathrm{e}_2, ..., \mathrm{e}_N]$ are calculated, $\mathfrak{D}_j$ is obtained as follows:

$$[\alpha_{j(1)}, ..., \alpha_{j(j-1)}, \alpha_{j(j+1)}, ..., \alpha_{j(N)}] = \frac{\varphi\left(\mathrm{e}_j, \left[\mathrm{e}_1^{\mathrm{T}}, ..., \mathrm{e}_{j-1}^{\mathrm{T}}, \mathrm{e}_{j+1}^{\mathrm{T}}, ..., \mathrm{e}_N^{\mathrm{T}}\right]\right)}{\sqrt{d_k}} \tag{19}$$

$$\omega_i = \frac{\exp\left(\alpha_{j(i)}\right)}{\sum_{i \neq j} \exp\left(\alpha_{j(i)}\right)}, i \in [1, ...j-1, j+1, ..., G] \tag{20}$$

$$\mathfrak{D}_j = \sum_{i \neq j} \omega_i \mathrm{e}_i \tag{21}$$

where $\varphi\left(a, [b,\ ..., z]\right) = [a \cdot b,\ ...,\ a \cdot z]$, and d_k represents the network dimensions. These are introduced in detail in [32].

Finally, the action-value function is

$$Q^j(\mathrm{X}, \mathrm{A}) = h_j\left(\mathrm{e}_j, \mathfrak{D}_j\right) \tag{22}$$

where h_j denotes the light-orange feed forward network with an input layer, two hidden layers of ReLu neurons, and a linear output layer.

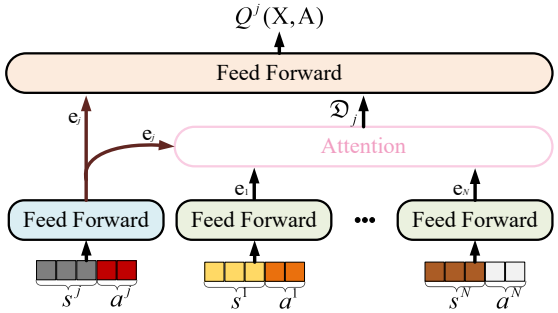


Fig. 4. The architecture of action-value network to address the variant number of agents cannot deploy on fixed neural network problem.

Actor architecture (include μ_θ and $\mu_{\theta'}$): To address the dynamic environment, the feed forward neural network is deployed in policy function-approximated neural network (shortened to policy network). Mathematically, the mapping relationship from the input state to the output action of policy network of the j-th agent can be defined as

$$\begin{aligned} I_1 &= s^j, \\ I_{u+1} &= \mathrm{ReLu}\left(I_u \cdot \mathrm{W}_u^j + \mathrm{B}_u^j\right),\ u = 1,\ ...,\ U-1 \\ \zeta &= \mathrm{ReLu}\left(I_U \cdot \mathrm{W}_U^j + \mathrm{B}_U^j\right), \\ a^j &= \tanh\left(\zeta \cdot \mathrm{W}_O^j + \mathrm{B}_O^j\right), \end{aligned} \tag{23}$$

where I_u represents the input of the u-th hidden layer; W_u^j and B_u^j are the weight and bias matrices, respectively, of the u-th hidden layer of the j-th agent; U denotes the total number of hidden layers; ζ denotes the latent feature extracted by all the U hidden layers from the input s^j ; and W_O^j and B_O^j are the weight and bias matrices, respectively, of the output layer. Specifically, the policy network consists of an input layer, four hidden layers of ReLu neurons, and a tanh output layer.

IV. Numerical Results

A. Experimental Setup

In the case study, a transformer that services eight households with one EV each is considered. The capacity of the transformer is 50 kVA. Four EV owners have HTS with $Z = 100$, and the other 4 EV owners have MTS. The charging power limit of the EVs is 6.3 kW [33]. As recommended by [34], the t_{arr}^j and $SOC_{t_{arr}^j}^j$ of the j-th EV owner are modeled as random variables. The time interval Δt is set as 1 h. Realistic data is used in this study to evaluate the performance of the proposed approach. The electricity price and basic load data have been provided by PJM [35] and the national grid [36], respectively. The data are divided into a

200-day training set (Jan 1, 2017 to Jul 19, 2017) and a 100-day test set (Jul 20, 2017 to Oct 27, 2017). The training set is utilized to train the model. When the training process is completed, the performance of the model is evaluated using the test set. The related parameters of the transformer [37], EV [33], and commuting behaviors [38] are listed in Table I. The $SOC_{t_{arr}}$ is a sample from a normal distribution $\mathcal{N}\left(0.4,0.1^2\right)$ and is bounded by 0.2 and 0.6. The arrival time t_{arr} is sampled randomly from the set $\{17,18,19\}$. The latest time to leave home t_{dep} is set as eight.

