Data-Driven Multi-Agent Deep Reinforcement Learning for Distribution System Decentralized Voltage Control With High Penetration of PVs

Cao, Di; Zhao, Junbo; Hu, Weihao; Ding, Fei; 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.2021.3072251

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Data-Driven Multi-agent Deep Reinforcement Learning for Distribution System Decentralized Voltage Control with High Penetration of PVs

Di Cao, *Student Member, IEEE*, Junbo Zhao, *Senior Member, IEEE*, Weihao Hu, *Senior Member, IEEE*, Fei Ding, *Senior Member, IEEE*, Qi Huang, *Senior Member, IEEE*, Zhe Chen, *Fellow, IEEE*, Frede Blaabjerg, *Fellow, IEEE*

***Abstract*—This paper proposes a novel model-free/data-driven centralized training and decentralized execution multi-agent deep reinforcement learning (MADRL) framework for distribution system voltage control with high penetration of PVs. The proposed MADRL can coordinate both the real and reactive power control of PVs with existing static var compensators and battery storage systems. Unlike the existing DRL-based voltage control methods, our proposed method does not rely on a system model during both the training and execution stages. This is achieved by developing a new interaction scheme between the surrogate modeling of the original system and the multi-agent soft actor critic (MASAC) MADRL algorithm. In particular, the sparse pseudo-Gaussian process with a few-shots of measurements is utilized to construct the surrogate model of the original environment, i.e., power flow model. This is a data-driven process and no model parameters are needed. Furthermore, the MASAC enabled MADRL allows to achieve better scalability by dividing the original system into different voltage control regions with the aid of real and reactive power sensitivities to voltage, where each region is treated as an agent. This also serves as the foundation for the centralized training and decentralized execution, thus significantly reducing the communication requirements as only local measurements are required for control. Comparative results with other alternatives on the IEEE 123-nodes and 342-nodes systems demonstrate the superiority of the proposed method.**

***Index Terms*—Voltage regulation, Gaussian process regression, network partition, multi-agent deep reinforcement learning, distribution network, PVs.**

I. Introduction

There has been increasing penetration of distributed energy resources, especially PVs into the distribution system. However, due to the uncertainty and volatility of PVs, voltage violation is becoming a concern. Numerous approaches have been proposed to regulate the voltage. They can be classified into four categories according to different communication infrastructure-enabled controls: centralized [1-3], local [4], distributed [5-6], and decentralized control [7-9]. Centralized control strategy requires extensive communication links and can obtain global optimization results. However, it suffers from a computational bottleneck and the communications are not that reliable for today's distribution systems. Local control strategy only relies on local measurements and can react fast without high communication links. Due to a lack of coordination, only suboptimal solutions can be obtained and not all constraints are fully satisfied. The distributed control strategy achieves coordination between various control devices with limited communication links. The consensus-based methods are usually used, which are vulnerable to communication delays. Decentralized control combines the advantages of centralized and distributed methods by adopting zonal control and inter-zone coordination [10].

To achieve decentralized or distributed control, network partition is a first step [11-12]. In [7], a particle swarm optimization (PSO) algorithm is used for voltage regulation of each cluster. Although the PSO algorithm is easy to implement, it cannot guarantee a global optimum. To cope with the uncertainties of DERs and load demand, a robust optimization (RO) for distribution system optimization is proposed [9]. Note that the solution of RO is achieved under the worst scenario, yielding conservative outcomes. Stochastic programming (SP) is also employed for uncertainty management [13]. It depends on pre-sampling scenarios of uncertainty realizations, and therefore accurate knowledge of the distributions of random variables is needed. This is quite challenging to obtain. SP also suffers from heavy computational burdens. It is worth noting that accurate information on system topology and parameters is required for these methods. However, the model quality of the practical distribution systems is rather poor, especially under high penetration of the behind-the-meter PVs [14].

The model assumption can be mitigated with the help of advanced machine learning-based control methods. They allow extracting knowledge from data to cope with uncertain patterns. The learned knowledge is scalable and thus can be exploited for optimization in new situations. Among these, the deep reinforcement learning (DRL) is widely used since it can learn

This work was supported by the National Key Research and Development Program of China (2018YFE0127600). Corresponding author: Weihao Hu.

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

J. Zhao is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762 (e-mail: junbo@ece.msstate.edu).

F. Ding is with the Power systems Engineering Center, National Renewable Energy Laboratory, Golden, United States (e-mail: fei.ding@nrel.gov)

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 Nuclear Technology and Automation Engineering, Chengdu University of Technology, Chengdu, China. (e-mail: hwong@uestc.edu.cn)

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

an optimal control strategy even when the global optimum is not known. To this end, several DRL-based voltage control algorithms have been developed [15-18]. However, the learned control strategies need to be executed in a centralized manner with massive communications. [19] proposes an attention mechanism based multi-agent deep deterministic policy gradient (MADDPG) algorithm for the voltage control of DNs. [20] develops a MADDPG based approach for the autonomous voltage control of transmission network. The heuristic method is first utilized to partition the whole system to several sub-regions. After that, the multi-agent AVC problem is formulated as a Markov Game, which is then solved by the MADDPG algorithm by modeling each sub-region as an intelligent agent. The proposed method is implemented in a centralized training and decentralized execution framework. During the training process, the accurate information of the physical system is required for the calculation of the reward value [20-21]. This defeats the original idea of mitigating model issues.

This paper proposes a novel model-free centralized training and decentralized execution MADRL framework for distribution system voltage control considering high penetration of PVs. It has the following contributions:

• The proposed method is model-free during both training and execution stages, which distinguishes from existing DRL-based approaches [15-21] that need a physical model during the training stage. This is achieved via two steps: system identification and voltage regulation. Instead of relying on the original inaccurate model, the sparse pseudo-Gaussian process (SPGP) is utilized to capture the complex relationship between real and reactive power injections and voltage magnitudes from a few numbers of sparse measurements. The latter requires the observability of the system that can be achieved from the distribution system state estimation via smart meters. The learned surrogate model is integrated with MADRL to provide a reward signal during the training process.

• The proposed multi-agent soft actor critic (MASAC) enabled MADRL can be executed in a decentralized manner and this is very different from existing DRL-based methods that rely on centralized training and centralized execution. We rely on the real and reactive power sensitivity relationship with voltage to partition the distribution system into a couple of regions. Each region can be conveniently treated as an agent in our MADRL framework to learn the coordination strategies. The network partition and MADRL framework serve as the foundation for centralized training and decentralized execution. It allows us to address scalability issues and significantly reduce the communication requirements as only local measurements are required for control. This also enables fast control actions.

• The proposed MADRL can coordinate both the real and reactive power control of PVs with existing static var compensators and battery storage system (BSS) to minimize the voltage deviation while maintaining a minimum amount of active power curtailment of PVs.

The remainder of this paper is organized as follows. In section II, the problem formulation is presented. Section III shows the proposed model-free MADRL framework. Section IV demonstrates the effectiveness of the method by the simulation results and Section V concludes the paper.

II. Problem Formulation

The objective of voltage control is to reduce the voltage deviation while minimizing the active power curtailment of PV. The problem is formulated as follows:

$$\min_{Q_{PV}(j,t),Q_{SVC}(i,t),P_{cur}(j,t),P_{BSS}(i,t)} F(x) = \sum_{t=1}^{T}(\sum_{i=1}^{N}|V(i,t)-V_0|+\beta\sum_{j=1}^{G}P_{cur}(j,t)) \quad (1)$$

$$V_e(i,t)\sum_{j=1}^{N}(G(i,j)V_e(j,t)-B(i,j)V_f(j,t))+ \quad (2)$$
$$V_f(i,t)\sum_{j=1}^{N}(G(i,j)V_f(j,t)+B(i,j)V_e(j,t))+P(i,t)=0,\ i\in N$$

$$P(i,t)=P_{Load}(i,t)-P_{PV}(j,t)+P_{cur}(j,t)-P_S(t)-P_{BSS}(i,t),\ i\in N,\ j\in G \quad (3)$$

$$V_f(i,t)\sum_{j=1}^{N}(G(i,j)V_e(j,t)-B(i,j)V_f(j,t))- \quad (4)$$
$$V_e(i,t)\sum_{j=1}^{N}(G(i,j)V_f(j,t)+B(i,j)V_e(j,t))+Q(i,t)=0,\ i\in N$$

$$Q(i,t)=Q_{Load}(i,t)-Q_S(t)-Q_{PV}(j,t)-Q_{SVC}(i,t),\ i\in N,\ j\in G \quad (5)$$

$$V_{\min}\le V(i,t)\le V_{\max} \quad (6)$$

$$Q_{SVC.\min}\le Q_{SVC}(i,t)\le Q_{SVC.\max} \quad (7)$$

$$0\le P_{cur}(j,t)\le \delta P_{PV}(j,t),\ \ j\in G \quad (8)$$

$$(P_{PV}(j,t))^2+(Q_{PV}(j,t))^2\le(S_{PV}(j,t))^2,\quad j\in G \quad (9)$$

$$|P_{BSS}(i,t)|\le P_{BSS.\max} \quad (10)$$

$$\begin{cases}E(i,t+1)=E(i,t)+\eta_{ch}P_{BSS}(i,t), & if\ P_{BSS}(i,t)>0\\ E(i,t+1)=E(i,t)+P_{BSS}(i,t)/\eta_{dis}, & if\ P_{BSS}(i,t)\le 0\end{cases} \quad (11)$$

