

Multi-Agent Graph-Attention Deep Reinforcement Learning for Post-Contingency Grid Emergency Voltage Control

Y. Zhang, M. Yue

To be published in "IEEE Transactions on Neural Networks and Learning Systems"

January 2024

Interdisciplinary Science Department
Brookhaven National Laboratory

U.S. Department of Energy

USDOE Office of Electricity (OE), Advanced Grid Research & Development. Power Systems
Engineering Research

Notice: This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DE-SC0012704 with the U.S. Department of Energy. The publisher by accepting the manuscript for publication acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or any third party's use or the results of such use of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof or its contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Multi-Agent Graph-Attention Deep Reinforcement Learning for Post-Contingency Grid Emergency Voltage Control

Ying Zhang, *Member, IEEE*, Meng Yue, *Member, IEEE*, Jianhui Wang, *Fellow, IEEE*, Shinjae Yoo

Abstract—Grid emergency voltage control (GEVC) is paramount in electric power systems to improve voltage stability and prevent cascading outages and blackouts in case of contingencies. While most deep reinforcement learning (DRL)-based paradigms perform single agents in a static environment, real-world agents for GEVC are expected to cooperate in a dynamically shifting grid. Moreover, due to high uncertainties from combinatory natures of various contingencies and load consumption, along with the complexity of dynamic grid operation, the data efficiency and control performance of the existing DRL-based methods are challenged. To address these limitations, we propose a multi-agent graph-attention (GATT)-based DRL algorithm for GEVC in multi-area power systems. We develop graph convolutional network (GCN)-based agents for feature representation of the graph-structured voltages to improve the decision accuracy in a data-efficient manner. Furthermore, a cutting-edge attention mechanism concentrates on effective information sharing among multiple agents, synergizing different-sized sub-networks in the grid for cooperative learning. We address several key challenges in the existing DRL-based GEVC approaches, including low scalability and poor stability against high uncertainties. Test results in the IEEE benchmark system verify the advantages of the proposed method over several recent multi-agent DRL-based algorithms.

Index Terms— Multi-agent deep reinforcement learning, grid emergency control, graph convolutional network, attention, voltage control, dynamic power system.

I. INTRODUCTION

BLACKOUT incidents due to voltage instability and voltage collapse in an electric grid have been reported in the last decades, resulting in power losses of thousands of consumers [1]. Grid emergency voltage control (GEVC) serves as a vital solution to reducing the chance of occurrence and impact of power outages and blackouts. GEVC employs corrective actions, e.g., under-voltage load shedding (UVLS),

and optimizes the production or absorption in the grid to prevent the system operation from continuing to deteriorate and bring voltages back to a normal state [2]. However, highly nonlinear dynamics in the grid operation, which are solved by a set of differential-algebraic equations (DAEs) on many dynamic components installed in the grid [3], increase the difficulty of decision-making in GEVC significantly.

Furthermore, effective GEVC strategies are required to guarantee that voltages satisfy a series of time-variant constraints, i.e., the transient voltage recovery criteria (TVRC) [4]. The voltages drop significantly for several seconds once contingencies happen, and the control strategies following the TVRC need to lift these voltages to certain levels at certain moments. Per the TVRC, the post-contingency voltages are required to return to at least 0.7, 0.8, 0.9, and 0.95 levels of the nominal values within 0, 0.33, 0.5, and 1.5 s, respectively. Traditional control schemes adopt the rule-based load-shedding relay and are designed by the utilities to follow the TVRC in offline studies based on some typical operational scenarios [5]. However, in varying operating conditions, the effectiveness and adaptivity of these pre-determined control strategies are not guaranteed [6].

A. Related Works

To provide timely and coordinated control, GEVC is developed by considering the grid model. The solving methodologies can be divided into model-based and model-free categories. In the model-based category, model predictive control [6]–[9] predicts the voltage trajectory and adopts traditional optimization methods to acquire the control solutions. However, the model-based procedure in nonlinear, large-scale dynamic power systems, which consist of diverse dynamic components, is computationally intensive. To meet the requirements of communication speed and privacy, decentralized grid control approaches are further developed [10]–[12]. The decentralized GEVC schemes apply traditional optimization or heuristics techniques and then act on local devices based on local observation, without requiring a central controller. However, for the post-contingency dynamic power system with highly uncertain fault scenarios, the GEVC model complexity makes it challenging to provide an efficient online strategy [13].

Reinforcement learning (RL)-based model-free approaches for adaptive emergency control capture the interest of the latest research [14]–[17]. Different from the traditional GEVC schemes, the learning-based methods can adapt to various fault scenarios and thus can be very potentially implemented online. Early attempts on RL applied temporal difference (TD)

Y. Zhang is with the Department of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK 74078, USA (e-mail: imyinzhang@ieee.org).

M. Yue is with the Interdisciplinary Science Department, Sustainable Energy Division, Brookhaven National Laboratory, Upton, NY 11973, USA (e-mail: yuemeng@bnl.gov). S. Yoo is with Computational Science Initiative, Brookhaven National Laboratory, Upton, NY 11973, USA (e-mail: sjyoo@bnl.gov).

J. Wang is with the Department of Electrical and Computer Engineering, Southern Methodist University, Dallas, TX 75205, USA (e-mail: jianhui@smu.edu).

This work was supported under Agreement 36291 by the Advanced Grid Modeling Program, Office of Electricity of the U.S. Department of Energy.

learning [18] and Q-learning methods [19], [20] to adaptive emergency control. To improve the learning efficiency in high-dimensional state spaces, deep RL (DRL) exploits neural networks (NNs) to represent the complex nonlinear function of the control policy. For example, [21] and [22] develop deep Q-network (DQN)-based algorithms and provide discrete actions for load shedding to stabilize the voltage profile after a fault happens. To accelerate the learning procedure in DQN, [23] develops the parallel augmented random search technique to further the scalability of the Q-learning. From another perspective, the authors of [21] then addressed an overestimation bias of the Q-learning by double DQN (DDQN), which is proposed in [24]. However, the poor sample efficiency issue, which [25] points out as a “dominant concern in RL”, is not addressed in the above Q-learning-based algorithms.

Recent advances in electric power system control are towards an adaptive intelligent system with the advent of multi-agent RL (MARL) techniques. In a multi-area power system, multiple agents can be assigned to the subnetworks based on empirical practice of system operators, and each of them makes decisions for the corresponding subnetwork according to observations in the subnetwork [26]. To this end, multi-agent DQN [27], deep deterministic policy gradient (DDPG) [28], twin delayed deep deterministic policy gradient (TD3) [29], and soft-actor-critic (SAC) [30] are explored in decentralized grid optimization and control applications. It should be noted that these applications fall into the category of static power system control without considering the operation of dynamic components in the grid. For example, [30] proposes a multi-agent SAC algorithm for static voltage regulation by centralized training in critics and decentralized conduction of each actor. In contrast to the DRL applications in static power system control [27]–[32], there remain gaps in the related MARL research for dynamic grid emergency control. Three main challenges lie in the MARL-based GEVC schemes, which have not been fully addressed, as summarized below.

1) Due to the existence of various dynamic components in the grid, voltages on all buses are fast time-variant after a contingency happens. Also, control actions applied to the grid further intensify the voltage dynamics. As a result, complex spatiotemporal features exist in the nodal voltage observations from the multi-area power system. Most of the existing DRL-based works adopt fully connected networks (FCNs) such as [21], [22], [26]–[30], [33] as agents. Nevertheless, they fail to exploit the topological structures in the multi-area power grid, resulting in low data efficiency and generalization for the grid control problem with different subnetwork topologies.

2) Dynamic power system operation in practice is highly uncertain since it combines various uncertainty factors, such as unknown fault types, fault durations, fault locations, and time-varying load consumption [20], [21]. The high uncertainties make it very difficult for the control policies to obtain stable adaptive learning performance.

3) Despite the growing research efforts in the field of multi-agent systems (MAS) control, achieving cooperative multi-agent GEVC in power systems remains open [32]. In parallel with the MARL applications in power system control, various

control methods for MASs are investigated, such as leader-to-formation stability (LFS) [34]–[36] and game theory [37], [38], to ensure the coordination and stability of the optimal control among the agents. For instance, LFS is introduced to investigate the stability of control policy to guarantee that the closed-loop system is leader-to-formation stable, through RL or adaptive dynamic programming (ADP) [36]. However, due to the sensitivity of dynamic power system operation represented by large-scale DAEs to control variables, the interaction of multiple agents demands data-efficient solutions for cooperative learning. Control policies provided by independent Q learning (IQL) are easily stuck in the local optimality [39].