In the scenario mentioned above, the training process of the proposed approach is divided into four stages. The first stage has five populations, and each population has two agents. Each of the five populations is trained with 1000 episodes. Then, agents in different roles from the five populations are assembled to form 25 new populations. The first stage ends after the 25 new populations are evaluated with 5000 episodes. In the second stage, we select the agent sets with the top-two reward values for each role. Thus, there are three mix combinations within the same role and nine match combinations for the different roles, i.e., nine new populations. The nine populations are trained with 4000 episodes, and 81 populations are evaluated subsequently with 5000 episodes. This follows the same theory as in the first stage. The second stage also terminates after the evaluation is completed. The third stage involves a process identical to that in the second stage, except that there are 10000 training episodes. In the fourth stage, i.e., final stage, 5000 training episodes are provided to a population that has been selected from the third stage evaluation. The parameters used in the training are shown in Table II.

TABLE I
PARAMETERS OF MODEL MENTIONED IN SECTION II

Symbol	Value	Symbol	Value
n, m	0.8	θ_{A}	40℃
τ_{TO}	6.86 hours	$\Delta\theta_{\mathrm{TO,R}}$	53℃
τ_w	0.08 hours	$\Delta\theta_{\mathrm{HS,R}}$	27℃
NLL	1.8×10^5 hours	R	4.1
$\mathrm{W_{CP}}$	0.1 \$ · hour/kWh	$\mathrm{W_{LOL}}$	8305\$
$SOC_{t_{arr}}$	clip($\mathcal{N}\left(0.4,0.1^2\right)$, 0.2,0.6)	t_{arr}	{17,18,19 }
$E_{\max}$	24 kWh	P_{dsg} P_{chg}	-6.3 kW, 6.3 kW

TABLE II
THE PARAMETERS OF THE PROPOSED APPROACH

Parameters	Value
Neuron numbers of hidden layers	128
Reward discount factor	0.95
Memory capacity	1e6
Learning rate of actor	1e-4
Learning rate of critic	1e-4
Soft replacement	1e-3
Batch sizes for updating	256
Training episodes	2e4
The variance of normal distribution	0.2
The bound of noise	0.3
Policy update frequency	2
Optimizer	Adam

B. Comparisons with baselines

Comparative tests are carried out with various benchmark approaches to evaluate the performance of the proposed approach.

1) Illustrations of benchmark approaches:

• Uncontrolled approach: In this approach, the EVs are charged with the maximum charging power once these arrive at home.

• TD3 [24]: Each EV controller is modeled as a TD3 agent, which is trained separately to maximize its reward function without coordination with other EVs.

• Vanilla-MATD3 [23], [24]: It is a simplified version of the proposed approach. The difference between vanilla-MATD3 and the proposed approach is the training mechanism. Specifically, the neural networks of vanilla-MATD3 are trained in a conventional manner, rather than using the ECL-based training mechanism utilized in the proposed approach.

• Optimization of deterministic information scenario with NLopt [39]: The electricity price, basic load, and commuting behaviors of EV owners are assumed to be known beforehand. Then, NLopt is utilized to solve a deterministic optimization problem based on global information. This approach yields optimized results based on deterministic information. However, it is infeasible to achieve this in a realistic scenario owing to the randomness of the electricity price, basic load, and commuting behavior.

2) Performances on training set: The cumulative costs obtained by different RL-based control strategies are shown in Fig. 5. For an unbiased comparison, only the cumulative costs of the final training stage (i.e., Algorithm 1, line 12) of the proposed approach are plotted in Fig. 5. At the beginning of the training process, the EV controllers are incapable of making decisions to obtain low cumulative costs. Therefore, they explore the action space to accumulate experience. As the training process progresses, the EV controllers gradually learn the management strategy to achieve lower costs. Although the TD3 approach converges at approximately 10000 episodes, its cost curve shows the highest value and oscillates substantially. In this study, the working state of the transformer and the dissatisfaction of the EV owners are determined jointly by the charging/discharging powers of all the EVs. In other words, the state transition probability of each EV controller depends not only on itself, but also on other EV controllers. Such scenario can straightforwardly cause the independent EV controller to violate the Markov assumptions in a multi-agent scenario, resulting in a significantly low training effect [40]. Because the agents are only trained separately for the TD3 approach, this will lead to the absence of coordination among the EV controllers and subsequently make it difficult to guarantee the Markov assumptions. Thereby, TD3 fails to learn the valid policy. In contrast, the centralized training mechanism helps the vanilla-MATD3 approach to learn a coordinated charging management and ensures the Markov assumption owing to the utilization of the global critic during the training process. Therefore, it learns a strategy that is better and more stable than that developed by the TD3 approach. The proposed approach further enhances the stability and control performance of the learned strategy. This is achieved by systematically integrating the centralized training with the ECL mechanism. The centralized training helps the proposed approach satisfy the Markov assumption.

Furthermore, the ECL mechanism enables the proposed approach to achieve better control performance and adapt to more challenging charging tasks with a larger number of EVs that need to be managed. As a result, a more stable control strategy with better control performance is developed by the proposed approach. The results demonstrate the effectiveness of the proposed approach.