$$E_{\min}\le E(i,t)\le E_{\max} \quad (12)$$

Equation (1) is the objective function, where $V(i,t)$ and V_0 represent the voltage of node *i* at time *t* and the rated voltage, respectively; $Q_{PV}(j,t)$, $P_{cur}(j,t)$, $Q_{SVC}(i,t)$, and $P_{BSS}(i,t)$ are control variables, which represent the reactive power and the active power curtailment of PV connected to node *j*, reactive power of SVC connected to node *i* during time *t*, and the active powr injection of the BSS connected to node *i* at *t*, respectively. In (1), $\sum_{i=1}^{N}|V(i,t)-V_0|$ represents the sum of voltage deviation of all nodes during *t*; $\sum_{j=1}^{G}P_{cur}(j,t)$ is the active power curtailment of all PVs during hour *t*; β represents the coefficient to balance the weight between voltage deviation and power curtailment. (2) and (4) represent the active and reactive power flow constraints at bus *i*, where $V_e(i,t)$ and $V_f(i,t)$ are the real and imaginary components of the complex voltage at bus *i* during *t*; $G(i,j)$ and $B(i,j)$ are the real and imaginary components of the complex admittance matrix elements. Equations (3) and (5) denote the active and reactive power injections at bus *i* during *t*, where $P(i,t)$ and $Q(i,t)$ represent the active and reactive power injections at bus *i* during hour *t*; $P_{Load}(i,t)$ and $Q_{Load}(i,t)$ are the active and reactive power of load demand at bus *i* during *t*; $P_{PV}(j,t)$ is the active power injection of the PV connected to node *i* at *t*; $P_S(t)$ and $Q_S(t)$ represent the active and reactive power injected at a slack bus during *t*. Equation (6) denotes the

constraint of voltage at each node, where $V_{\min}$ and $V_{\max}$ are the lower and upper bounds. (7) denotes that the reactive power of SVC should be within its capability, where $Q_{SVC.\min}$ and $Q_{SVC.\max}$ are the lower and upper bounds. (8) indicates that the active power curtailment of each PV should be within its allowed range, where δ is the maximum curtailment ratio. (9) is the relationship between the active power of PV and the reactive power of PV inverter during *t*, where $S_{PV}(j,t)$ represents the apparent power of PV inverter connected to node *i*. (10) denotes that the charging power of BSS should be within its capability, where $P_{BSS.\max}$ is the charging power limit. (11) represents the energy balance of BSS, where $E(i,t)$ is the energy level of the BSS connected to node *i* at time *t*; η_{ch} and η_{dis} are the charging and discharging coefficients, respectively. (12) denotes that the energy level of BSS should be within allowable range, where $E_{\min}$ and $E_{\max}$ are the lower and upper bounds, respectively.

There are two main challenges in solving the above optimization problem in practice: 1) the accurate information (i.e., line parameters) of the DNs is difficult to obtain for practical distribution systems with high penetration of distributed energy resources and 2) limited communication resources and time delay make the centralized control method challenging to obtain satisfactory control performance. To this end, this paper proposes a model-free decentralized control framework based on surrogate model and MADRL algorithm for the voltage control of DNs.

III. Proposed Model-Free MADRL Framework for Voltage Control

The proposed model-free MADRL framework consists of four main components, namely 1) surrogate modeling via SPGP; 2) network partition; 3) formulation of multiple sub-regions voltage regulation as a Markov game; 4) application of MASAC algorithm to solve the developed Markov game.

A. Gaussian Process for Surrogate Modeling

It has been elaborated in the introduction that existing DRL-based voltage control algorithms have to use the original power flow model for reward calculations during the training process. To deal with that, we propose to develop a surrogate model that yields the same input-output relationship as the power flow equations. In this paper, we advocate the use of SPGP as it only needs a few numbers of measurements and can achieve good performance.

A Gaussian process (GP) is denoted as $f(x) \sim GP(m(x), k(x,x'))$, which is specified by a mean function $m(x)=E[f(x)]$ and covariance $k(x,x') = \mathrm{cov}(f(x), f(x'))$ [22]. Typically, $m(x)$ is assumed to be 0 since we have no prior knowledge about it. For the regression problem $y = f(x) + \varepsilon$, where x is the input and y represents the response by $f(x)$ corrupted by noise $\varepsilon \sim N(0, \sigma_n^2)$. The prior distribution of the observed value is

$$y \sim N(0, K(X,X) + \sigma_n^2 I_n) \tag{13}$$

where $K_{ij} = k(x_i, x_j)$ represents the kernel function to express the correlation between x_i and x_j; I_n is an $N \times N$ identity matrix. GP aims to forecast f_* given new input x_*. The joint prior distribution of *y* and the predicted value f_* is

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N(0, \begin{bmatrix} K(X,X)+\sigma_n^2 I_n & K(X,x_*) \\ K(x_*,X) & k(x_*,x_*) \end{bmatrix}) \tag{14}$$

where $K(X,x_*)$ denotes the *N*-dimensional covariance vector between the training points *X* and test points x_*; $k(x_*,x_*)$ is the variance of x_*. Then the posterior distribution of f_* is

$$f_* \mid X, y, x_* \sim N(\bar{f}_*, \mathrm{cov}(f_*)) \tag{15}$$

$$\bar{f}_* = K(x_*,X)[K(X,X)+\sigma_n^2 I_n]^{-1} y \tag{16}$$

$$\mathrm{cov}(f_*) = k(x_*,x_*) - K(x_*,X)\cdot[K(X,X)+\sigma_n^2 I_n]^{-1} K(X,x_*) \tag{17}$$

where $\bar{\mu}_* = \bar{f}_*$ and $\bar{\sigma}_{f_*}^2 = \mathrm{cov}(f_*)$ represent the mean and covariance of f_*. GP may suffer from a heavy computational burden when dealing with a large amount of data. To this end, SPGP is proposed. The pseudo set $\bar{D} = (\bar{X}, \bar{f})$ is used to substitute for the real data set, where $\bar{X} = \{\bar{x}_i\}_{i=1}^m$ and $\bar{f} = \{\bar{f}_i\}_{i=1}^m$ represent the input and output of the pseudo points. The observed output corrupted by noise is

$$y \mid x_*, \bar{X}, \bar{f} \sim N(\bar{k}_m(x_*)^T \bar{K}_{mm}^{-1} \bar{f}, \bar{k}(x_*,x_*) - \bar{k}_m(x_*)^T \bar{K}_{mm}^{-1} \bar{k}_m(x_*) + \sigma_n^2) \tag{18}$$

where $[\bar{k}_m(x_*)]_i = K(\bar{x}_i, x_*)$, $[\bar{K}_{mm}]_{ij} = K(\bar{x}_i, \bar{x}_j)$. Then the posterior distribution of f_*^{SP} is derived as [22]:

$$\bar{f}_*^{SP} = \bar{k}_m(x_*)^T \bar{Q}_{mm}^{-1} \bar{K}_{mn} (\Lambda + \sigma_n^2 I)^{-1} y \tag{19}$$

$$\mathrm{cov}(f_*^{SP}) = k(\bar{x}_*, \bar{x}_*) - \bar{k}_m(x_*)^T \times (\bar{K}_{mm}^{-1} - \bar{Q}_{mm}^{-1}) \bar{k}_m(x_*) + \sigma_n^2 \tag{20}$$

where

$$[\bar{K}_{nm}]_{ij} = K(x_i, \bar{x}_j),\ \Lambda = diag(\lambda)$$
$$\lambda_n = K_{nn} - \bar{K}_{nm}^T \bar{K}_{mm}^{-1} \bar{K}_{mn},\ \bar{Q}_{mm} = \bar{K}_{mm} + \bar{K}_{mn} (\Lambda + \sigma_n^2 I)^{-1} \bar{K}_{nm} \tag{21}$$

In this paper, the square exponential covariance function is used as the kernel function. It can be expressed as:

$$k(x,x') = \sigma_f^2 \exp(-\frac{1}{2}(x-x')^T M^{-1} (x-x')) \tag{22}$$

where $M = diag(l^2)$ and l represents the variance scale; σ_f^2 is the signal variance. The hyper-parameter set can be denoted as $\theta_s = \{M, \sigma_f^2, \sigma_n^2\}$. When the hyper-parameter set is fixed and the optimal location of the pseudo set is determined, the mean and covariance of the predicted value can be obtained according to (19) and (20). In our MADRL, the developed (19) is used to interact with MADRL for award calculation. Since it is inferred from data, no physical power flow model is needed.