B. Contributions

To address the above challenges, this paper proposes a novel graph-attention (GATT)-based MARL algorithm for decentralized GEVC in multi-area power systems. In the proposed algorithm, a novel soft actor-attention-critic structure is adopted to provide control actions. We develop graph convolutional network (GCN)-powered agents to capture the feature representation of the graph-structured states. By embedding topological information of each subnetwork in the agents, GCN [40] helps the agents understand the subgraph structure and accelerate the learning process. Moreover, an attention mechanism is adopted for the interactions among the agents to select more relevant information from all the subnetworks. The powerful combination of the GCN model and the attention mechanism enables stable and cooperative learning for the dynamic control problem. All agents are trained offline to learn the cooperative control strategy against randomly generated scenarios with the abovementioned high uncertainties. After that, the online decentralized control that only receives the local states can be achieved. The main contributions are listed below.

- The proposed method is data-efficient by simultaneously concentrating the spatial topology connection and temporal features of the agent inputs to make adaptive decisions. As a result, the required number of interactions with the complex power system environment in the training procedure decreases.
- The scalability and sampling efficiency of the proposed method are guaranteed jointly by the state-of-the-art attention-based MARL structure and an off-policy SAC training paradigm.
- Comparative studies with several other MARL algorithms illustrate the outstanding control performance and largely improved learning stability of the proposed method in diverse dynamic operating scenarios with high uncertainties.

C. Paper Organization

The rest of the paper is organized as follows: Section II introduces TVRC-represented voltage recovery requirements and problem formulation for GEVC in dynamic power system operation. Section III presents the proposed GATT-based MARL algorithm, followed by the case study shown in Section IV. The conclusion is provided in Section V.

II. GEVC PROBLEM FORMULATION

A power grid is a highly nonlinear dynamic system, and various dynamic components, such as synchronous machines,

excitation systems, and turbine governors, are installed at generator nodes (buses). The dynamic system is modeled as DAEs, which depict the dynamic component operation and network connection among the generator and load nodes:

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{y}(t)) \quad (1)$$

$$0 = g(\mathbf{x}(t), \mathbf{y}(t)) \quad (2)$$

where $\mathbf{x}(t)$ is the dynamic state vector; $\mathbf{y}(t)$ denotes an algebraic state vector, typically all the nodal voltages in the grid [9]; $f(\cdot)$ denotes various differential equations (DEs) for dynamic components, such as generators, thermal turbine governors, static exciters, etc.; $g(\cdot)$ denotes algebraic equations (AEs) in the grid. For example, the fourth-order DEs of a synchronous generator is written as [3]

$$\begin{cases} \frac{dw_k}{dt} = w_k - w_s \\ \frac{d\delta_k}{dt} = \frac{P_m - P_e - D_k(w_k - w_s)}{M_k} \\ \frac{dE'_{qk}}{dt} = \frac{(-E'_{qk} - E_{fdk})}{T'_{do_k}} \\ \frac{dE_{fdk}}{dt} = \frac{-E_{fdk} - K_a(V_{ref} - V_k)}{T_{a_k}} \end{cases} \quad (3)$$

where E_{fdk} and E'_{qk} denote the equivalent excitation voltage and the internal voltage corresponding to generator k , $k = \{1, 2, \dots, \mathcal{K}\}$, respectively; δ_k and w_k denote the generator rotor angle and angular velocity; T'_{do_k} is the time constant of excitation circuits, and T_{a_k} is the regulator time constant; K_a is the regulator gain, and V_{ref} and w_s are the reference voltage and angular speed in per unit; M_k and D_k denote the machine parameters; P_m and P_e are the mechanical input and electrical output powers.

The dynamic state variables in the dynamic power system operation are expressed as

$$\mathbf{x} = \{\mathbf{w}, \boldsymbol{\delta}, \mathbf{E}'_q, \mathbf{E}_{fd}\} \quad (4)$$

where \mathbf{w} and $\boldsymbol{\delta}$ are the vectors of angular speeds and rotor angles for all the generators; and \mathbf{E}_{fd} denotes the internal voltage behind the transient reactance of the direct axis; \mathbf{E}'_q is the quadrature-axis induced emf.

These dynamic components installed at the generator buses are linked via the power transmission network to supply customer loads on all the load buses of the grid. The algebraic equation (2) for the network operation is specifically formulated as [8]:

$$\begin{bmatrix} \mathbf{G}(t) & -\mathbf{B}(t) \\ \mathbf{B}(t) & \mathbf{G}(t) \end{bmatrix} \begin{bmatrix} \mathbf{V}_x(t) \\ \mathbf{V}_y(t) \end{bmatrix} = \begin{bmatrix} \mathbf{I}_x(t) \\ \mathbf{I}_y(t) \end{bmatrix} \quad (5)$$

where $V_{xk}(t) + jV_{yk}(t)$ is the complex voltage on bus k , and $I_{xk}(t) + jI_{yk}(t)$ denotes the current injection into bus k from generators and loads; \mathbf{G} and \mathbf{B} denote the nodal conductance and susceptance matrices in the transmission lines.

After the occurrence of a contingency, i.e., a fault, GEVC provides corrective control to prevent the system voltage collapse. Furthermore, the GEVC problem is modeled as a nonlinear and non-convex optimization to solve control variables in the above dynamic power system. These control

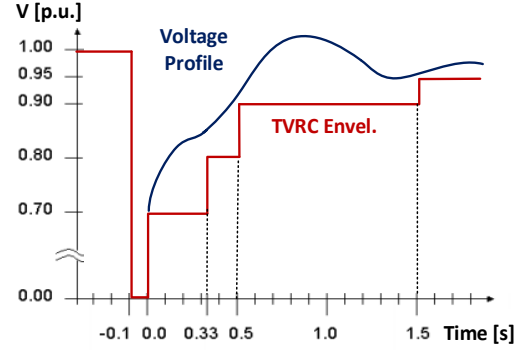


Fig.1. Transient voltage recovery criteria in post-fault power systems [4]

variables are defined as $\mathbf{u}(t)$, and $\mathbf{u}(t) = \{u_i(t)\}$ and $i = \{1, 2, \dots, N_c\}$ are controllable nodes. For $t \in [t_0, t_0 + T]$, the time-series GEVC model can be formulated as [8]:

$$\min_{\mathbf{u}} \int_{t_0}^{t_0+T} C(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t)) dt \quad (6)$$

$$\begin{cases} \dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t)) & (6a) \\ 0 = g(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t)) & (6b) \\ \underline{\mathbf{S}} \leq H(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t)) \leq \bar{\mathbf{S}} & (6c) \\ \underline{\mathbf{u}} \leq \mathbf{u}(t) \leq \bar{\mathbf{u}} & (6d) \end{cases}$$

where $\underline{\mathbf{S}}$ and $\bar{\mathbf{S}}$ denote the bounds of the operating requirements about $\mathbf{x}(t)$ or $\mathbf{y}(t)$; the constraints of the control variables are depicted in (6d), and the maximum of control magnitudes that the controller can implement is usually predetermined by the utilities [2]; $C(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t))$ denotes the cost function.

In this paper, we set the ratio of the load shedding at the controllable load buses as the control variables [21], [23]. Moreover, the cost function at each control conduction time t aims to calculate the accumulated load shedding in all the N_c control buses and can be expressed as:

$$C(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t)) = \sum_{i=1}^{N_c} u_{Li}(t) P_{Li} \quad (7)$$

where P_{Li} denotes the initial load on the i th load bus before the load shedding.

GEVC should follow the TVRC to design the time-series control strategy [4], [21]–[24]. A typical TVRC envelope curve is shown in Fig.1. To satisfy these TVRC criteria, the operating requirements of voltage levels are constrained in (6c) during the time horizon. These voltage constraints can be expressed in the form of the deviations of voltage magnitude vector from the TVRC as

$$\Delta \mathbf{V}(t) \geq 0 \quad (8)$$

where the deviations of voltage magnitudes are calculated by

$$\Delta \mathbf{V}(t) = \begin{cases} \mathbf{V}(t) - 0.7 & t \in [T_{fc}, T_{fc} + 0.33] \\ \mathbf{V}(t) - 0.8 & t \in [T_{fc} + 0.33, T_{fc} + 0.5] \\ \mathbf{V}(t) - 0.9 & t \in [T_{fc} + 0.5, T_{fc} + 1.5] \\ \mathbf{V}(t) - 0.95 & t \in [T_{fc} + 1.5, t_0 + T] \end{cases} \quad (9)$$

where T_{fc} represents the time moment of clearing a fault.