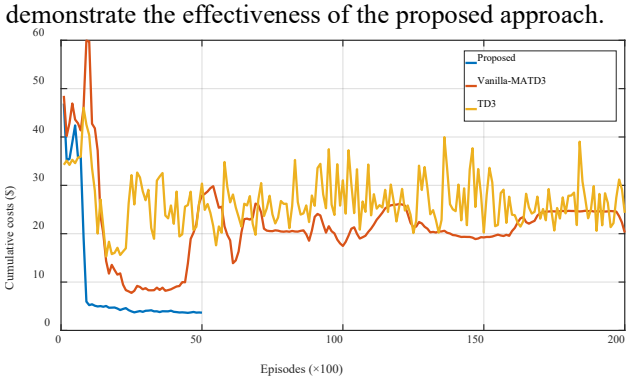


Fig. 5. Each 100-episodes average cumulative costs during the training process for different RL-based approach.

3) Performances on the test set: The cumulative costs obtained by different control strategies on the test set are shown in Fig. 6. The percentage terms on the right indicate the cost reduction ratio of the corresponding approach compared with the uncontrolled approach. The figure reveals that the cost obtained by the TD3 approach is not reduced. This is because the separate training mechanism cannot ensure the Markov assumptions. In contrast, the Markov assumption is satisfied by vanilla-MATD3 owing to the centralized training mechanism and the global critic structure. Therefore, it achieves better performance, and its cumulative cost is reduced by 59.35%. The ECL mechanism further enhances the performance of the proposed approach. Furthermore, it enables the approach to achieve a performance that is close to that of the deterministic information solutions, based only on local information. However, the deterministic information approach assumes the uncertain variables to be known beforehand and require complete communication links. These are difficult to obtain in practice.

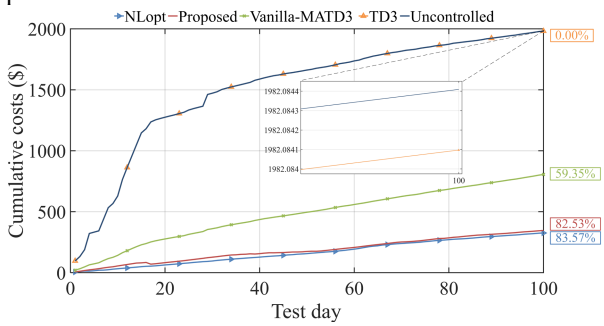


Fig. 6. Cumulative costs on the test days for the different approaches.

The cumulative cost is composed of three parts: the LOL cost, charging preference cost, and charging cost. The detailed costs obtained by different control strategies on the test set are shown in Fig. 7. The figure reveals that the TD3 and uncontrolled approaches cannot consider all the three objectives simultaneously. The LOL cost is excessively high for these. Although vanilla-MATD3 can simultaneously minimize the costs of the three objectives, the proposed approach achieves a better balance among these objectives and a lesser cost. This demonstrates the effectiveness of the proposed approach. The optimization of the deterministic information scenario has a negative LOL cost. This implies that it is profitable for the EV owners to reduce the LOL cost. This mechanism is reasonable because the EV charging behavior is controlled to maximize the interests of the transformer at the expense of those of the EV owners [3].

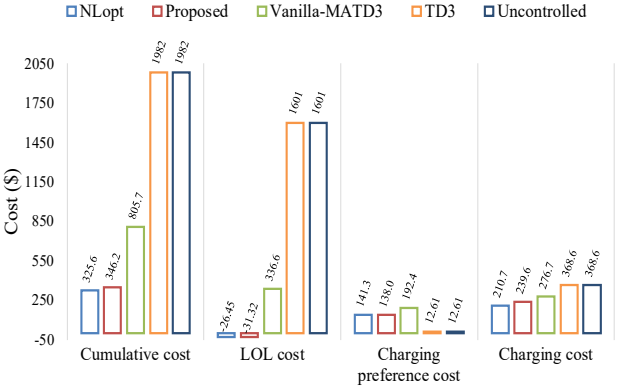


Fig. 7. Detailed cost on the test days for the different approaches.

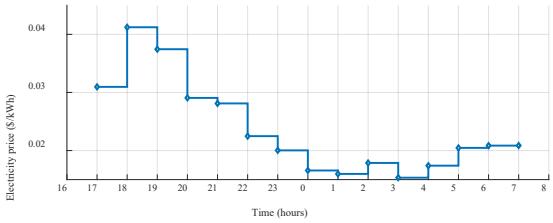
C. *Select a day on the test set for evaluating the proposed approach*

To further evaluate the performance of the proposed approach, detailed results for a test day are shown in Fig. 8. The arrows in Fig. 8(c) and (e) indicate the time of arrival at home of different EVs. The $\theta^{tot}_{\mathrm{HS},t}$ mentioned in Fig. 8(g) denotes the θ_{HS} influenced by $\mathbf{L}^{tot}_{t}$.