B. Network Partition

A network partition aims to divide the whole network into several sub-regions such that the centralized voltage regulation problem is separated into several sub-problems that can be solved in a distributed manner. In this paper, the voltage sensitivity based on electrical distance is used to aggregate the nodes with the similar property while considering the regional voltage regulation ability. The electrical distance d_{ij} is defined based on the voltage-active power sensitivity and voltage-reactive power sensitivity matrix as follows [8]:

$$d_{ij} = d_{ij}^{VP} + d_{ij}^{VQ} \tag{23}$$

where d_{ij}^{VP} and d_{ij}^{VQ} represent the electrical distance defined based on voltage-active power sensitivity and voltage-reactive power sensitivity, respectively. They are expressed as

$$d_{ij}^{VP} = S_{ii}^{VP} + S_{jj}^{VP} - S_{ij}^{VP} - S_{ji}^{VP},\ d_{ij}^{VQ} = S_{ii}^{VQ} + S_{jj}^{VQ} - S_{ij}^{VQ} - S_{ji}^{VQ} \quad (24)$$

where S_{ij}^{VP} and S_{ij}^{VQ} represent the sensitivity of the voltage of node *i* to the injected active and reactive power of node *j*, respectively. The smaller the electrical distance d_{ij} is, the tighter the electrical connection between node *i* and *j* will be. The voltage regulation capability is defined as:

$$c = \min\{c_1, c_2, ..., c_k\},\ c_k = clip((\sum_{i \in k} Q_i S_{ij}^{VQ} + P_i S_{ij}^{VP}) / \Delta V_j, 0, 1) \quad (25)$$

where *c* represents the voltage regulation capability of the whole network; c_k is the regional voltage capability of the cluster *k*; ΔV_j denotes the voltage deviation of the node with the maximum voltage violation value; Q_i and P_i represent the maximum reactive power and active power curtailment that can be used by node *i* for voltage regulation, respectively. The clip function restricts the range of c_k within [0,1]. The performance index of the clustering is defined based on modularity index:

$$\rho = \frac{1}{a}\sum_i\sum_j [W_{ij} - \frac{l_i l_j}{a}] * \delta(i,j) * c,\ \delta(i,j) = \begin{cases} 1, \text{ if } i, j \in k \\ 0, \text{ else} \end{cases},\ W_{ij} = 1 - \frac{d_{ij}}{\max\limits_{m,n \in N} d_{mn}} \quad (26)$$

where $a = \sum_i \sum_j W_{ij}$ represents the sum of the weights for all nodes; $l_j = \sum_j W_{ij}$ represents the weight of node *j*. For the performance index, the larger the value is the closer the electrical connection inside the cluster and a looser connection among clusters. A tabu algorithm can be used to search for the optimal clusters while maximizing the defined performance [8].

C. Formulation of Markov Games

After network partition, the centralized voltage regulation problem is separated into several sub-problems. In this section, the control of multiple sub-networks is formulated as a Markov Game (MG), whose components are defined as follows:

- **Agents**: In the MG, each agent represents a sub-network.
- **Environment**: The surrogate model that calculates the voltage deviation according to states and actions by agents.
- **State**: The global state set at *t*, S_t, includes all agents' states.

For agent *j*, the state s_t^j includes the local observations of the *j*th sub-network. For the tests when PVs and SVCs are utilized for voltage regulation, the state of agent *j* is defined as $s_t^j = (P_{Load}(i,t), Q_{Load}(i,t), P_{PV}(j,t))$. When PVs and BSSs are used, the state is $s_t^j = (P_{Load}(i,t), Q_{Load}(i,t), P_{PV}(j,t), E(i,t))$.

- **Action**: The action set at *t*, A_t, includes all agents' actions.

For agent *j*, the action a_t^j contains the control variables within the *j*th sub-network. When PVs and SVCs are utilized for voltage regulation, the action of agent *j* is defined as $a_t^j = (\alpha_{PV}(j,t), \alpha_{SVC}(i,t), \alpha_{cur}(j,t))$. Then, the control variables in (1) can be obtained by:

$$\begin{aligned} Q_{PV}(j,t) &= \alpha_{PV}(j,t)\sqrt{(S_{PV}(j,t))^2 - (P_{PV}(j,t))^2},\ -1 \le \alpha_{PV}(j,t) \le 1 \\ Q_{SVC}(i,t) &= \alpha_{SVC}(j,t) Q_{SVC.\max},\ -1 \le \alpha_{SVC}(j,t) \le 1 \\ P_{cur}(j,t) &= \alpha_{cur}(j,t) P_{PV}(j,t),\ 0 \le \alpha_{SVC}(j,t) \le \delta \end{aligned} \quad (27)$$

By contrast, when PVs and BSSs are utilized, the action is defined as $a_t^j = (\alpha_{PV}(j,t), \alpha_{cur}(j,t), \alpha_{BSS}(i,t))$. Then, the control variable $P_{BSS}(i,t)$ can be obtained by $P_{BSS}(i,t) = \alpha_{BSS}(i,t) P_{BSS.\max}$. To avoid the violation of constraint (12), the action that is actually performed is

$$P_{BSS}(i,t) = \begin{cases} \eta_{dis}(E_{\min} - E_t(i,t)), \text{ if } P_{BSS}(i,t) < 0 \text{ and } E(i,t) + P_{BSS}(i,t)/\eta_{dis} < E_{\min} \\ (E_{\max} - E_t(i,t))/\eta_{ch}, \text{ if } P_{BSS}(i,t) > 0 \text{ and } E(i,t) + \eta_{ch} P_{BSS}(i,t) > E_{\max} \\ P_{BSS}(i,t), \text{ others} \end{cases} \quad (28)$$

where the charging/discharging actions of BSS are between $[\eta_{dis}(E_{\min} - E_t(i,t)), (E_{\max} - E_t(i,t))/\eta_{ch}]$. Then inequality constraints (10) and (12) can be satisfied. It is worth noting that since the proposed DRL is the off-policy algorithm, the bounded actions can also be used for training.

- **Reward**: r_t represents the immediate reward the agent obtains when action A_t is executed under state S_t. All agents share the same reward in this study, i.e.,

$$r_t = -(\sum_{i=1}^{N} |V(i,t) - V_0| + \beta \sum_{j=1}^{G} P_{cur}(j,t)) + \eta \quad (29)$$

where $V(i,t)$ is calculated by the surrogate model; $P_{cur}(j,t)$ is calculated by (27) according to action $\alpha_{cur}(j,t)$ and state $P_{PV}(j,t)$; η is the penalty term if voltage violation exists.

- **State transition**: The state transition is denoted as $s_{t+1}^j = f(s_t^j, a_t^j, \omega_t)$. When PVs and SVCs are utilized for voltage regulation, the state transitions of $P_{Load}(i,t)$, $Q_{Load}(i,t)$, and $P_{PV}(j,t)$ are mainly affected by the randomness of environment ω_t since the accurate values of load demand and PV generation are unknown. When PVs and BSSs are utilized, the state transition for the energy level of BSS is controlled by $P_{BSS}(i,t)$ and can be explicitly modeled by (11). It is difficult to find the accurate distribution of ω_t as it is affected by uncertain variables, such as load demand and PV generations. To this end, a model-free approach is proposed in this paper.

At each time-step, each agent obtains a local observation of its corresponding sub-network s_t^j, based on which action a_t^j is made. Then all agents receive an immediate reward r_t according to the curtailment PV output and the voltage deviation calculated by the surrogate model.

D. Solutions via MASAC Algorithm

The MASAC algorithm is used to solve the MG by modeling each sub-region as a SAC agent within the centralized training framework. Each SAC agent employs two functions for different purposes: the actor function takes the local observation of each sub-region s_t^j as input and outputs the action a_t^j; the critic function takes the global information (S_t, A_t) as input and outputs the judgment of the actor's

decision. All agents are trained in a centralized manner to develop a coordinated control strategy [23].

Consider an MG with *N* agents and policy set $\pi = \{\pi_1, ..., \pi_N\}$, where policy π_i represents the actor function of agent *i* parameterized by μ_i. Different from standard DRL algorithm that aims to maximize the cumulative reward, the objective of the actor function of SAC is to maximize the sum of the expected reward and a entropy term $\sum_{t=0}^{T} E[r_t + \alpha \mathrm{H}(\pi_i(\cdot \mid s_t^i))]$, where $\mathrm{H}(\pi^{\mu_i}(\cdot \mid s_t^i)$ is the entropy term; α represents the temperature parameter used to balance the two terms in the objective function. The introduced entropy term encourages the agents to explore more widely in the policy space by acquiring more diverse behaviors [24]. The parameters of the actor function are adjusted according to [25]:

$$\nabla_{\mu_i} J(\mu_i) = E_{S_t, A_t \sim D}[\nabla_{\mu_i} \log(\pi^{\mu_i}(a_t^i \mid s_t^i)) \rho_i(S_t, a_t^1, ..., a_t^N)] \quad (30)$$

$$\rho_i(S_t, a_t^1, ..., a_t^N) = -\alpha \log(\pi^{\mu_i}(a_t^i \mid s_t^i)) + Q_i^{\pi}(S_t, a_t^1, ..., a_t^N) - b(S_t, a_t^{\setminus i}) \quad (31)$$

$$b(S_t, a_t^{\setminus i}) = E_{a_t^i \sim \pi^{\mu_i}(s_t^i)}[Q_i^{\pi}(S_t, (a_t^i, a_t^{\setminus i}))] \quad (32)$$

where $Q_i^{\pi}(S_t, a_t^1, ..., a_t^N)$ is the value of current action; $b(S_t, a_t^{\setminus i})$ is the baseline term, indicating the value of the average action for agent *i*; $Q_i^{\pi}(S_t, a_t^1, ..., a_t^N) - b(S_t, a_t^{\setminus i})$ represents the advantage of current action as compared to the baseline term.