III. PROPOSED MARL ALGORITHM

This section proposes a multi-agent GATT-based DRL algorithm for GEVC in multi-area dynamic power systems. We develop a novel soft actor-attention-critic MARL architecture. After assigning agents to the multi-area power grid, each agent is modeled as GCN-powered actor and attention-critic networks. The graph information is captured by the GCN-powered actor to extract effective topological information in the voltage observation of each subarea. To learn effectively in multi-agent environments, a centrally computed attention mechanism for interaction among the agents helps the critic of each agent to selectively pay attention to information from other agents for global policy evaluation. Finally, the proposed method realizes cooperative learning by a Q-learning objective function covering a maximum-entropy term.

A. Markov Games of Multiple Agents

The decentralized GEVC problem is described as each localized controller (agent) making decisions on the controllable buses at each action time step to recover the voltage trajectories above the TVRC constraints. The multi-agent GEVC problem can be characterized as a Markov game, which is a multi-agent extension of Markov decision process (MDP). A Markov game for N_a agents can be defined by a set of states \mathcal{S} , the action for each agent \mathcal{A}^i , and the state for each agent \mathcal{S}^i , $i = 1, 2, \dots, N_a$. To determine the actions, agent i uses a stochastic policy $\pi_i: \mathcal{S}^i \rightarrow P(\mathcal{A}^i)$, and then this process produces the next state according to a transition probability function $\Gamma: \mathcal{S} \times \mathcal{A}^1 \times \dots \times \mathcal{A}^N \rightarrow P(\mathcal{S})$. During the state transition, each agent computes its reward r^i , which is a function of the state and agent's action to evaluate the policy efficacy. Moreover, agent i aims to maximize its cumulative reward within the time horizon T , expressed as

$$\max_{\pi} R_i(\pi_i) = \sum_{t=0}^T \gamma r_i^t \quad (10)$$

where $\gamma \in [0,1]$ denotes the discount factor, and r_i^t denotes the reward at the time step t for agent i .

The time-series decision in the MDP can be performed in DRL by training with the historical dataset, see several DRL-based power system control methods [20]-[23], where the NN agents learn the control policies via observations and interactions with the environment. In this paper, the Markov game for the GEVC problem is implemented in MARL by interpreting the voltages, the control variables, and the objective function in the optimization model (6) into the state, actions, and reward for each agent. Meanwhile, multi-area post-contingency dynamic power system operation acts as the MARL environment. The detailed designs are provided below:

1) **Action:** Without loss of generality, agent i manages all n_i controllable nodes in the assigned subnetwork. The

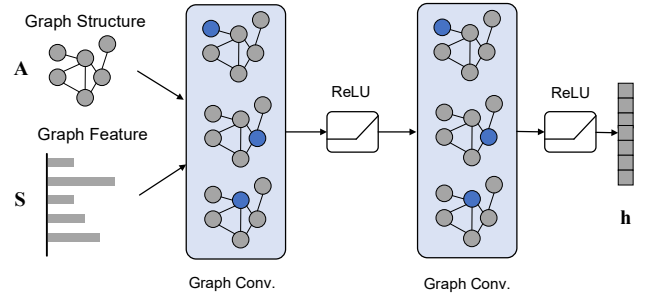


Fig.2. The structure of GCN.

decision variables in the GEVC problem are represented by discrete quantities [2]. Moreover, the integer actions are decided to shed ($a \neq 0$) or not shed ($a = 0$) customer loads on the k th controllable node, $k = \{1, 2, \dots, n_i\}$. At action time step t , the control variables in agent i are expressed as

$$\mathbf{a}_i(t) = [u_{L1}(t), u_{L2}(t), \dots] \quad (11)$$

where $u_{Lk} \in \{0, \beta\%$ is the discrete action at the k th controllable load bus at t [23]; β as the ratio of load shedding decided to shed on a controllable node.

2) **State:** The voltages at all the buses of subnetwork i are denoted at t as $\mathbf{O}_t^i = \{\Delta V_j(t)\}$ with $j \in \{1, 2, \dots, N_i\}$, where N_i is the number of the buses in this subnetwork. Moreover, the state vector for agent i is expressed as:

$$\mathbf{s}_t^i = [\mathbf{O}_{t-N_r-1}^i, \dots, \mathbf{O}_{t-1}^i] \quad (12)$$

where $\mathbf{O}_\tau^i \in \mathbb{R}^{N_i \times 1}$ is the observation set at the previous observation time step τ , $\tau \in \{t - \mathcal{T} - 1, \dots, t - 1\}$, and \mathcal{T} is the number of the temporal features. Note that \mathbf{s}_t^i consists of the topological information about the voltages in this subnetwork, which should be extracted for more data-efficient decision-making.

3) **Reward Function:** After conducting the control actions \mathbf{a}_i , the reward at time t in agent i is calculated as:

$$r_i^t = \begin{cases} \sum_{j=1}^{N_i} \min \{\Delta V_j(t), 0\} & \text{If } \forall j, \Delta V_j(t) < 0 \\ c \sum_{k=1}^{n_i} P_{Lk} (1 - u_{Lk}(t)) / P_{total} & \text{Otherwise} \end{cases} \quad (13)$$

where the voltage deviation $\Delta V_j(t)$ is calculated by (9); P_{total} is the total number of all the controllable loads in agent i , and the remaining loads are computed in terms of percentages of P_{total} , and P_{Lk} is the initial load before taking all the actions at node k . Therefore, the high positive reward points to a solution with a relatively large total amount of the remaining loads and meanwhile satisfying TVRC constraints.

B. GCN for Graph Feature Representation

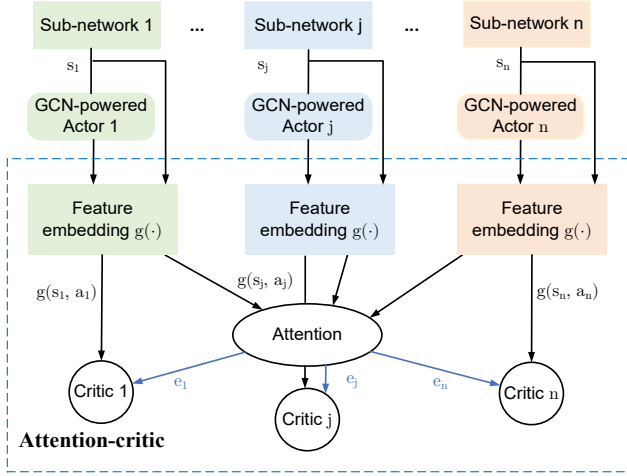


Fig.3. The structure of the proposed algorithm.

Voltage states in a multi-area power grid contain essential topological information of different subnetworks. However, the existing MARL algorithms for grid control usually use FCNs to process a one-dimensional state from different topologies without exploring the topological information embedded in the nodal voltages. In contrast, a GCN is capable of learning these latent node-level features from the graph-structured voltage states. Moreover, the topological information captured from the graph-structured voltages can accelerate the policy learning process of agents. Therefore, in the proposed DRL method, we adopt the GCN model for feature representation in each subnetwork for decision-making in GEVC.

GCN, which is developed from the spectral graph theory, is widely used in the graph signal processing domain [40]. Specifically, the GCN model generates feature representations using the adjacent matrix of each network topology, making it easier to capture the topological information. Define a graph as $G(\mathcal{V}, \mathcal{E})$ in the node set \mathcal{V} , and \mathcal{E} denotes the edge set. The adjacency matrix of the graph is denoted as $\mathbf{A} = \{A_{ij}\}$, and $i, j = 1, \dots, N$.

The GCN model adopts a classic two-layer structure by stacking convolutional layers [40], shown in Fig.2. Denote the output of GCN as $\mathbf{h} \in \mathbb{R}^{N \times F}$, and F is the number of the output feature at each node. The two-layer GCN operation can be expressed as:

$$\mathbf{h} = \text{GCN}(\mathbf{S}, \mathbf{A}) = \sigma(\tilde{\mathbf{A}}\text{ReLU}(\tilde{\mathbf{A}}\mathbf{S}\mathbf{W}^{(0)}))\mathbf{W}^{(1)} \quad (14)$$

where $\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}\hat{\mathbf{A}}\mathbf{D}^{-1/2}$, and $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, \mathbf{I} denotes an identity matrix with the same dimension as \mathbf{A} ; \mathbf{D} is the degree matrix of the graph, where the (i, i) element is derived by $\hat{D}_{ii} = \sum_{j=1}^n \hat{A}_{ij}$; $\mathbf{W}^{(0)}$ and $\mathbf{W}^{(1)}$ are the weight matrices of the first and second layers, respectively; $\sigma(\cdot)$ denotes a nonlinear activation function, and $\text{ReLU}(\cdot)$ is the rectified linear unit.