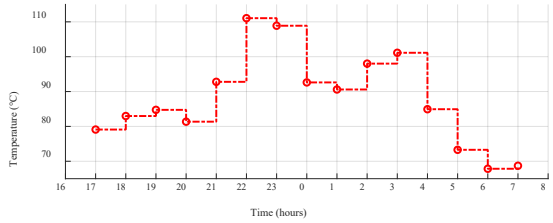
It is evident from Fig. 8(c) and (e) that each EV charges at relatively low electricity prices and discharges at relatively high electricity prices. However, the presence of CP causes the EV owners with HTS and MTS to have different charging modes. Comparing the two figures, the EV owners with HTS 1) tend to charge in an earlier charging period although the electricity price for that period is relatively high and 2) have full energy earlier than the EV owners with MTS. At the three instances when the electricity price is relatively high (17:00, 18:00, and 19:00), the EV owners with MTS opt for a high discharging power to obtain economic advantages. However, the EV owners with HTS have high charging and low discharging power at the same time. This phenomenon indicates that a HTS imparts EV owners with a tendency to access more battery-energy earlier (compared with the EV owners having MTS) before they are likely to encounter uncertain events. In addition, each EV battery has full energy when the EV departs (see Fig. 8(d) and (f)).

The transformer LOL cost also influences the EV-charging behaviors. It is highly important to understand the change rule of transformer LOL to comprehend its influence on EV-charging behaviors. In Fig. 9, each triangle denotes the difference in LOL between two hottest-spot temperatures when the temperature difference is fixed as 5 °C. For example, the first triangle denotes the LOL value at 75 °C minus that at 70 °C. As Fig. 9 shows, the difference in LOL between two temperatures increases exponentially with the increase in the hottest-spot temperature. This trend implies that for transformers having high hottest-spot temperatures, each increase in temperature would cause more damage than earlier. Therefore, whether the hottest-spot temperature can be reduced is highly important for minimizing the transformer LOL. As shown in Fig. 8(c) and (e), the instances when EV owners with HTS are on the verge of being charged fully (i.e. 22:00, 23:00, and 0:00) and those when EV owners with MTS start charging (i.e. 21:00, 22:00 and 23:00) involve a low charging power. This presents a phenomenon wherein if all the EVs are charged at short intervals, the charging power of each EV would not be significantly high. This is because if all the EVs perform at

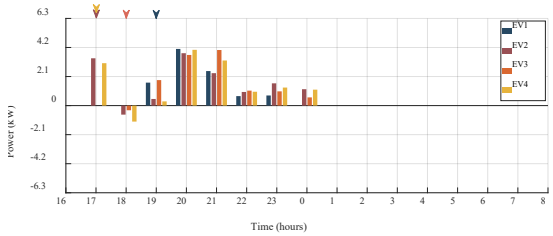
a high charging power in a short time, the excessive load would result in a significantly high probability of a rapid increase in the hottest-spot temperature. This is illustrated by the blue temperature curve generated by the uncontrolled approach in Fig. 8(g). The increased temperature would eventually cause an increase in transformer LOL (see Fig. 9). Thus, maintaining the total charge power of all the EVs within a relatively appropriate range in the charging period helps ease the growth of the hottest-spot temperature curve. This is illustrated by the orange temperature curve generated by the proposed approach in Fig. 8 (g).



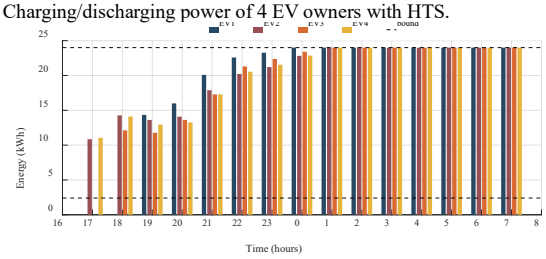
(a) Real-time electricity price.



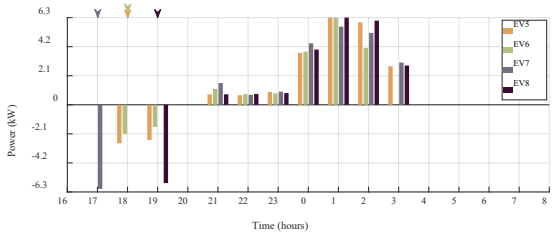
(b) The real-time winding hottest-spot temperature influenced by the sum of basic load $\theta_{\mathrm{HS},t}^{bas}$.



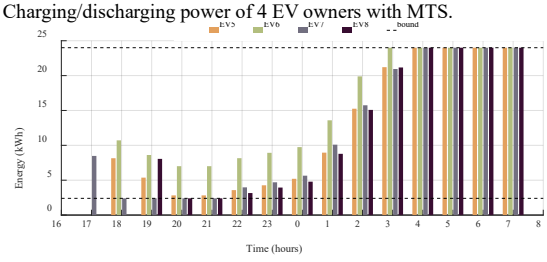
(c) Charging/discharging power of 4 EV owners with HTS.