In (31), the action value of agent *i* $Q_i^{\pi}(S_t, a_t^1, ..., a_t^N)$ is calculated by the critic function $Q_i^{\pi}(\cdot)$, which optimizes its parameters following a Q-learning iteration. The parameters of the critic function are optimized by minimizing the following loss objective [25]:

$$L = (y_t - Q_i^{\pi}(S_t, a_t^1, ..., a_t^N))^2 \quad (33)$$

$$y_t = r_t^i + \gamma E_{a_{t+1}^i \sim \pi^{\mu_i}}[-\alpha \log(\pi^{\mu_i}(a_{t+1}^i \mid s_{t+1}^i)) + Q_i'^{\pi'}(S_{t+1}, a_{t+1}^1, ..., a_{t+1}^N)] \quad (34)$$

where y_t represents the target value; $\pi^{\mu_i'}$ and $Q_i'^{\pi'}$ are the target actor and critic introduced to stabilize the training process. The parameters of the target networks are typically updated by slowly tracking the online ones. SAC algorithm follows the calculation of y_t in the double Q-learning method by adopting a pair of critics $(Q_{i,1}^{\pi}, Q_{i,2}^{\pi})$, yielding

$$y_t = r_t^i + \gamma E_{a_{t+1}^i \sim \pi^{\mu_i}}[-\alpha \log(\pi^{\mu_i}(a_{t+1}^i \mid s_{t+1}^i)) + \min_{n=1,2} Q_{i,n}'^{\pi}(S_{t+1}, a_{t+1}'^1, ..., a_{t+1}'^N)] \quad (35)$$

where $Q_{i,n}'^{\pi}(\cdot)$ represents the *n*th target critic function of agent *i*. $\pi^{\mu_j'}(\cdot)$ is the target actor function of agent *j*. By maintaining a pair of critics and taking the minimum value between them, the overeastimation can be efficiently reduced.

The SAC algorithm also introduces the experience replay mechanism to promote training stability. The proposed MASAC algorithm consists of *N* agents, each with an *M*-sized replay buffer $D_i = \{m_1^i, ..., m_M^i\}$. At each time-step, the transition experience $m_t^i = \{s_t^i, a_t^i, r_t^i, s_{t+1}^i\}$ is stored in the memory D_i. When the number of stored experiences reaches the upper limit, the old memory is replaced with the new ones. In the training process, a mini-batch of experience is sampled from the replay buffer for calculating gradients and performing optimization.

To cope with the complex nonlinearities in the voltage regulation problem, deep neural networks (DNNs) are used to approximate the actor and critic functions. Then, the parameters of functions are replaced by the weight matrix and bias vector. For agent *i*, the parameter set to be optimized can be represented as $\theta_i = \left\{\theta_i^{\mu}, \theta_i^{\mu'}, \theta_i^{Q_1}, \theta_i^{Q_1'}, \theta_i^{Q_2}, \theta_i^{Q_2'}\right\}$, where θ_i^{μ} and $\theta_i^{\mu'}$ represent the parameters of the actor and target actor networks; $\theta_i^{Q_1}$ and $\theta_i^{Q_1'}$, $\theta_i^{Q_2}$ and $\theta_i^{Q_2'}$ are the parameters of the critic and target critic networks of $Q_{i,1}^{\pi}$ and $Q_{i,2}^{\pi}$.

E. Centralized Training with Surrogate Model

The proposed method contains two sets of parameters: the parameters of the surrogate model θ_s and the MADRL algorithm θ_c. The training procedure is shown in Table I.

TABLE I Centralized Training of Model-Free MADRL

Algorithm Training of proposed MADRL

1: Randomly initialize θ_s and obtain the prior model
2: Update θ_s and determine the location of the pseudo set
3: Output the posterior model as the surrogate model
4: Randomly initialize parameters of NNs θ_c
5: for episode =1,2,…, *M* do
6: Receive initial observation s_0^i for each agent
7: for *t*=1,2,…*T* do
8: determine control action $a_t^i = \pi_i(s_t^i \mid \theta_i^{\mu})$ for each agent, execute actions $A_t = (a_t^1, ..., a_t^N)$, calculate reward r_t by surrogate model according to (19) and (29), and observe the new state s_{t+1}^i
9: for agent *i* = 1,…, *N* do
10: store transition $m_t^i = \{s_t^i, a_t^i, r_t^i, s_{t+1}^i\}$ in memory D_i
11: if memory capacity is full, do
12: sample a random mini-batch *B* of transitions from memory *D*
13: update critic networks according to (36) and (37)
14: update actor-networks according to (38) and (39)
15: update target networks according to (40)
16: end if
17: end for
18: end for
19: end for

The parameter set of the surrogate model is represented as $\theta_s = \{M, \sigma_f^2, \sigma_n^2\}$. The input of the algorithm is $P(i,t)$, $Q(i,t)$, and $V(i,t)$ and the output is θ_s. Firstly, the hyper-parameters are initialized and the prior model is obtained. Then, optimize the parameters through the maximum likelihood method and determine the optimal location of the pseudo set. After that, the posterior model is used as the surrogate model.

The parameter set of the control model is represented as $\theta_c = \{\theta_1, ..., \theta_N\}$, where $\theta_i = \left\{\theta_i^{\mu}, \theta_i^{\mu'}, \theta_i^{Q_1}, \theta_i^{Q_1'}, \theta_i^{Q_2}, \theta_i^{Q_2'}\right\}$ represents the parameter set of all the actor and critic NNs of agent *i*. The input includes $P_{Load}(i,t)$ and $Q_{Load}(i,t)$ of each node, $P_{PV}(j,t)$ of all the PVs, and the reward r_t calculated by the surrogate

model. The output is θ_c. The training of DNN is based on the DRL algorithm and the centralized training framework.

In the beginning, θ_c is randomly initialized. The parameters of the target NNs are copied from online NNs. Then, the algorithm is trained for *M* episodes to update the parameters. An episode includes 24 time-steps, each corresponding to an hour. At each time-step, each agent makes decision a_t^i based on its local observation s_t^i, yielding an immediate reward r_t calculated by the surrogate model via (19) and (29); then the system transfers to the next state s_{t+1}^i. Next, the transition experience $m_t^i = \{s_t^i, a_t^i, r_t^i, s_{t+1}^i\}$ is stored in its memory D_i. From $0 - M_1$ episodes, the actions are randomly selected to fully explore the environment. From $M_1 - M$ episodes, the actions of each agent are selected according to its policy function π^{μ_i} and the parameters are optimized utilizing the transition stored in replay buffer. Specifically, a mini-batch of experiences are sampled from *D* to calculate the gradient of the NNs. The update equations of the critic networks are:

$$L(\theta_i^{Q_n}) = \frac{1}{B}\sum_{k=1}^{B}(Q_{i,n}^{\pi}(S_k, a_k^1, ..., a_k^N) - y)^2, \quad n = 1, 2 \tag{36}$$

where *B* represents the size of the mini-batch. The loss function is minimized by optimizing the parameters of critic NNs based on the gradient rule:

$$\theta_i^{Q_n} \leftarrow \theta_i^{Q_n} + \eta_Q \nabla_{\theta_i^{Q_n}} L(\theta_i^{Q_n}), \quad n = 1, 2 \tag{37}$$

where η_Q represents the learning rate for critic networks. The parameters of the actor NNs are updated via

$$\nabla_{\mu_i} J(\mu_i) = E_{S_t, A_t \sim D}[\nabla_{\mu_i} \log(\pi^{\mu_i}(a_t^i \mid s_t^i)) \rho_i(S_t, a_t^1, ..., a_t^N)] \tag{38}$$

Then, the gradient rule is applied:

$$\theta_i^{\mu} \leftarrow \theta_i^{\mu} + \eta_{\mu} \nabla_{\theta_i^{\mu}} J(\theta_i^{\mu}) \tag{39}$$

where η_μ represents the learning rate for actor networks. The parameters of target NNs are optimized via

$$\theta_i^{\mu'} \leftarrow \tau\theta_i^{\mu} + (1-\tau)\theta_i^{\mu'}, \theta_i^{Q_n'} \leftarrow \tau\theta_i^{Q_n} + (1-\tau)\theta_i^{Q_n'}, \quad n = 1, 2 \tag{40}$$

where $\tau << 1$ represents the tracking coefficient.

F. Real-Time Decentralized Execution

TABLE II Decentralized Execution

Algorithm Decentralized control by proposed MADRL

1: Load the parameters of actor-network of each agent θ_i^{μ}
2: for time step t=1,2,…*T* do
3: for agent *i* = 1,…, *N* do
4: obtain the local observation s_t^i
5: determine action a_t^i according to $a_t^i = \pi_i(s_t^i \mid \theta_i^{\mu})$
6: end for
7: execute the concatenated actions $A_t = (a_t^1, ..., a_t^N)$
8: end for

The parameters of NNs will be fixed when the training procedure is completed and only the actor NNs are kept for real-time control. Each actor NNs takes the local information of its corresponding sub-network as inputs and outputs control decisions in real-time. Since the critic NNs explicitly model the policy of other agents during training, the actor learns a coordinated strategy and can exhibit cooperative behavior using only local information. This allows us a decentralized execution using only local measurements, achieving a significant reduction of communication requirements, and enabling its scalability to large-scale systems.

IV. Numerical Results

In this section, simulations are carried out on the IEEE 123-node and practical 342-node systems [26] to evaluate the performance of the proposed model-free MADRL method. Comparative tests with other regression algorithms are first provided, followed by comparison results with other voltage control methods. For the modified 123-bus system, the parameters of the installed PVs and SVCs are listed in Table III. In this study, the total number of nodes for each cluster is limited in 14-38. The optimal partition results are shown in Fig. 1. The performance index is 0.09. For the PVs, actual data lasting 360 days in Xiaojin, a county in the Sichuan province of China are used. The PV output data are separated into the training set (300 days) and test set (10 days). The hyper-parameters of the control model are listed in Table IV. Both the surrogate model and the control model are implemented in Python. A workstation with an Intel i9-7900X CPU and an NVIDIA GeForce 2080Ti GPU is used for the simulation.