For the GEVC problem, the GCN deals with the graph features of the temporal voltages in \mathcal{T} historical observations

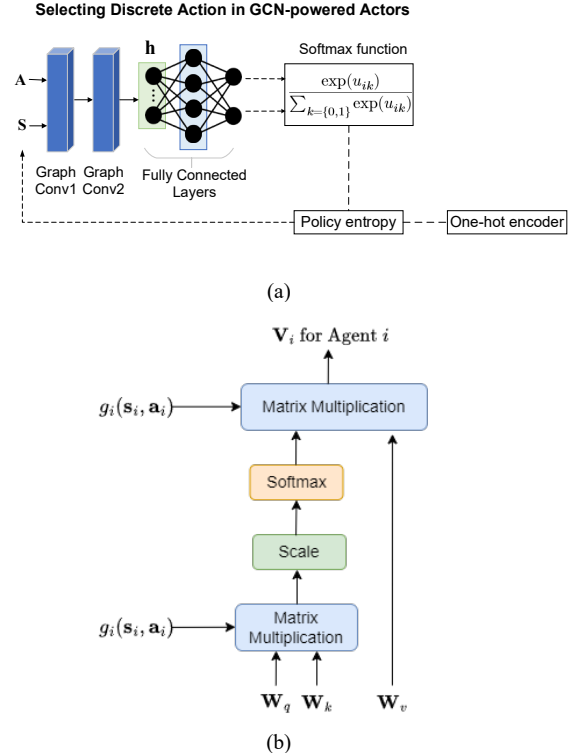


Fig.4. a) GCN-powered actor NN and b) The architecture of the adopted attention mechanism

in the corresponding power network. Moreover, in the proposed MARL method, the GCN model can be trained together with the actor-critic model in an end-to-end manner, which is efficient in learning the control policy. We will elaborate on their training procedure in the next subsections.

C. Graph-Attention-Based Multi-Agent Deep Reinforcement Learning Algorithm

The architecture of the proposed GATT-based MARL method is shown in Fig.3. Each agent in this algorithm consists of two modules: GCN-powered actor networks and attention-critic networks. After embedding the GCN model into the actor network, each actor has its own state representations that can be further leveraged in Q-learning to predict the control policy. Moreover, the attention-critic network helps these agents efficiently learn and evaluate the policy for cooperative learning via a central attention mechanism.

In each agent, the critic is approximated by two types of NNs, including a state value function V parameterized by ψ and a soft Q-function Q parameterized by θ for policy evaluation, while the actor is expressed as a policy function π parameterized by ϕ . For example, in agent i , $\pi_{\phi}^i(\cdot | s_i)$ is the GCN-powered policy based on its individual current observation s_i . For clarity, the subscript t for each time step in the related terminology is omitted. Next, we describe the actor and critic networks in detail.

The Q-learning objective in the proposed method maximizes the expected reward return and the entropy of the policy at the current observation. To this end, a soft Bellman

equation is adopted to integrate the policy entropy $\mathcal{H}(\cdot | \mathbf{s}) = -\alpha \log(\pi(\cdot | \mathbf{s}))$, expressed as [41]

$$Q_i(\mathbf{s}_i, \mathbf{a}_i) = r(\mathbf{s}_i, \mathbf{a}_i) + \gamma \mathbb{E}_{(\mathbf{s}_i, \mathbf{a}_i) \sim \rho_\pi} [Q_i(\mathbf{s}'_i, \mathbf{a}'_i) - \alpha \log(\pi(\mathbf{a}_i | \mathbf{s}_i))] \quad (15)$$

where γ denotes a discount parameter, and α is a temperature parameter for balancing between the reward and the policy entropy.

1) GCN-Powered Actor

Most of the existing DRL-based control methods, such as [21], [22], [26]–[30], [33], stacking previous time-series voltages into a state vector, lack the capability of interpreting the graphic features of the states. To capture the topological representation of the graph-structured voltage observations in the power grid, we adopt the GCN-powered actors to provide control actions for GEVC.

We link the topology information of voltages on all the nodes of a subnetwork into the agent state. For agent i , the input to the GCN is $\mathbf{S}_i \in \mathbb{R}^{N_i \times \mathcal{T}}$ and $\mathbf{A}_i \in \mathbb{R}^{N_i \times N_i}$, where N_i is the number of nodes in subnetwork i . The actor for agent i learns the policy, and the control actions are its output. Moreover, the policy function π of actor i is approximated by a GCN-powered neural network as follows:

$$\mathbf{h}_i = \text{GCN}(\mathbf{S}_i, \mathbf{A}_i) \quad (16)$$

$$\mathbf{a}_i = p_\phi^i(\mathbf{s}_i) = p_{i,L}(\dots p_{i,1}(\mathbf{h}_i)) \quad (17)$$

$$p_{i,l}(\mathbf{o}_{l-1}) = \sigma(\mathbf{W}_l^a * \mathbf{o}_{l-1} + \mathbf{b}_l^a), l = 2, 3, \dots, L \quad (18)$$

where $\sigma(\cdot)$ is the ReLU activation function, and $p_{i,l}(\cdot)$ denotes the mapping of the l th layer of the NN, with \mathbf{o}_{l-1} as the output on the $(l-1)$ th layer; \mathbf{W}_l^a and \mathbf{b}_l^a denote the NN weights and bias.

The original actor networks are not applicable to the discrete action settings. However, the settings in the GEVC problem involve discrete actions. To obtain the discrete actions, we modify the actor to attach a softmax function to estimate the probability of each feasible action, shown in Fig.4(a). The softmax function is employed in the last layer of the actor NNs and converts a vector of action points into the probability distribution of the K discrete actions. For each controllable node, the probability of the k th action choice on each controllable load bus is calculated as

$$P_k = \exp(u_k) / \sum_k \exp(u_k) \quad (19)$$

where $k = 1, 2, \dots, K$.

2) Attention-Critic Networks

An attention mechanism [42] is originally proposed in natural language processing and consists of an attention NN that calculates attention weights for each part of the input to enhance the precision of the output. In the proposed MARL method, an attention-critic NN is proposed to share weighted information among agents, which integrates a modified attention network into the critic Q function in DRL. By enabling the critic network to concentrate on the most pertinent information across all subnetworks, irrespective of the information's location or magnitude within the input, this new

critic structure facilitates dynamic focus. As a result, the proposed attention-critic NN is capable of weighing the contributions of other agents for cooperative control, providing a notable advantage over previous approaches, e.g., [27], [28], that did not prioritize or differentiate the information from each agent. Moreover, the proposed attention-critics linearly increase input space with respect to the number of agents to enable more scalable learning. As in Fig. 3, besides the local observations and actions, the impacts of other agents are captured by the attention system, and the agents' features are concatenated, which is a weighted sum of encodings from other agents, \mathbf{e}_i , as another input to the Q network.

The proposed attention-critic Q function for agent i is denoted as $Q_i(\mathbf{s}_i, \mathbf{a}_i)$. The Q function adopts a NN with L fully connected layers to approximate the critic function for the policy evaluation, and $Q_i(\mathbf{s}_i, \mathbf{a}_i)$ is represented by:

$$Q_i(\mathbf{s}_i, \mathbf{a}_i) = f_i(g_i(\mathbf{s}_i, \mathbf{a}_i), \mathbf{e}_i) \quad (20)$$

$$f_{i,l}(\mathbf{o}_{l-1}) = \sigma(\mathbf{W}_l^c * \mathbf{o}_{l-1} + \mathbf{b}_l^c), l = 2, 3, \dots, L \quad (21)$$

where $f_{i,l}(\cdot)$ denotes the mapping of the l th layer of the NN, and \mathbf{o}_{l-1} is the output on the $(l-1)$ th layer; $g_i(\mathbf{s}_i, \mathbf{a}_i)$ is a multi-layer perceptron embedding function, and it allows the parameter sharing to learn effectively in the multi-agent environment where the size of state and action for individual agents might be different; \mathbf{e}_i denotes a weighted sum of the Q function outputs from other agents, indicating their contribution to agent i , and next we will introduce \mathbf{e}_i in detail.