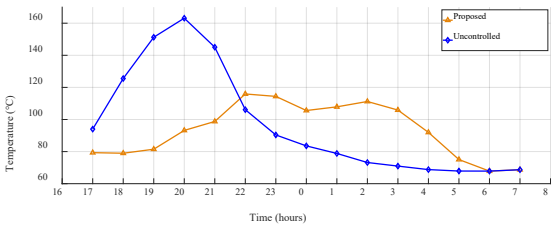
(d) The energy of 4 EV owners with HTS.



(e) Charging/discharging power of 4 EV owners with MTS.



(f) The energy of 4 EV owners with MTS.



(g) The real-time winding hottest-spot temperature influenced by the sum of total load $\theta_{\mathrm{HS},t}^{tot}$.

Fig. 8. Detailed scheduling results of the proposed approach on one test day.

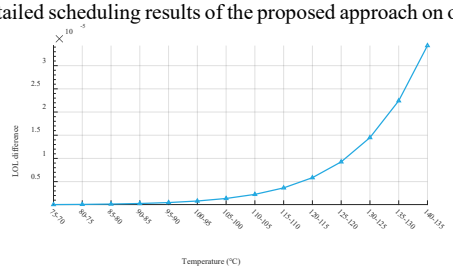


Fig. 9. The trend of LOL difference when the hottest-spot temperature difference is fixed at 5mpe

D. Discussion of weighting factor w

Further tests are carried out to investigate the impact of the selection of weighting factor on the charging behavior developed by the proposed approach. In this test, *w* is set to 0, 0.5, and 1 for cases 1, 2, and 3, respectively. The cumulative cost and its detailed cost (i.e., the transformer LOL cost, charging preference cost, and charging cost) achieved by the proposed approach when *w* is set to different values are presented in Fig. 10. The figure shows that the charging preference costs decrease when we increase *w*. This is because the individual objective becomes more important than the overall objective as we increase *w*. Because of the decrease in the importance of the overall objective, the LOL costs increase as *w* increases. This results in the increase in the cumulative cost. The results demonstrate that the proposed approach enables EV owners to balance the weight between their individual preferences and overall objective by adjusting *w*. This is different from the centralized optimization approach with a summing objective function, which may hinder the optimization of the various charging preferences of different EV owners.

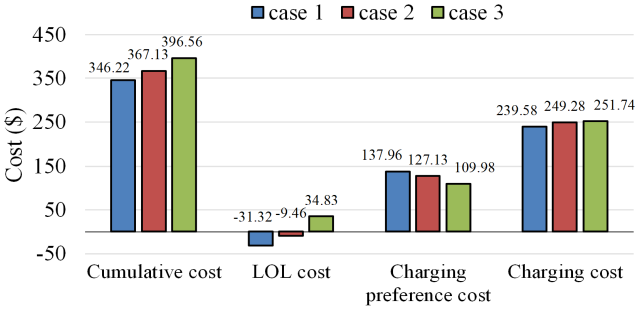


Fig. 10. Impact of the weighting factor *w* on the optimization results.

E. Scalability of proposed approach

To demonstrate the superiority of the proposed approach in terms of addressing a large population of agents, further tests are carried out when the number of EVs whose charging needs to be managed is larger. We consider six cases in this test, wherein the number of EVs is set to 8, 12, 16, 24, 32, and 64 separately. The experimental parameters of the eight-agent case have been mentioned earlier. To provide fair comparisons among these six cases, the EV penetration keeps the same level. In this setting, the transformers with ratings of 75, 100, 150, 200, and 400 kVA

services 12, 16, 24, 32, and 64 EVs, respectively. The basic load $\mathbf{L}_t^{bas}$ and economic value of transformer W_{LOL} also varies proportionally [3]. In addition, these six cases have identical commuting behaviors of EV owners. The case with 12 EVs involves 8 EV owners with HTS and 4 with MTS. The case with 16 EVs involves 8 EV owners with HTS and 8 with MTS. The case with 24 EVs involves 16 EV owners with HTS and 8 with MTS. The case with 32 EVs involves 16 EV owners with HTS and 16 with MTS. The case with 64 EVs involves 32 EV owners with HTS and 32 with MTS.