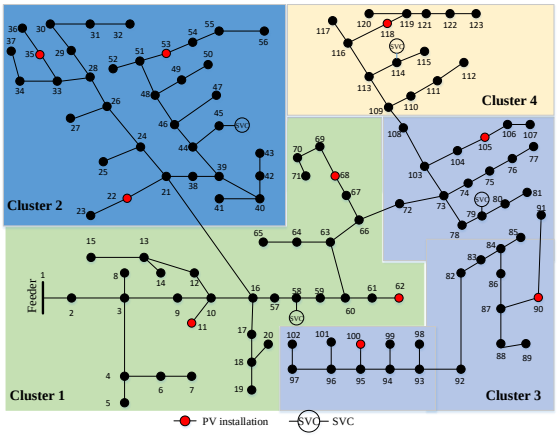

Fig. 1. The partition results of the modified IEEE 123-bus system.

Table III Parameters of controllable devices

Type	Capacity	Location
SVC	0.3MVar	45, 58, 79, 114
PV	1.2MW/1.26MVA	11, 22, 35, 53, 62, 68, 90, 100, 105, 118

Table IV Parameters of the control model

Parameter	Value
Batch size for updating NN	256
Replay buffer size	24000
Temperature parameter	1.25e-3
Discount factor	0
Soft update coefficient	0.001
Learning rate for actor/critic-network	0.001/0.001
Neuron number of hidden layer	100/100/100

Table V MAE comparisons under different training instances

Training instances	LR	DNN	SPGP
200	3.02e-3	1.29e-3	9.70e-4
500	2.23e-3	7.74e-4	4.88e-4
1500	1.96e-3	4.48e-4	2.58e-4

A. Performance Evaluation of the Surrogate Model

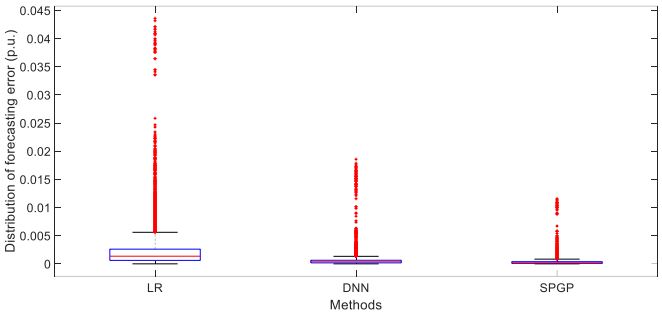

Fig. 2. The distributions of the prediction errors for three methods.

Comparative tests are carried out among the linear regression (LR), DNN, and SPGP to evaluate the forecasting accuracy of the proposed surrogate model when there are few recorded samples. 3000 instances of data with $\boldsymbol{P}$, $\boldsymbol{Q}$, and $\boldsymbol{V}$ are generated by the AC power flow. Among these, 1000 instances are used as the test set to evaluate the performance of the trained model. The mean absolute error (MAE) is used as performance index.

Table V displays the MAE of each method under a different number of training instances. DNN has three hidden layers, the neuron number of which are 400, 200, and 200, respectively. The batch-size and learning rate are selected as 32 and 0.001, respectively. The DNN is trained for 30000, 17500, and 17500 epochs to learn the relationship between $\boldsymbol{P}$, $\boldsymbol{Q}$, and $\boldsymbol{V}$ when 200, 500, and 1500 instances of data are utilized for training. The proposed SPGP adopts the square exponential covariance function. It can be observed that when the training set consists of only 200 instances of data, the forecasting accuracy of the SPGP method has better performance than the LR and DNN. The DNN yields poor performance on the test data because it involves too many parameters to be optimized and therefore suffers from the over-fitting issues. The performance of the LR method is the worst because it is unable to capture the complex nonlinear relationship between the power injection and the voltage magnitude of each node. The improvement of forecasting accuracy of the three methods can be observed when we gradually increase the number of training instances. Note that the proposed SPGP can always achieve the best performance. The distributions of the prediction error for various methods are shown in Fig. 2 when the number of training instances is selected as 1500. It can be observed that the prediction error of the SPGP method is distributed in a very small range that is close to zero. The maximum prediction errors of the LR, DNN, and the SPGP are 0.044, 0.019, and 0.012, respectively.

B. Performance Evaluation of the Control Model

To demonstrate the benefits of the proposed model-free MADRL control method, comparative results with other ones are carried out. First, the evoluation of cumulative reward during the training procedure of the proposed method is compared with other methods, including 1) **multi-agent twin delayed deep deterministic policy gradient (MATD3) algorithm** [23][27], where each sub-region is modeled as a TD3 agent and all agents are tranined in a centralized manner according to the reward signal calculated by the surrogate model. The neuron number of each hidden layer, replay buffer size, soft update coefficient, and discount factor are set to the same values as those in the proposed method. The batch size, policy update frequency, target policy smoothing coefficient, and learning rate for actor/critic network are set as 128, 2, 0.2, and 0.001/0.002, respectively; 2) **the SAC method adopting a centralized control framework** (SAC-C) [24], where the whole distribution system is modeled as a SAC agent, which is trained by continuous interaction with the surrogate model. The learning rate for the actor/critic network are set to 3e-4/3e-4, while the other parameters are the same as the proposed method.

1) Centralized Training Evaluation

All agents are trained with 50000 episodes to learn a coordinated voltage regulation strategy. The evolution of the cumulative reward during the training procedure is shown in Fig. 3, when β is set to 0.1. In the first 1000 episodes, the parameters of NNs are fixed and the actions of agents are

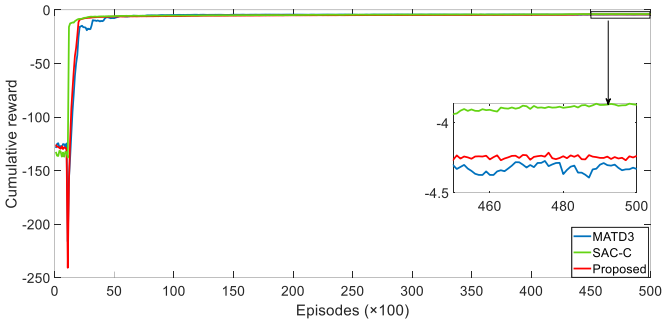

Fig. 3. The evolution of the cumulative reward for different methods during the training procedure on the IEEE 123-node system.

randomly selected to explore the environment. From 1000 episodes onward, the actions are selected by the NNs whose parameters are optimized concurrently. It can be observed that the agents obtain a low reward value at the beginning. This demonstrates that the initialized policy is incapable of making good decisions to obtain a high reward. However, the cumulative reward continuously increases and this indicates that the agents gradually evolve and learn a policy to obtain a high reward. Thanks to the entropy term that can improve the exploration capability of the SAC algorithm, the proposed method finally converges to a higher value than the MATD3 method. Since the SAC-C method takes the global information as inputs, it can learn a better control strategy than the MADRL method based on only local information. The results validate the integration of the surrogate model with MADRL during the training stage.

TABLE VI Comparison results for different strategies on test data

Method	Ave. Dev.	Ave curt. (MW)	Max. rise	Max. drop	Para. dep.
Original	2.14%	-	6.02%	5.88%	-
SAC-D	3.20%	1.02	9.42%	5.81%	✓
MATD3	0.16%	0.24	0.70%	0.80%	×
SP	0.16%	1.61	0.91%	0.97%	✓
Proposed	0.15%	0.20	0.67%	0.84%	×
MASAC	0.14%	0.21	0.68%	0.78%	✓
SAC-C	0.12%	0.12	0.63%	0.78%	✓

2) Evaluation of the Learned Strategy on Test Data

When the training process is completed, comparative tests with more benchmark methods are carried out on 10 days test data to evaluate the generalization ability of the learned strategy, including 1) **the original method** without reactive power

control and PV curtailment for voltage regulation; 2) **the SAC algorithm** [23] in a decentralized training manner (SAC-D), where each sub-region is controlled by SAC agents and the agents are trained separately and sequentially based on local observations to minimize the regional voltage deviation. The parameter setting of the SAC-D method is the same as the proposed method; 3) **the stochastic programming method (SP)** [13] based on global information, where 300 scenarios are generated according to the assumed distribution of the predicted PV outputs and load demand. Note that the scenario reduction method is applied to select 20 representative ones; 4) **the MASAC method** [25] implemented in a centralized training and decentralized executing manner. *Note that the SAC-D, MASAC, SAC-C methods use the Z-bus method [28] for the calculation of the immediate reward during training, and therefore the accurate distribution system physical model is*

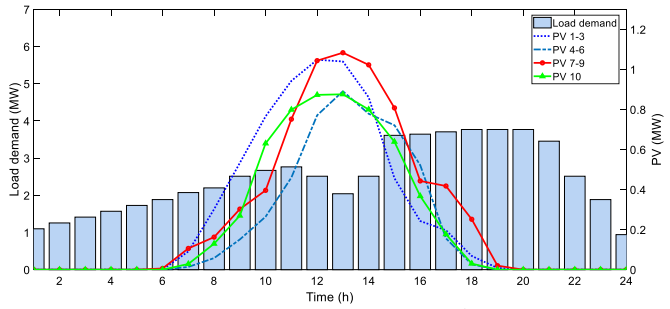

Fig. 4. Load demand and PV generations of the test day.