The weighted sum of the Q function outputs from other agents \mathbf{e}_i is obtained from a classic *attention* system, illustrated as Fig.4(b). Given a query \mathbf{q} and a set of key-value pairs (\mathbf{K}, \mathbf{V}) , the attention system aims to compute a weighted sum of the values dependent on the query and the corresponding keys [42]. Moreover, a query, a key, and a value of the attention in each critic are computed by multiplying the embedded observation by their corresponding transformation matrices. These matrices in the attention model are denoted as $\{\mathbf{W}_k, \mathbf{W}_q, \mathbf{W}_v\}$. In the case that multiple attention heads are used, each head adopts a separate set of these parameters and aggregates the contributions from all other agents to agent i . Moreover, the contributions from all the heads are concatenated as a single vector.

Let \mathbf{V}_i denote the key in agent i and is a function of $g_i(\mathbf{s}_i, \mathbf{a}_i)$ and then linearly transformed by a shared matrix \mathbf{W}_v [37]:

$$\mathbf{V}_i = \mathbf{W}_v g_i(\mathbf{s}_i, \mathbf{a}_i) \quad (22)$$

The weighted sum of the Q function outputs from other agents \mathbf{e}_j is calculated as:

$$\mathbf{e}_i = \sum_{j \in \{1, 2, \dots, N_a\} \setminus i} \alpha_j \text{LReLU}(\mathbf{V}_j) \quad (23)$$

where a leaky ReLU activation function $\text{LReLU}(\cdot)$ is added for improving the efficiency of the attention mechanism; α_j is obtained by comparing the similarity between the embedding

Algorithm 1 Offline Training of GATT

```

1  Inputs: learning rates and smoothing parameter  $\tau$ 
2  Initialize actor and critic networks for each agent,
   and memory buffer.
   for each epoch do
3     $t = 0$ : set random fault scenario and load
       levels and initialize power system dynamics.
       While  $t <$  total simulation time
         for each control step do
4           Obtain states for each agent.
5           Get actions from GCN-powered actor
              policy:  $\mathbf{a}_t \sim \pi_{GCN}(\mathbf{a}_t | \mathbf{s}_t)$ 
         end for
6         Apply the actions to the grid and update
           the loads.
7         Based on the current operation status and
           loads, update grid dynamic at  $t$ .
8         Collect next states  $\mathbf{s}_{t+1}$  and reward of all
           the agents. Store the experience into  $\mathcal{D}_i$ .
9          $t = t + 1$ 
       end while
10        Update each attention-critic by (27) and (28).
11        Update each GCN-powered actor by (31).
12        Update the parameters of each target critic:

```

$$\begin{aligned}\bar{\psi} &\leftarrow \tau\psi + (1 - \tau)\bar{\psi} \\ \bar{\theta} &\leftarrow \tau\theta + (1 - \tau)\bar{\theta}\end{aligned}$$

end for

of agent i and other agent j , and a softmax function is adopted to quantify this similarity on the value:

$$\alpha_j \propto \exp(g_j(\mathbf{s}_j, \mathbf{a}_j)^T \mathbf{W}_k^T \mathbf{W}_q g_i(\mathbf{s}_i, \mathbf{a}_i)) \quad (24)$$

D. Training Process and Decentralized Online Control

The proposed control algorithm consists of central offline training and decentralized online implementation. During the training, the central attention mechanism receives information from all the agents and extracts the attentive latent information to the critic of each agent. The offline training process of the proposed method is summarized in Algorithm 1. Moreover, the central DRL training is implemented offline, and the offline attention-based interaction allows each actor to provide cooperative control strategies to multi-area power system dynamics based on its local states. This results in efficient decentralized online inference for the GEVC problem.

The replay buffer technique is adopted for efficient training in the DRL process. Each agent stores the experiences in a buffer tuple \mathcal{D}_i , denoted as $\{\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}'_i, r_i\}$, which denote the states, actions, future states, and reward, respectively. Mini-batch experience for each agent is randomly sampled from \mathcal{D}_i .

1) Updating Critic Networks

The attention-critic parameters in all the critics are denoted as $\theta = \{\theta_1, \theta_2, \dots, \theta_{N_a}\}$, each of which is from the NN

defined by (20). These parameters are updated together by minimizing the following joint regression loss function with the information sharing:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N_a} \mathbb{E}_{(\mathbf{s}, \mathbf{a}, r, \mathbf{s}') \sim \mathcal{D}} \left[\left(Q_i(\mathbf{s}_i, \mathbf{a}_i) - \hat{Q}_i(\mathbf{s}_i, \mathbf{a}_i) \right)^2 \right] \quad (25)$$

where we calculate the soft Bellman squared residual for all the $(\mathbf{s}_i, \mathbf{a}_i)$ pairs in the sampled replay buffer from \mathcal{D}_i ; the target Q -function $\hat{Q}_i(\mathbf{s}_i, \mathbf{a}_i) = r(\mathbf{s}_i, \mathbf{a}_i) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p} [V_{\bar{\psi}}(\mathbf{s}'_i)]$, and $V_{\bar{\psi}}(\mathbf{s}'_i)$ is the output of the target value network; $\bar{\psi}$ represents the weights in the target value networks.

The value network V is trained by minimizing the following loss function:

$$\mathcal{L}_V(\psi) = \mathbb{E}_{\mathbf{s}_i \sim \mathcal{D}_i} \left[\frac{1}{2} (r(\mathbf{s}_i, \mathbf{a}_i) - \mathbb{E}_{\mathbf{a}_i \sim \pi(\phi)} [Q_i(\mathbf{s}_i, \mathbf{a}_i) - \alpha \log(\pi(\mathbf{a}_i | \mathbf{s}_i))])^2 \right] \quad (26)$$

The network parameters of these critics are updated by the gradient calculation on the loss functions (25) and (26):

$$\theta_i \leftarrow \theta_i - \lambda_Q \widehat{\nabla}_{\theta_i} \mathcal{L}_Q(\theta_i) \quad (27)$$

$$\psi_i \leftarrow \psi_i - \lambda_V \widehat{\nabla}_{\psi_i} \mathcal{L}_V(\psi_i) \quad (28)$$

where λ_Q and λ_V are learning rates.

The parameters of the target critics, $\bar{\theta}$ and $\bar{\psi}$, are subsequently updated by an exponential moving average of θ and ψ , respectively.

2) Updating Actor Networks

According to the policy evaluation of the attention-critic, the policy updates in each actor are performed. To provide the discrete actions, the discrete actions are represented by a one-hot encoder to compute the policy entropy $\mathcal{H}(\cdot | \mathbf{s}_i) = -\alpha \log(\pi(\cdot | \mathbf{s}_i))$. The policy parameters in the actor are learned by minimizing the expected Kullback–Leibler (KL)-divergence, resulting in the following policy objective [41]:

$$\mathcal{L}_\pi(\phi_i) = \mathbb{E}_{\mathbf{s}_i \sim \mathcal{D}_i} [\mathbb{E}_{\mathbf{a}_i \sim \pi} [\alpha \log(\pi(\mathbf{a}_i | \mathbf{s}_i) - Q_i(\mathbf{s}_i, \mathbf{a}_i))] \quad (29)$$

The output of the discrete actor is the exact action distribution, and thus the expectation in (29) can be directly calculated in the objective of the policy changes, expressed as:

$$\mathcal{L}_\pi(\phi_i) = \mathbb{E}_{\mathbf{s}_i \sim \mathcal{D}_i} [\pi(\mathbf{s}_i)^T (\alpha \log \pi(\mathbf{a}_i | \mathbf{s}_i) - Q_i(\mathbf{s}_i, \mathbf{a}_i))] \quad (30)$$

In this way, the original SAC algorithm for continuous actions is extended to provide discrete actions for the dedicated GEVC application. The actor network parameters are updated by:

$$\phi_i \leftarrow \phi_i - \lambda_\pi \widehat{\nabla}_{\phi_i} \mathcal{L}_\pi(\phi_i) \quad (31)$$

where the policy gradient $\widehat{\nabla}_{\phi_i} \mathcal{L}_\pi(\phi_i)$ is calculated by applying the policy gradient theorem to (30) [43].