The cumulative cost incurred by the different control strategies under the different cases are shown in Fig. 11. With the gradual addition of more EVs that need to be managed, the optimization capability of vanilla-MATD3 decreases gradually. In particular, the 12-agent case is the watershed for vanilla-MATD3. Vanilla-MATD3 has better control performance than the uncontrolled approach in the cases with at most 12 agents in the environment. However, when the number of agents exceeds 12 (i.e., in the 16-, 24-, 32-, and 64-agent cases), vanilla-MATD3 does not display the capability to optimize compared with the uncontrolled approach. Apparently, with the increase in the number of agents, the control strategy needs to optimize the charging decision in a larger action space and more complex coupling relationship between EV charging. This generates significant challenges for the optimization of vanilla-MATD3. In contrast, the proposed approach can effectively maintain the control performance when the number of agents is increased, and always displays performance that is closest to that of NLopt. The cumulative costs for the proposed approach are reduced by 57.03%, 72.04%, 81.38%, 79.31%, 77.71%, and 70.53%, respectively, compared with those obtained with vanilla-MATD3. These simulation results demonstrate the superiority of the proposed approach in addressing a large population of agents.

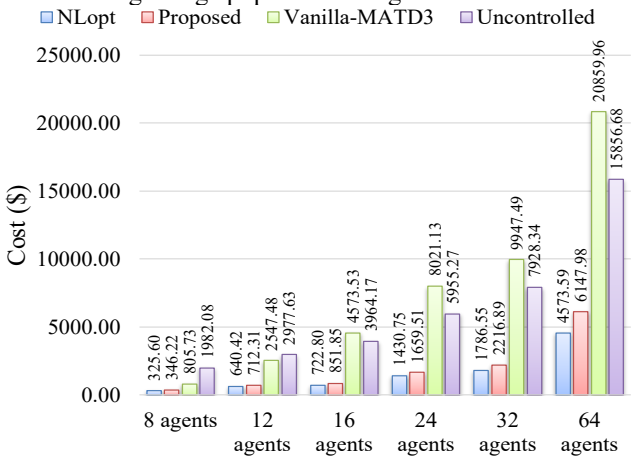


Fig. 11. Cumulative costs on the test days under the six different numbers of EVs case.

V. Conclusions

In this study, the proposed approach simultaneously optimizes the operation of transformers and the demand of EV owners by managing EV charging/discharging. It proposes a centralized ECL-based training and a decentralized deployment approach to manage larger-scale EV charging in a fully decentralized manner. The core concept underlying the ECL-based training mechanism is to learn the more straightforward task first and then, gradually increase the task difficulty. The advantage of this training approach is that it provides the agents a process to adapt gradually to the target task. This can reduce the learning difficulty of agents and thereby, yield higher performance for the target task. After the training, the advantage of a decentralized deployment can protect the EV owner's privacy, reduce the related communication cost, and prevent single-point failure. The simulations and comparative results reveal the following: 1) The training process of the proposed approach is the most stable and achieves the best performance with the training set, among the RL-based approaches. 2) In the test set, the performance of the proposed decentralized approach is close to that of the NLopt centralized approach based on the deterministic global information. 3) The interpretability of EV charging/discharging decisions further demonstrates the effectiveness of the proposed approach. 4) The proposed approach can adaptively adjust to the different preferences of EV owners by setting different weighting factors *w*. 5) The six cases with different numbers of EVs demonstrate that the proposed approach has good scalability.

VI. References

[1] 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, May 2020.

[2] Q. Gong, S. Midlam-Mohler, V. Marano *et al*., "Study of PEV charging on residential distribution transformer life," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 404-412, Mar. 2012.

[3] M. R. Sarker, D. J. Olsen and M. A. Ortega-Vazquez, "Co-optimization of distribution transformer aging and energy arbitrage using electric vehicles," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2712-2722, Nov. 2017

[4] M. Soleimani and M. Kezunovic, "Mitigating transformer loss of life and reducing the hazard of failure by the smart EV charging," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, pp. 5974-5983, Sept.-Oct. 2020.

[5] W. Fu, J. D. McCalley and V. Vittal, "Risk assessment for transformer loading," *IEEE Transactions on Power Systems*, vol. 16, no. 3, pp. 346-353, Aug. 2001.

[6] M. Humayun, M. Z. Degefa, A. Safdarian *et al*., "Utilization improvement of transformers using demand response," *IEEE Transactions on Power Delivery*, vol. 30, no. 1, pp. 202–210, Feb. 2015.

[7] P. Mohseni and R.G. Stevie, "Electric Vehicels: Holy Grail or Fools Gold," in Proc. IEEE PES General Meeting, Alberta, Canada, July 2009.

[8] F. L. D. Silva, C. E. H. Nishida, D. M. Roijers *et al*., "Coordination of electric vehicle charging through multiagent reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2347-2356, May. 2020.

[9] S. S. Karimi Madahi, H. Nafisi, H. Askarian Abyaneh *et al*., "Co-Optimization of energy losses and transformer operating costs based on smart charging algorithm for plug-In electric vehicle parking lots," *IEEE Transactions on Transportation Electrification*, vol. 7, no. 2, pp. 527-541, June 2021.

[10] S. Li, W. Hu, D. Cao *et al*., "A multi-agent deep reinforcement learning-Based approach for the optimization of transformer life using coordinated electric vehicles," *IEEE Transactions on Industrial Informatics*, doi: 10.1109/TII.2021.3139650.