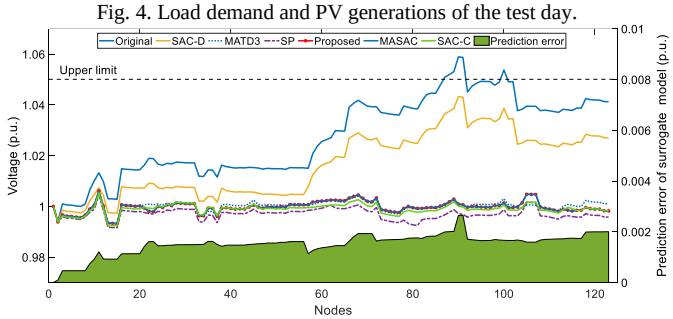

Fig. 5. Voltage distributions for different strategies when *t=12:00*.

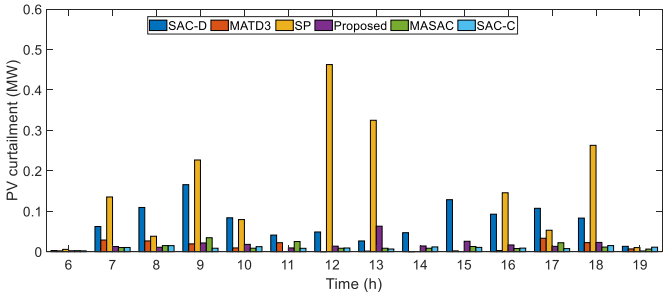

Fig. 6. PV curtailment for different strategies from *6:00* to *19:00* on test day.

required. **In this paper, the perfect physical model for them is assumed while our proposed method is based on the surrogate model learned by the data**. The key idea is that even without the accurate physical model, our proposed MADRL can achieve similar performances or outperform them. The comparison results for all methods are shown in Table VI. For the parameter dependency index, "✓" represents that the corresponding approach relies on the exact physical model when taking control decisions while "×" means model-free. There are two types of control frameworks, where SAC-D, MATD3, proposed, and MASAC methods make decisions according to the information collected from its sub-network; by contrast, SP and SAC-C inform decisions using global information.

It can be observed from the Table VI that when SAC-D agents are trained separately, they do not provide appropriate decisions to regulate the voltage within an acceptable range if only local information is used. This is due to the lack of coordination between agents. By contrast, the centralized SAC-C method fully leverages the coordination and yields much better performance than the SAC-D. The MASAC method can also effectively control the voltage with only local information. Note that during the centralized training and decentralized execution framework, the coordinated control strategy has been learned and this justifies its effectiveness. The performance of the proposed method is only slightly worse than that of the MASAC method but being better than the SP. It should be noted that the proposed method does not need any physical model while MASAC needs an accurate physical power flow model, which is very difficult to achieve in practice. Both the MATD3 and the proposed methods are model-free decentralized approaches. Thanks to the enhanced exploration capability by the entropy term in SAC algorithm, the proposed method achievs a better performance than the MATD3 algorithm. The active power curtailments of PVs for different strategies are also shown in Table VI. Due to the lack of coordination, the SAC-D method curtails more PV generations to reduce the voltage deviations. By contrast, the MATD3, the proposed method, and the MASAC method learn a coordinated strategy during the centralized training, and thus can achieve better voltage regulation performance while curtailing much less PV generations. The SAC-C method can coordinate the reactive power and PV curtailments based on global information, yielding the best voltage regulation performance with least PV curtailment. Since the SP method calculates a pre-determined decision based on sampled scenarios, it tends to be conservative and curtails more PV generations. As a result, the effectiveness of the proposed model-free MADRL is validated.

A sunny day in the test set that suffers from voltage violation issues is also selected from the test set to verify the effectiveness of the proposed method. The load demand and PV generation of the selected day are plotted in Fig. 4. The voltages of all nodes using various control strategies when *t*=12:00 are shown in Fig. 5. We can observe that the voltages at nodes 87-91 and 100 go beyond the upper bound if no reactive control and PV curtailment are applied. When the SAC-D method is used, the voltages return to the allowed range. However, it suffers from large deviation due to the lack of coordination. By contrast, the proposed model-free MADRL, MATD3 and the model-based MASAC with only local measurements can achieve a control performance that is close to that of the centralized SAC-C. Most of the prediction errors for nodes at this moment are less than 0.002 p.u., demonstrating that the surrogate model can provide accurate reward signal to guide the learning of the control strategies. The PV curtailments for different control strategies on this test day are shown in Fig. 6. The proposed MADRL, the MATD3, the MASAC, and the SAC-C can achieve good voltage control performance with little PV curtailments. This justifies their economic benefits.

3) Evaluation of the Generalization Ability of the Learned Strategy

Further tests are carried out under large PV outputs and fluctuations to evaluate the generalization ability of the learned strategy. In particular, a varying PV output profile in 90 seconds due to cloud dynamic is shown in Fig. 7. In the 30th and 90th

second, the active power generations of all PVs reach 92% of their rated power, which is an extreme situation and not seen by the agent during the training. The voltage distributions of node 90 using different voltage control methods are shown in Fig. 8. The controllable devices are operated following pre-determined scheduling for the SP method since it is time-consuming to resolve the optimization problem. The SAC-D, MATD3, MASAC, SAC-C, and the proposed model-free MADRL methods can take decisions in milliseconds since the learned knowledge during training is scalable to newly encountered situations. It can be observed from the figure that if no reactive power control and PV curtailments are applied, node 90 has an over-voltage issue. The voltage comes down when the SAC-D method is applied but its voltage falls out the lower limit and suffers from overshoot issues due to the lack of coordination between controllable devices. Although the SP method can adjust the voltage to the allowed range, it suffers from large fluctuations. The proposed model-free MADRL, the MATD3, MASAC, and the SAC-C method can adjust the scheduling of various control devices based on the real-time observations, therefore yielding much better performance than other methods. This demonstrates that the voltage control strategy learned by the proposed method can generalize to extreme situations with large PV outputs and fluctuations. The SAC-C method achieves the best control performance since accurate global information is assumed, see t=15-50s in Fig. 8 for example. The proposed approach can achieve a performance that is close to the MASAC and the SAC-C method. Again, our method is model-free while the accurate power flow models must be assumed for SAC-C and MASAC method. In summary, our model-free MADRL with only local measurements can achieve comparable performance with those methods that have accurate global information and physical models.

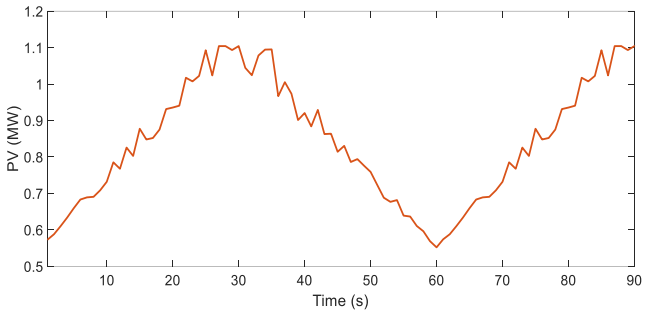

Fig. 7. Varying PV output profiles in 90 seconds.

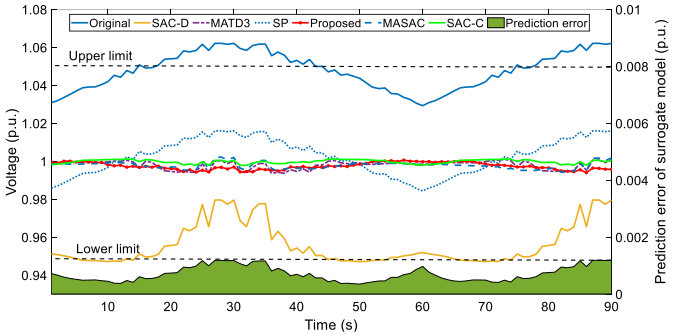

Fig. 8. Voltage distributions of node 90 for different control strategies.

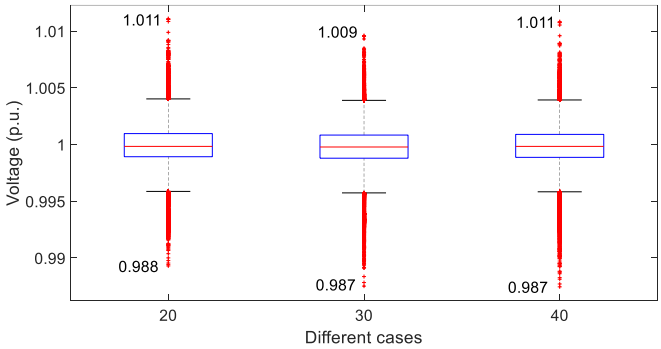

Fig. 9. Voltage distributions of all nodes on test data by the proposed method under different cases.

Further tests are carried out to evaluate the generalization ability of the learned strategy when actual load demand suffers from large deviations from the forecasted values. In this test, the actual load demands of certain number of nodes are assumed to have 20% deviation from the forecasted ones. At each moment, the nodes that suffer from large deviations are randomly selected from all nodes. The voltage distributions of all nodes on 10 days test data achieved by the proposed method are shown in Fig. 9 when the numbers of the randomly selected nodes with large forecasting error are set as 20, 30, and 40. The average voltage deviations on the three cases are 0.16%, 0.17%, and 0.17%, respectively. It can be observed from the figure that the maximum voltage rise and drop under three cases are larger than those obtained by the proposed method under normal conditions, see Table VI. This may due to the fact that the extreme situations are different from the situations seen by the agent during the training. However, the proposed method can adjust the voltage to allowed ranges and achieve average voltage deviations that are close to normal conditions, demonstrating the generalization ability of the learned strategy.