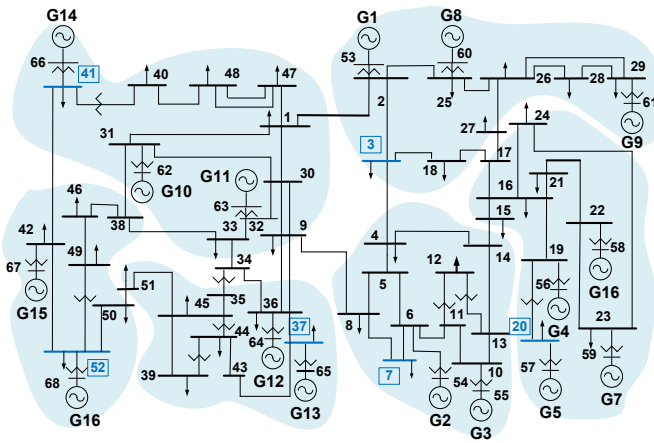


Fig. 5. The IEEE 16-machine 68-bus power system and network partition.

TABLE I
HYPERPARAMETER SETTING IN DRL

Hyperparameter	Value
The number of hidden layers	2
Learning rates for actors	0.001
Learning rates for critics	0.001
Discount factor γ	0.99
Replay buffer size	8000
Mini-batch size	3072
Optimizer	Adam
Target smoothing para. τ	0.005
The number of attention heads	4

IV. CASE STUDY

We test the proposed algorithm on the IEEE 16-machine 68-bus dynamic power systems [44]. The multi-area 68-bus system is abstracted from a real inter-connected power system of New England test system (NETS) with New York power system (NYPS), shown as Fig.5. There are six subnetworks in this multi-area power system, and each has a controllable load bus. The power system dynamics are simulated by solving the DAEs per 0.01s as the DRL environment. Dynamic components, including sub-transient generators, thermal turbine governors, static exciters, etc., are modeled and simulated in [45] as the environment. It is assumed that the controllable load nodes are located at buses 3, 7, 20, 37, 41, and 52 to provide control actions, and the total number of agents is 6. We also assume that the control action at each action time step can shed at most 20%, 10%, 10%, 5%, 5%, and 10% of the initial loads [2], according to the different initial load amounts on these six buses.

To exhibit the adaptivity of the DRL methods to various operating conditions, highly uncertain scenarios are randomly generated by combining different pre-fault load conditions (90%~110% of the base loads), fault duration time floating in

[0.06s, 0.1s], fault types, and fault locations. The simulation time is set as 6.5 seconds, and the number of action steps is 5 [8]. After a fault happens, the control actions provided by the DRL algorithm are conducted every second. The action searching space in each agent is 2^5 , and the joint action space in the multi-area power system is 2^{30} . The nodal voltages in all the buses inside a subnetwork are the observation and input to the DRL agent. Moreover, the latest 10 timesteps of all the nodal voltages in each subnetwork are the input to the GCN layer, i.e., the number of features is 10.

Table I lists the adopted hyperparameters in the proposed algorithm. The GCN model used in the actors is designed to be equipped with two hidden layers, and the size of each hidden layer is 32. The size of the fully connected layers is 128. All the training and tests are performed in Python 3.7 and run on a 3.2 GHz, 32 GB of RAM, Intel Core i9 computer.

A. Result Analysis

We compare the proposed algorithm with the multi-agent DQN [21] and SAC algorithms [30]. We also test a recent attention-embedded MARL algorithm, i.e., MAAC [39].

The data efficiency of different multi-agent DRL algorithms is quantified by 1) the required number of interactions with the environment and 2) the average reward (i.e., moving average rewards per successive 400 episodes [27]). We compare the data efficiency of these DRL algorithms in the offline training process of these methods, as shown in Table II. It can be observed that the reward of the proposed approach converges the fastest. Compared with MAAC, the proposed GCN-embedded model for feature representation accelerates the learning process in a data-efficient manner, and with the least number of training episodes (i.e., 112 episodes), the proposed method quickly converges. Meanwhile, the proposed method achieves higher rewards since it is equipped with the GCN model and the attention mechanism for cooperative learning simultaneously and provides more efficient control strategies. In contrast, the DQN-based method eventually reaches a lower reward level, due to its relatively poor exploration capability. To sum up, the proposed method illustrates superior performance in terms of convergence rate and reward improvement, compared with other DRL-based methods.

Furthermore, in online implementation, we test 1000 fault scenarios that are randomly set to investigate the control performance under uncertain environments. We compare the rewards of these DRL-based algorithms in all the test scenarios. We calculate the differences between the proposed algorithm and each of the other DRL-based algorithms (i.e., DQN, SAC, or MAAC) at the test scenario i by:

$$\Delta r_i = r_{pro,i} - r_{other,i} \quad (32)$$

where positive Δr_i denotes the superiority of the proposed algorithm over others. The higher reward value implies more efficient shedding performances of the time-series actions.

TABLE II
COMPARISON OF OFFLINE TRAINING PERFORMANCE IN DIFFERENT ALGORITHMS

MADRL Algorithms	Offline Training	
	No. Episode before convergence	Average reward after 800 episodes
MA-DQN	954	1063.1
MA-SAC	765	4161.5
MAAC	547	4392.2
Proposed	112	4769.7

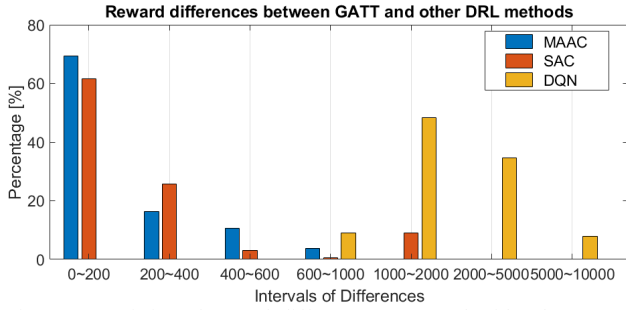


Fig. 6. A statistics of reward differences compared with other DRL-based methods in all the online tests. The x-axis shows the intervals that the differences between DQN (or SAC, MAAC) and the proposed method fall into.

Moreover, the statistics of the differences are shown in Fig.6. The higher reward value implies more efficient shedding performances of the time-series actions. It can be shown that the rewards of MAAC are closer to those of the proposed method, and the reward differences of 69.49% of them fall into the interval of 0~200. In contrast, the DQN-based method has the overall highest difference from the proposed approach since the actions determined by DQN violate the voltage constraints constantly, resulting in negative reward returns. Therefore, the proposed method with the pre-training paradigm exhibits outstanding improvements in control performance in terms of robustness and adaptiveness.

B. Performance Evaluation of Voltage Control

To further illustrate the advantages of the proposed method, the control performances of these different DRL-based methods after a fault happens are compared in a new test scenario.

In this scenario, the fault is single-phase grounded and happens on line 15-16 at 1.0s with a duration time of 0.1s. Fig. 7 shows the post-fault voltage profiles at buses 4, 15, and 16 using different control strategies from the GATT, MAAC, SAC, and DQN schemes. If no control actions are cast into the grid, this fault will lead to significant voltage violations, such as those on buses 4 and 15, shown as the green lines in Fig.7. In contrast, the proposed algorithm demonstrates an efficient voltage control performance, because it can predict the dynamic trend of the voltage states accurately by the GCN model and further approximate the nonlinear relationships between the actions and states. It can be observed that our method can lift the voltage curves above the TVRC envelope during the entire time horizon, which indicates the cooperative

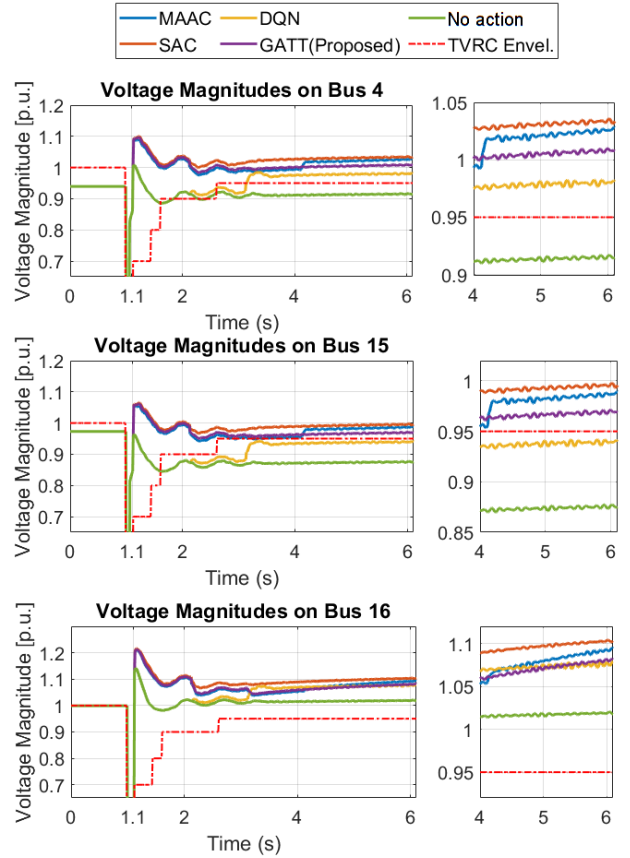


Fig.7. Comparison of different voltage trajectories using actions from DQN, SAC, MAAC, and GATT algorithms when a fault happens at 1s on the line 15-16. The zoom-in figures from 4s to 6s are provided to show the details of voltage variations with different control policies. Moreover, the trajectories with no control actions are depicted as the baseline.

control strategy can always satisfy the voltage constraints for GEVC. Moreover, the voltage curves led by the decision-making of the proposed GATT algorithm, shown as purple lines, are closer to the TVRC envelope than those of the MAAC method, which shows that fewer loads on the controllable buses are shed. Therefore, from the perspective of power supply quality and control costs, the proposed method outperforms other MARL-based methods.