[11] X. Shi, Y. Xu, Q. Guo *et al*., "A distributed EV navigation strategy considering the interaction between power system and traffic network," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3545-3557, July 2020.

[12] H. Pourbabak, J. Luo, T. Chen *et al*., "A novel consensus-based distributed algorithm for economic dispatch based on local estimation of power mismatch", *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5930-5942, Nov. 2018.

[13] E.R. Muñoz, G. Razeghi, L. Zhang *et al*., "Electric vehicle charging algorithms for coordination of the grid and distribution transformer levels," *Energy*, vol. 113, pp. 930-942, Oct. 2016.

[14] A. D. Hilshey, P. D. H. Hines, P. Rezaei *et al*., "Estimating the impact of electric vehicle smart charging on distribution transformer aging," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 905-913, Jun. 2013.

[15] N. Sadeghianpourhamami, J. Deleu and C. Develder, "Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 203-214, Jan. 2020.

[16] D. Cao, W. Hu, J. Zhao *et al*., "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4120-4123, Sept. 2020.

[17] M. Samvelyan, T. Rashid, C.S. de Witt *et al.*, "The StarCraft Multi-Agent Challenge," in *Proceedings of the 18th International Conference on*

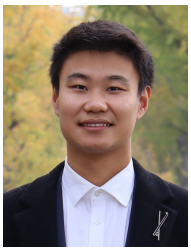
Autonomous Agents and MultiAgent Systems (AAMAS), Montreal, Canada, May 2019, pp. 2186–2188.
[18] L. Yu, Y. Sun, Z. Xu *et al.,* "Multi-agent deep reinforcement learning for HVAC control in commercial buildings," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 407-419, Jan. 2021
[19] IEEE Std C57.91-2011: 'IEEE guide for loading mineral-oil-immersed transformers and step-voltage regulators', 2011.
[20] Z. Wan, H. Li, H. He *et al.*, "Model-free real-Time EV charging scheduling based on deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5246-5257, Sept. 2019.
[21] A. Alsabbagh, B. Wu and C. Ma, "Distributed electric vehicles charging management considering time anxiety and customer behaviors," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2422-2431, Apr. 2021.
[22] Morgenstern, Oskar, and John Von Neumann, "Theory of games and economic behavior," *Princeton university press*, 1953.
[23] R. Lowe, Y. Wu, A. Tamar *et al*., "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, California, USA, Dec. 2017, pp. 6379-6390.
[24] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 2018, pp. 1587–1596.
[25] Q. Long, Z. Zhou, A. Gupta *et al*., "Evolutionary population curriculum for scaling multi-agent reinforcement learning," in *Proceedings of International Conference on Learning Representations (ICLR)*, Millennium Hall, Addis Ababa ETHIOPIA, 2020.
[26] J. L. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, no. 1, pp. 71–99, Jul. 1993.
[27] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-533, Feb. 2015.
[28] T. P. Lillicrap, J. J. Hunt, A. Pritzel *et al.*, "Continuous control with deep reinforcement learning," in *Proceedings of International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, May 2016.
[29] I. Osband, B. Van Roy, D. J. Russo *et al.*, "Deep exploration via randomized value functions," *Journal of Machine Learning Research*, vol. 20, no. 124, pp. 1–62, 2019.
[30] M. Plappert, R. Houthooft, P. Dhariwal *et al.*, "Parameter space noise for exploration," in *Proceedings of International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.
[31] M. Fortunato, M. G. Azar, B. Piot *et al.*, "Noisy networks for exploration," in *Proceedings of International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.
[32] A. Vaswani, N. Shazeer, N. Parmar, *et al*., "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, California, USA, Dec. 2017, pp. 6000–6010.
[33] K. Chaudhari, N. K. Kandasamy, A. Krishnan, *et al.*, "Agent-based aggregated behavior modeling for electric vehicle charging load," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 856-868, Feb. 2019.
[34] D. Cao, W. Hu, J. Zhao *et al.,* "Reinforcement learning and Its applications in modern power and energy systems: a review," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1029-1042, Nov. 2020.
[35] "Historical hourly electricity price from PJM", 2017. [Online]. Available: https://www.pjm.com/.
[36] "Historical hourly load data", 2017. [Online]. Available: https://www.nationalgridus.com/.
[37] R. Vicini, O. Micheloud, H. Kumar *et al.*, "Transformer and home energy management systems to lessen electrical vehicle impact on the grid," *IET Generation, Transmission & Distribution*, vol. 6, no. 12, pp. 1202-1208, Dec. 2012.
[38] S. Li, W. Hu, D. Cao *et al.*, "Electric vehicle charging management based on deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, doi: 10.35833/MPCE.2020.000460.
[39] "Library for nonlinear optimization", [Online]. Available: https://nlopt.readthedocs.io/en/latest/.
[40] G. J. Laurent, L. Matignon and N. Le Fort-Piat, "The world of independent learners is not markovian," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, vol. 15, no. 1, pp. 55–64, 2011.