With those comparative results, we can conclude that, 1) the proposed method can learn an strategy to reduce voltage deviation while minimizing the PV curtailment from the training data; 2) the learned strategy can generalize to test data that are unseen by the agent during training; 3) the learned strategy can also deal with extreme situations in the test stage; 4) the proposed model-free decentralized approach can achieve control performance that is close to that by the centralized method with accurate physical model.

C. Tests on 342-nodes Low Voltage Network Test Systems (LVNTS)

Tests are also carried out on 342-node unbalance LVNTS to evaluate the performance of the proposed method. LVNTS is a representative of low voltage distribution systems that are deployed in North America [26]. There are 48 PVs installed in the system. The rated real power and apparent power of PV inverters are 500 kW and 550 kVA, respectively. The whole distribution system is divided into 6 regions according to the responsibility region. Each region is modeled as an DRL agent that is in charge of 8 PVs.

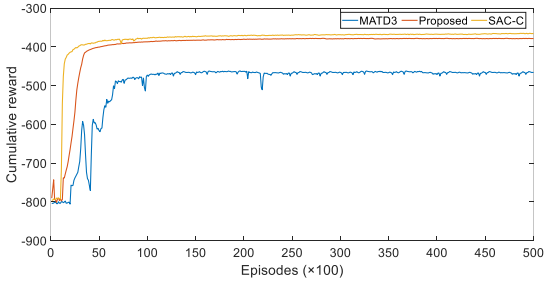

Fig. 10. The evolutions of the cumulative rewards for different methods during the training procedure on 342-node system.

The evolutions of the cumulative rewards of the MATD3, the proposed method, and the SAC-C method during the training process are shown in Fig. 10 when β is set to 1. It can be seen that the proposed MASAC learns a better control strategy than the MATD3 method. This is because the entropy term in SAC algorithm makes it more efficient in exploring the policy space. Since the SAC-C method makes decisions based on global observation, it converges to a higher cumulative reward. This is consistent with what we have observed in Fig. 3.

TABLE VII Comparison results for different strategies

Method	Ave. Dev.	Ave curt. (MW)	Max. rise	Max. drop	Para. dep.
Original	2.40%	-	7.1%	0.39%	-
SAC-D	1.76%	17.01	8.22%	9.04%	✓
MATD3	1.53%	6.52	4.98%	5.94%	×
SP	1.32%	10.02	4.98%	4.99%	✓
Proposed	1.32%	7.41	4.98%	4.43%	×
MASAC	1.32%	7.32	4.98%	3.98%	✓
SAC-C	1.29%	4.11	4.98%	4.48%	✓

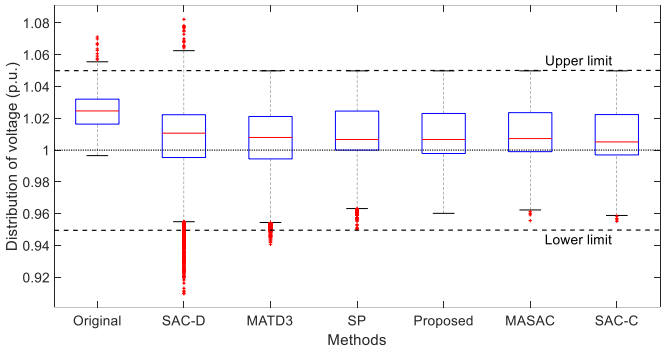

Fig. 11. Voltage distributions under different strategies on this test day.

The comparison results for different control strategies on 10 days test data are listed in Table VII. It can be observed that the maximum voltage rise goes beyond the upper limit when no reactive power control and PV curtailment are applied. When the SAC-D method is utilized, the average voltage deviation can be reduced. However, due to the lack of coordination, both the maximum voltage rise and drop go beyond the allowed ranges. The MATD3 algorithm can further reduce the voltage deviation via the centralized training and decentralized execution framework. However, the voltage constraints are occasionally violated. This is because the huge control space of the large system makes it difficult to find a satisfactory control strategy. By contrast, the entropy term improves the exploration ability of SAC agent, thus improving the control performance of the proposed method. The proposed method can achieve similar control performance with that obtained by the MASAC method, which assumes the knowledge of the accurate physical model of distribution system. The SAC-C method achieves the least voltage deviation among various control strategies. However, it depends on complete two-way communication links and also requires the accurate system model. Since the SAC-D, MATD3, proposed, and MASAC methods inform decisions based on regional information, they choose to curtail more active power of PV to reduce voltage deviation. The SAC-C method curtails the least active power of PV owing to that global information can help it better coordinate the reactive power control and PV curtailment. Compared with the SAC-D method, the centralized training and decentralized execution framework utilized in MATD3, the proposed, and MASAC methods enhance the coordination between agents, thus less PV curtailment of those methods can be observed.

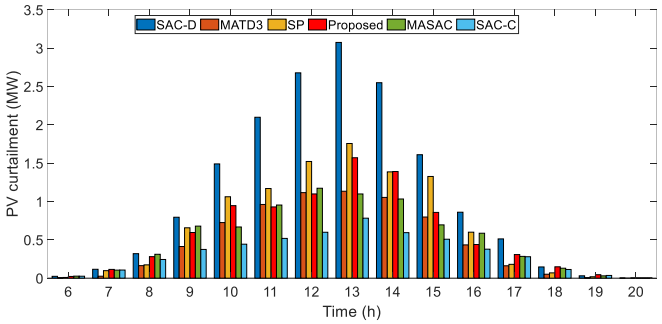

Fig. 12. PV curtailments for different strategies from *6:00* to *20:00* on test day.

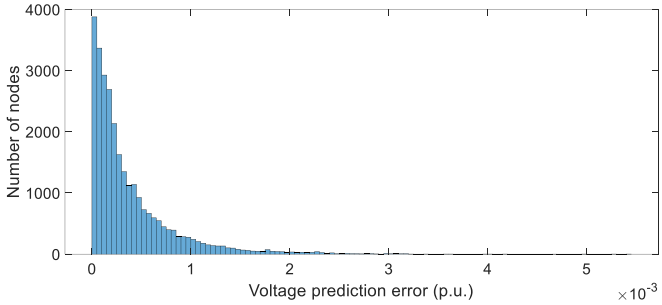

Fig. 13. The distributions of voltage prediction errors achieved by the surrogate model on test day.

To further evaluate the performance of the proposed method, a sunny day in the test set is selected. The voltage distributions for different control strategies on this day are shown in Fig. 11. It can be concluded that the proposed method can adjust the voltage to allowed limit as well as reduce the voltage deviation.

TABLE VIII Comparison results of the proposed method when β is set to different values.

β	Cum. reward	Ave. dev.	Ave. curt. (MW)
0.5	374.8	1.32%	8.21
1	378.1	1.32%	7.41
2	379.9	1.34%	1.83

The control performance by the proposed method is close to the model-based centralized control approach. This is consisent with those shown in Table VII. The PV curtailments for different strategies from *6:00* to *20:00* on this test day are shown in Fig. 12. All methods curtail more active power of PVs to adjust the voltage when the PV generations are high. The SAC-D method curtails the most PV generations owing to the lack of coordination. By contrast, the proposed method curtails less PV generations thanks to the coordination between agents. The SAC-C method curtails the least active power of PVs, which is achieved via the coordination based on the global information. The distributions of voltage prediction errors by the surrogate model are shown in Fig. 13. Most of the prediction errors are distributed in the regions that are less than 3e-3 p.u., demonstrating the effectiveness of the surrogate model.

Tests are also carried out to evaluate the impact of β on the learning performance. The comparative tests by the proposed method are shown in Table VIII when β is set to different values on 342-node system. It can be observed that when we increase β, the proportion of the PV curtailments in the reward increases.

Therefore, the agents learn to curtail less PVs to maximize the reward value, which leads to the increase of average voltage deviations. In practice, the value of β can be selected according to the preference of the operator.

Table IX Parameters of controllable PVs and BSSs

Type	Capacity	Location
BSS	1MWh	45, 58, 79, 114
PV	1.2MW/1.26MVA	11, 22, 35, 53, 62, 68, 90, 100, 105, 118

Table X Parameters settings in the presence of BSSs

Parameter	Value
Batch size for updating NN	32
Replay buffer size	1000000
Temperature parameter	5.0e-3
Discount factor	0.5
Soft update coefficient	0.001
Learning rate for actor/critic-network	0.0003/0.0003
Neuron number of actor networks	128/128
Neuron number of critic networks	128/128/128

TABLE XI Comparison results for different strategies on test data in the presence of BSSs

Method	Ave. Dev.	Ave curt. (MW)	Max. rise	Max. drop	Para. dep.
Original	2.42%	-	5.62%	5.87%	-
SAC-D	2.67%	2.34	6.99%	3.58%	✓
Proposed	0.20%	1.21	1.32%	0.86%	×
MASAC	0.18%	1.39	0.84%	0.84%	✓
SAC-C	0.31%	2.14	0.92%	0.95%	✓

D. Sequential Tests Considering BSSs

Sequential tests are also carried out on the IEEE 123-node systems [26] to evaluate the performance of the proposed method when BSSs and PV inverters are utilized for voltage regulation. The parameters of controllable devices are shown in Table IX. The lower and upper bounds for the energy level of BSS are set as 0.2 and 0.8, respectively. The maximum charging/discharging power limit of BSS is 0.3 MW with the charging/discharging coefficients η_{ch} and η_{dis} as 0.9. For the surrogate model, 1500 instances of data with $\boldsymbol{P}$, $\boldsymbol{Q}$, and $\boldsymbol{V}$ are used for the SPGP algorithm. The setting of the surrogate model is the same as the single-shot control problem. The parameter settings of the MADRL are shown in Table X.