It should be noted that in this case, the DQN-based control scheme cannot satisfy the requirements of the emergency voltage control, resulting in voltage violations with respect to the TVRC envelope. For instance, voltages on bus 4 with a duration of about 0.5s are below the TVRC envelope owing to the relatively poor prediction capability of FCN used in Q-learning. Moreover, the voltages on bus 15 after the first action at 1.1s drop below the TVRC envelope quickly, which implies that the actions provided by DQN, which is ineffective, fail to lift the voltage profiles to an expected level following TVRC.

The performance improvements of the proposed algorithm are illustrated in two aspects:

1) *Voltage Constraint Satisfaction.* The voltages after the control actions are executed are expected to satisfy the time-varying TVRC constraints in (9) during all the time horizons.

TABLE III
PERFORMANCE EVALUATION OF DIFFERENT DRL CONTROL METHODS
IN CASE OF FIG.7

Evaluation Metrics	Satisfy TVRC constraints?	Remaining Loads	Rewards
No action (Baseline)	No	100%	-21127.9
DQN	No	91.77%	-1191.1
SAC [27]	Yes	96.90%	4793.4
MAAC [29]	Yes	97.52%	4835.2
GATT (Proposed)	Yes	98.57%	4919.5

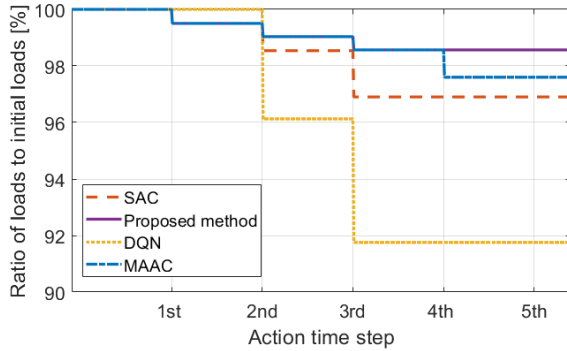


Fig. 8. Changing trends of the total online remaining loads from multi-agent DQN, SAC, and MAAC on all the controllable load buses in Case of Fig.7.

2) *Remaining Load Amounts.* Meanwhile, more loads are expected to be preserved on the controllable buses for better customer service and lower shedding costs. In particular, the ratio of remaining loads to the total controllable loads is adopted to evaluate the on-line load situation in different scenarios.

The reward, voltage constraint satisfaction, and remaining loads for different DRL algorithms are given in Table III. The total rewards of these DRL algorithms in this test case are 4793.4, 4835.2, and 4919.5, respectively. In contrast, DQN obtains the lowest reward among these DRL algorithms, i.e., -1191.1, since the control actions produce the voltages violation of the TVRC constraints, and meanwhile, more loads are shed.

Fig. 8 shows the load shedding amount at all the controllable load nodes using different DRL control approaches, and the changing trends of the total remaining loads in all the controllable nodes are illustrated in all the action steps. Among these DRL methods, the actions provided by the proposed algorithm preserve the highest percentage of controllable loads, i.e., 98.57% of the initial loads.

Moreover, we investigate the decision-making time of the proposed method for multi-timestep emergency control. During the online test, the decision-making of the actions on each agent takes 0.58 milliseconds per action timestep on average. Therefore, the well-trained model can provide proper control actions quickly in online implementation, making it promising to be applied to real-time GEVC in multi-area power systems.

V. CONCLUSION

This paper proposes a GATT-based MARL framework for dynamic emergency control confronting various contingencies in multi-area power systems. The proposed approach develops GCN models to extract topological information in a graph-aware way to make control decisions and accelerate the learning procedure. Moreover, an attention mechanism, which works among the agents for cooperative learning, helps each agent select more relevant information from other agents. Under the above data-efficient actor-attention-critic structure, the proposed method maximizes the averaged reward and policy entropy. Comprehensive case studies demonstrate that the proposed method significantly improves the control performances in terms of high adaptation and learning stability, in contrast to the existing MARL-based algorithms for the GEVC problem. Future work will focus on safe reinforcement learning to further strengthen the training stability [46] and meanwhile explore the connections between the cooperative control, DRL, and game theory [47] for power system emergency control.

REFERENCES

- [1] W. R. Lachs and D. Sutanto, "Voltage instability in interconnected power systems: a simulation approach," *IEEE Trans. Power Syst.*, vol. 7, no. 2, pp. 753–761, May 1992, doi: 10.1109/59.141782.
- [2] R. M. Larik, M. W. Mustafa, and M. N. Aman, "A critical review of the state-of-art schemes for under voltage load shedding," *Int. Trans. Electr. Energy Syst.*, vol. 29, no. 5, p. e2828, 2019, doi: 10.1002/2050-7038.2828.
- [3] D. P. S. Kundur, "Power System Stability and Control," 10, pp.7-1, 2007.
- [4] Q. Li, Y. Xu, and C. Ren, "A Hierarchical Data-Driven Method for Event-based Load Shedding Against Fault-Induced Delayed Voltage Recovery in Power Systems," *IEEE Trans. Ind. Inform.*, pp. 1–1, 2020, doi: 10.1109/TII.2020.2993807.
- [5] D. Lefebvre, C. Moors, and T. Van Cutsem, "Design of an undervoltage load shedding scheme for the Hydro-Quebec system," in *2003 IEEE Power Engineering Society General Meeting (IEEE Cat. No.03CH37491)*, Jul. 2003, vol. 4, pp. 2030-2036 Vol. 4. doi: 10.1109/PES.2003.1270926.
- [6] M. Larsson and D. Karlsson, "Coordinated system protection scheme against voltage collapse using heuristic search and predictive control," *IEEE Trans. Power Syst.*, vol. 18, no. 3, pp. 1001–1006, Aug. 2003, doi: 10.1109/TPWRS.2003.814852.
- [7] J. Y. Wen, Q. H. Wu, D. R. Turner, S. J. Cheng, and J. Fitch, "Optimal coordinated voltage control for power system voltage stability," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 1115–1122, May 2004, doi: 10.1109/TPWRS.2004.825897.
- [8] L. Jin, R. Kumar, and N. Elia, "Model Predictive Control-Based Real-Time Power System Protection Schemes," *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 988–998, May 2010, doi: 10.1109/TPWRS.2009.2034748.
- [9] Z. Li, G. Yao, G. Geng, and Q. Jiang, "An Efficient Optimal Control Method for Open-Loop Transient Stability Emergency Control," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2704–2713, Jul. 2017, doi: 10.1109/TPWRS.2016.2629620.
- [10] D. A. Panasetsky and N. I. Voropai, "A Multi-agent approach to coordination of different emergency control devices against voltage collapse," in *2009 IEEE Bucharest PowerTech*, Jun. 2009, pp. 1–7. doi: 10.1109/PTC.2009.5281995.
- [11] Sk. R. Islam, D. Sutanto, and K. M. Muttaqi, "Coordinated Decentralized Emergency Voltage and Reactive Power Control to Prevent Long-Term Voltage Instability in a Power System," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2591–2603, Sep. 2015, doi: 10.1109/TPWRS.2014.2369502.
- [12] H. Shahbazi and F. Karbalaei, "Decentralized Voltage Control of Power Systems Using Multi-agent Systems," *J. Mod. Power Syst. Clean Energy*, vol. 8, no. 2, pp. 249–259, Mar. 2020, doi: 10.35833/MPCE.2018.000628.