Sichen Li is currently working toward the Ph.D. degree in control science and engineering from the University of Electronics Science and Technology of China, Chengdu, China.

His research interests includes demand response and the applications of machine learning in power system operation and control.

Weihao Hu (Senior Member, IEEE) received the B.Eng. and M.Sc. degrees in electrical en- gineering from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2007, respectively, and the Ph.D. degree in energy technology from Aalborg University, Aalborg, Denmark, in 2012.

He is currently a Full Professor and the Director with the Institute of Smart Power and Energy Systems, University of Electronics Science and Technology of China, Chengdu, China. He was an Associate Professor and the Vice Program Leader of Wind Power System Research Program with the Department of Energy Technology, Aalborg University. He has led/participated in more than 15 national and international research projects and he has more than 170 publications in his technical field. His research interests include artificial intelligence in modern power systems and renewable power generation.



Di Cao (Member, IEEE) received Ph.D. degree from University of Electronic Science and Technology of China (UESTC). He is currently a postdoctoral researcher at UESTC.

His research interest includes optimization of distribution network and applications of machine learning in power systems.

Zhenyuan Zhang (Senior Member, IEEE) received the B.S. degree from Chang'an University, Xi'an, China, in 2007, and the Ph.D. degree from the University of Texas at Arlington, Arlington, TX, USA, in 2015, all in Electrical Engineering. He is currently a Professor with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China. Since 2010, he has been the Project Associate with the IEEE/NFPA Arc Flash Research Project. His focus lies in smart grids and arc flash research, but he has also been involved in power system analysis, renewable energy, electrical safety analysis, and power market researches.

Qi Huang (Fellow, IEEE) was born in Guizhou, China. He received the B.S. degree in electrical engineering from Fuzhou University in 1996, the M.S. degree from Tsinghua University in 1999, and the Ph.D. degree from Arizona State University in 2003.

He is currently a Professor with Chengdu University of Technology (CDUT) and university of Electronic Science and Technology of China (UESTC). He is vice president of CDUT and the Director with the Sichuan State Provincial Lab of Power System Wide-area Measurement and Control. He is an IET Fellow and IEEE Fellow. His current research and academic interests include power system instrumentation, power system monitoring and control, and informatics for smart electric energy systems.

Zhe Chen (Fellow, IEEE) received the B.Eng. and M.Sc. degrees from the Northeast China Institute of Electric Power Engineering, Jilin, China, in 1982 and 1986, and the Ph.D. degree from the University of Durham, U.K., in 1997.

He is a Full Professor with the Department of Energy Technology, Aalborg University, Denmark. He is the Leader of the Wind Power System Research Program at the Department of Energy Technology, Aalborg University, and the Danish Principle Investigator for Wind Energy of the Sino-Danish Centre for Education and Research. He has led many research projects and has more than 500 publications in his technical fields. His research interests include power systems, power electronics, and electric machines, and main current research interests include wind energy and modern power systems.

Dr. Chen is an Editor for the IEEE TRANSACTIONS ON POWER SYSTEMS, an Associate Editor for the IEEE TRANSACTIONS ON POWER ELECTRONICS, a Fellow of the Institution of Engineering and Technology, London, U.K., and a Chartered Engineer in the U.K.

Frede Blaabjerg (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1995.

From 1987 to 1988, he was with ABB-Scandia, Randers, Denmark. He became an Assistant Professor in 1992, an Associate Professor in 1996, and a Full Professor of power electronics and drives in 1998. In 2017, he became a Villum Investigator. He is honoris causa at University Politehnica Timisoara, Timisoara, Romania, and Tallinn Technical University, Tallinn, Estonia. He has authored or coauthored more than journal papers in the fields of power electronics and its applications. He is the Co-Author of four monographs and an editor of ten books in power electronics and its applications. His current research interests include power electronics and its applications such as in wind turbines, PV systems, reliability, harmonics, and adjustable speed drives.

Dr. Blaabjerg was the Editor-in-Chief for the IEEE TRANSACTIONS ON POWER ELECTRONICS from 2006 to 2012. He has been a Distinguished Lecturer for the IEEE Power Electronics Society from 2005 to 2007 and for the IEEE Industry Applications Society from 2010 to 2011 and 2017 to 2018. During 2019–2020, he was the President of IEEE Power Electronics Society. He is also the Vice-President of the Danish Academy of Technical Sciences. He was the recipient of 32 IEEE Prize Paper Awards, the IEEE PELS Distinguished Service Award in 2009, the EPE-PEMC Council Award in 2010, the IEEE William E. Newell Power Electronics Award 2014, the Villum Kann Rasmussen Research Award 2014, the Global Energy Prize in 2019, and the 2020 IEEE Edison Medal. He is nominated during 2014–2019 by Thomson Reuters to be between the most 250 cited researchers in Engineering in the world.