When the training process is completed, comparative tests are carried out on test set to evaluate the performance of the proposed approach. The comparison results on 10 days' test data are shown in Table XI. It can observed that there is voltage violation issue when no control is applied. The SAC-D method also suffers from voltage security issue due to the lack of coordination. The centralized SAC-C method can reduce the voltage deviations and adjust the voltage to allowable range. However, since the control of BSS needs to consider future uncertainties, it is difficult for a single-agent to schedule multiple BSSs simultaneously. By contrast, the proposed MADRL based decentralized control method can achieve better control performance. This demonstrates the superiority of the proposed decentralized method as compared with the centralized ones in the presence of multiple BSSs. The performance of the proposed method is only slightly worse than that obtained by the MASAC method. However, MASAC method depends on the perfect physical model of the DNs, which are difficult to obtain in practice.

A sunny day is selected from the test set to further evaluate the performance of the proposed method. The voltage profiles achieved by different methods at *t=20:00* are shown in Fig. 14. It can be observed that when no control is applied, the voltage violates the lower limit. The SAC-D method suffers from overadjustment due to the lack of coordination. The MADRL methods with the centralized training and decentralized execution framework can effectively reduce the voltage deviation and achieve better performance than the SAC-C method. The MASAC method obtains the best control performance. Again, it requires accurate information of the physical system. The proposed model free method can achieve similar performance as that. The voltage distributions of node 90 for different control strategies are shown in Fig. 15. The results shown here are consisents with those in Table XI.

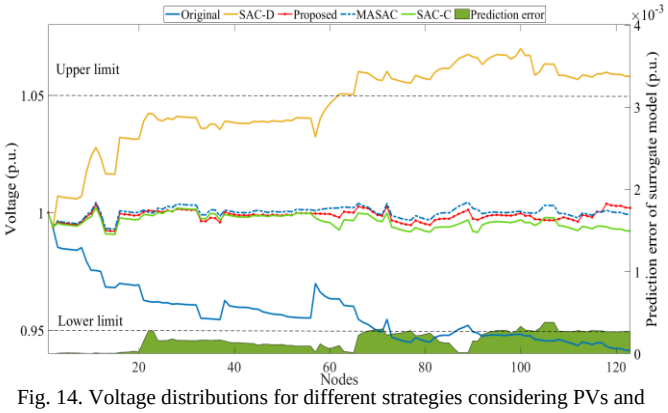

Fig. 14. Voltage distributions for different strategies considering PVs and BSSs when *t=20:00*.

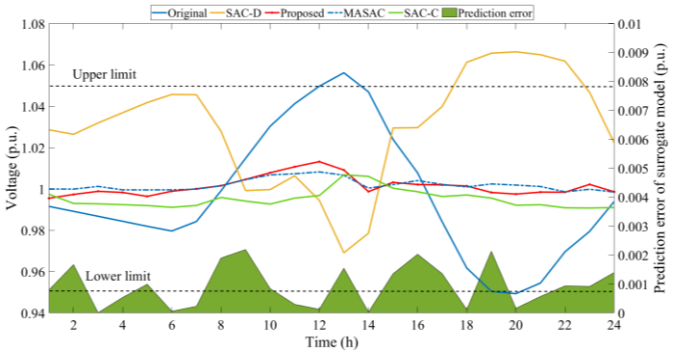

Fig. 15. Voltage distributions of node 90 for different control strategies considering PVs and BSSs.

V. Conclusions

This paper proposes a new model-free centralized training decentralized execution MADRL framework for distribution system voltage regulation with a high penetration of PVs. We first leverage the SPGP to build a surrogate model that learns the mapping relationship between the active and reactive power injections and voltage magnitude of each node using few-shot recorded data. This surrogate model is further integrated with the MADRL to assist the formulation of a coordinated control strategy. In particular, the voltage regulation problem is cast into the MADRL framework by partitioning the whole network to several sub-regions considering the regional voltage regulation ability and the electrical distance. Each sub-region is treated as an agent and solved by the DRL algorithm. All

agents are trained in a centralized framework to learn the coordinated control strategy guided by the reward given by the surrogate model. The proposed method can achieve real-time scheduling using only local information. Comparative results demonstrate that: 1) the proposed decentralized control strategy can achieve close performance as the centralized one; 2) the performance of the proposed model-free approach is similar to that relies on the perfect physical model; 3) the control strategy can be taken in real-time to mitigate the influence of violent PV fluctuations. The future works include 1) developing new method to set adaptive penalty coefficients so as to attain the desired trade-off between reward and constraint cost; 2) extending the proposed framework for networked microgrid control.

REFERENCES

[1] A. Kulmala, S. Repo, and P. Järventausta, "Coordinated voltage control in distribution networks including several distributed energy resources," *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 2010−2020, July 2014.

[2] J. Duan, *et al*., "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 814-817, Jan. 2020.

[3] J. Duan, *et al*., "A deep reinforcement learning based approach for optimal active power dispatch," 2019 IEEE Sustainable Power and Energy Conference (iSPEC), Beijing, China, 2019, pp. 263-267.

[4] S. Karagiannopoulos, P. Aristidou and G. Hug, "Data-Driven local control design for active distribution grids using off-line optimal power flow and machine learning techniques," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6461-6471, Nov. 2019.

[5] H. Xin, Y. Liu, Z. Qu, and D. Gan, "Distributed control and generation estimation method for integrating high-density photovoltaic systems," *IEEE Trans. Energy Conversion*, vol. 29, no. 4, pp. 988–996, Dec. 2014.

[6] M. Zeraati, M. E. Golshan, J. M. Guerrero. "Distributed control of battery energy storage systems for voltage regulation in distribution networks with high PV penetration." *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3582-3593, Jul. 2018.

[7] B. Zhao, Z. Xu, C. Xu, et al. "Network partition-based zonal voltage control for distribution networks with distributed PV systems." *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4087-4098, Sep. 2018.

[8] Y. Chai, L. Guo, C. Wang, et al. "Network partition and voltage coordination control for distribution networks with high penetration of distributed PV units," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3396-3407, May 2018.

[9] P. Li, C. Zhang, Z. Wu, et al. "Distributed adaptive robust voltage/var control with network partition in active distribution networks." *IEEE Trans. Smart Grid*, 2019.

[10] W. Zheng, W. Wu, B. Zhang, et al. "A fully distributed reactive power optimization and control method for active distribution networks", *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 1021-1033, 2016.

[11] R. J. Sánchez-Garcá, M. Fennelly, S. Norris, et al. "Hierarchical spectral clustering of power grids," *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2229-2237, 2014.

[12] E. Cotillasanchez, P. D. Hines, C. Barrows, et al. "Multi-attribute partitioning of power networks based on electrical distance", *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4979-4987, 2013.

[13] Y. Xu, Z. Y. Dong and R. Zhang et al., "Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems," *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4398-4408, Nov. 2017.

[14] G. Wang, V. Kekatos, A. J. Conejo, and G. B. Giannakis, "Ergodic energy management leveraging resource variability in distribution grids," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 4765–4775, Nov. 2016.

[15] W. Wang, N. P. Yu, Y. Q. Gao, and J. Shi, "Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems," *IEEE Trans. Smart Grid*, 2019.

[16] Q. Yang, G. Wang, A. Sadeghi, *et al*., "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2313-2323, May 2020.

[17] M. Al-Saffar, P. Musilek, "Reinforcement learning-based distributed BESS management for mitigating overvoltage issues in systems with high PV penetration," *IEEE Trans. Smart Grid*, vol. 11, no.4, pp. 2980-2994, 2020.

[18] H. Xu, A. Dominguez-Garcia, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 1990-2001, 2020.

[19] D. Cao, W. Hu, J. B. Zhao, Q. Huang, Z. Chen, F. Blaabjerg, "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Trans. Power Syst.*, vol. 35, no. 5, pp. 4120-4123, Sept. 2020.

[20] S. Wang, J. Duan, D. Shi, *et al.*, "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Trans. Power Syst.*, vol. 35, no. 6, pp. 4644-4654, Nov. 2020.

[21] D. Cao, J. B. Zhao, W. Hu, *et al*.., "Attention enabled multi-agent DRL for decentralized volt-VAR control of active distribution system using PV inverters and SVCs" *IEEE Transactions on Sustainable Energy*, early access

[22] E. Snelson, Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs", *Proc of the NIPS 18*, pp. 1257-1264, 2006.

[23] R. Lowe, *et al.*, "Multi-agent actor-critic for mixed cooperative-competitive environments." *Advances in Neural Information Processing Systems*, pp. 6379-6390, 2017.

[24] T. Haarnoja, A. Zhou, P. Abbeel, *et al*., "Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor." *International Conference on Machine Learning*, Stockholm, Sweden, July, 2018.

[25] I. Shariq, F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," *International Conference on Machine Learning*, Long Beach, CA, USA, June, 2019.

[26] IEEE PES, Distribution Test Feeders, Sep. 2010. [Online]. Available: https://site.ieee.org/pes-testfeeders/resources/.

[27] S. Fujimoto, *et al.*, "Addressing function approximation error in actor-critic methods." *arXiv preprint arXiv*:1802.09477, 2018.

[28] M. Bazrafshan, N. Gatsis, "Convergence of the Z-Bus method and existence of unique solution in single-phase distribution load-flow," *Proc. Global Conf. Signal & Information Proc*., Washington, DC, Dec. 2016.