- [13] Z. Yan and Y. Xu, "Data-Driven Load Frequency Control for Stochastic Power Systems: A Deep Reinforcement Learning Method With Continuous Action Search," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1653–1656, Mar. 2019, doi: 10.1109/TPWRS.2018.2881359.
- [14] H. Ma and D. J. Hill, "Adaptive Coordinated Voltage Control—Part I: Basic Scheme," *IEEE Trans. Power Syst.*, vol. 29, no. 4, pp. 1546–1553, Jul. 2014, doi: 10.1109/TPWRS.2013.2293577.
- [15] T. Van Cutsem and C. D. Vournas, "Emergency Voltage Stability Controls: an Overview," in *2007 IEEE Power Engineering Society General Meeting*, Tampa, FL, USA, Jun. 2007, pp. 1–10, doi: 10.1109/PES.2007.386089.
- [16] R. Sun and Y. Liu, "Hybrid Reinforcement Learning for Power Transmission Network Self-Healing Considering Wind Power," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–11, 2021, doi: 10.1109/TNNLS.2021.3136554.
- [17] A. Traue, G. Book, W. Kirchgässner, and O. Wallscheid, "Toward a Reinforcement Learning Environment Toolbox for Intelligent Electric Motor Control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 919–928, Mar. 2022, doi: 10.1109/TNNLS.2020.3029573.
- [18] J. Jung, Chen-Ching Liu, S. L. Tanimoto, and V. Vittal, "Adaptation in load shedding under vulnerable operating conditions," *IEEE Trans. Power Syst.*, vol. 17, no. 4, pp. 1199–1205, Nov. 2002, doi: 10.1109/TPWRS.2002.805023.
- [19] D. Ernst, M. Glavic, and L. Wehenkel, "Power systems stability control: reinforcement learning framework," *IEEE Trans. Power Syst.*, vol. 19, no. 1, pp. 427–435, Feb. 2004, doi: 10.1109/TPWRS.2003.821457.
- [20] J. Zhang, Y. Luo, B. Wang, C. Lu, J. Si, and J. Song, "Deep Reinforcement Learning for Load Shedding Against Short-Term Voltage Instability in Large Power Systems," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–12, 2021, doi: 10.1109/TNNLS.2021.3121757.
- [21] Q. Huang, R. Huang, W. Hao, J. Tan, R. Fan, and Z. Huang, "Adaptive Power System Emergency Control Using Deep Reinforcement Learning," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1171–1182, Mar. 2020, doi: 10.1109/TSG.2019.2933191.
- [22] Y. Zhang, M. Yue, and J. Wang, "Adaptive Load Shedding for Grid Emergency Control via Deep Reinforcement Learning," in *2021 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 01-05, IEEE, 2021.
- [23] R. Huang *et al.*, "Accelerated Deep Reinforcement Learning Based Load Shedding for Emergency Voltage Control," *ArXiv200612667 Cs Eess*, Jun. 2020, Accessed: Aug. 20, 2020. [Online]. Available: <http://arxiv.org/abs/2006.12667>
- [24] R. R. Hossain, Q. Huang, and R. Huang, "Graph Convolutional Network-Based Topology Embedded Deep Reinforcement Learning for Voltage Stability Control," *IEEE Trans. Power Syst.*, vol. 36, no. 5, pp. 4848–4851, Sep. 2021, doi: 10.1109/TPWRS.2021.3084469.
- [25] Y. Wu, E. Mansimov, R. B. Grosse, S. Liao, and J. Ba, "Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation," in *Advances in Neural Information Processing Systems*, 2017, vol. 30. Accessed: Mar. 21, 2022. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/361440528766bbaaa1901845cf4152b-Abstract.html>
- [26] Sk. R. Islam, D. Sutanto, and K. M. Muttaqi, "A Distributed Multi-Agent Based Emergency Control Approach Following Catastrophic Disturbances in Interconnected Power Systems," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 2764–2775, Jul. 2016, doi: 10.1109/TPWRS.2015.2469543.
- [27] Y. Zhang, X. Wang, J. Wang, and Y. Zhang, "Deep Reinforcement Learning Based Volt-VAR Optimization in Smart Distribution Systems," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 361–371, Jan. 2021, doi: 10.1109/TSG.2020.3010130.
- [28] S. Wang *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, doi: 10.1109/TPWRS.2020.2990179.
- [29] D. Cao, J. Zhao, W. Hu, F. Ding, Q. Huang, and Z. Chen, "Attention Enabled Multi-Agent DRL for Decentralized Volt-VAR Control of Active Distribution System Using PV Inverters and SVCs," *IEEE Trans. Sustain. Energy*, vol. 12, no. 3, pp. 1582–1592, Jul. 2021, doi: 10.1109/TSTE.2021.3057090.
- [30] M. Kamruzzaman, J. Duan, D. Shi, and M. Benidris, "A Deep Reinforcement Learning-based Multi-Agent Framework to Enhance Power System Resilience using Shunt Resources," *IEEE Trans. Power Syst.*, pp. 1–1, 2021, doi: 10.1109/TPWRS.2021.3078446.
- [31] H. Zhang, D. Yue, C. Dou, X. Xie, K. Li, and G. P. Hancke, "Resilient Optimal Defensive Strategy of TSK Fuzzy-Model-Based Microgrids' System via a Novel Reinforcement Learning Approach," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–11, 2021, doi: 10.1109/TNNLS.2021.3105668.
- [32] L. Xi, J. Wu, Y. Xu, and H. Sun, "Automatic Generation Control Based on Multiple Neural Networks With Actor-Critic Strategy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2483–2493, Jun. 2021, doi: 10.1109/TNNLS.2020.3006080.
- [33] J. Zhang, C. Lu, C. Fang, X. Ling, and Y. Zhang, "Load Shedding Scheme with Deep Reinforcement Learning to Improve Short-term Voltage Stability," in *2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, May 2018, pp. 13–18.
- [34] L. Consolini, F. Morbidi, D. Prattichizzo, and M. Tosques, "Leader-follower formation control of nonholonomic mobile robots with input constraints," *Automatica*, vol. 44, no. 5, pp. 1343–1349, 2008.
- [35] H. G. Tanner, G. J. Pappas, and V. Kumar, "Leader-to-formation stability," *IEEE Trans. Robot. Autom.*, vol. 20, no. 3, pp. 443–455, Jun. 2004.
- [36] W. Gao, Z. -P. Jiang, F. L. Lewis and Y. Wang, "Leader-to-Formation Stability of Multiagent Systems: An Adaptive Optimal Control Approach," *IEEE Transactions on Automatic Control*, vol. 63, no. 10, pp. 3581–3587, Oct. 2018.
- [37] J. R. Marden, Gü. Arslan, and J. S. Shamma, "Cooperative Control and Potential Games," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 39, no. 6, pp. 1393–1407, Dec. 2009, doi: 10.1109/TSMCB.2009.2017273.
- [38] K. G. Vamvoudakis, F. L. Lewis, and G. R. Hudas, "Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality," *Automatica*, vol. 48, no. 8, pp. 1598–1611, 2012.
- [39] S. Iqbal and F. Sha, "Actor-Attention-Critic for Multi-Agent Reinforcement Learning," in *International Conference on Machine Learning*, May 2019, pp. 2961–2970. Accessed: Aug. 13, 2021. [Online]. Available: <http://proceedings.mlr.press/v97/iqbal19a.html>
- [40] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *ArXiv160902907 Cs Stat*, Feb. 2017, Accessed: Aug. 29, 2020. [Online]. Available: <http://arxiv.org/abs/1609.02907>
- [41] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," *ArXiv180101290 Cs Stat*, Aug. 2018, Accessed: Mar. 12, 2020. [Online]. Available: <http://arxiv.org/abs/1801.01290>
- [42] A. Vaswani *et al.*, "Attention Is All You Need," *ArXiv170603762 Cs*, Dec. 2017, Accessed: Sep. 22, 2020. [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [43] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.
- [44] Hassan Bevrani; Masayuki Watanabe; Yasunori Mitani, "Appendix A: New York/New England 16-Machine 68-Bus System Case Study," in *Power System Monitoring and Control*, IEEE, 2014, pp.249-253.
- [45] Power System Toolbox. J. H. Chow, [Online]. Available: <https://www.ecse.rpi.edu/~chow/>
- [46] L. Zhang, R. Zhang, T. Wu, R. Weng, M. Han, and Y. Zhao, "Safe Reinforcement Learning With Stability Guarantee for Motion Planning of Autonomous Vehicles," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5435–5444, Dec. 2021.
- [47] A. Nowé, P. Vrancx, and Y.-M. De Hauwere, *Game Theory and Multi-agent Reinforcement Learning*. Berlin, Germany: Springer, 2012, pp. 441–470.