A Deep Reinforcement Learning-Based Multi-Agent Framework to Enhance Power System Resilience Using Shunt Resources

Md. Kamruzzaman, *Member, IEEE*, Jiajun Duan, *Member, IEEE*, Di Shi, *Senior Member, IEEE*, and Mohammed Benidris, *Senior Member, IEEE*

***Abstract*—Existing power system resilience enhancement methods, such as proactive generation rescheduling, movable sources dispatch, and network topology reconfiguration, do not explore the capability and flexibility of shunts to maintain voltage stability during and after disrupting events. Besides, existing methods rely on accurate system models that are not easily scalable for large integrated power grids. In this paper, a data-driven multi-agent framework based on a deep-reinforcement-learning algorithm is proposed to overcome the computation and scalability concerns related to precise system models and to plan for the deployment of shunts for power system resilience enhancement. Specifically, voltage violations due to outages of multiple lines during windstorms are taken as an example of a power system resilience improvement problem. Then, a multi-agent based hybrid soft actor critic (HSAC) algorithm is developed for offline siting and sizing as well as online controlling of shunt reactive power compensators to enhance voltage resilience. The HSAC algorithm is derived from the fundamental SAC algorithms that contain both continuous and discrete action spaces. The proposed multi-agent framework learns from previous experiences and eventually gets trained to determine proper locations and sizes for shunts to avoid voltage violations during multiple line failures. The proposed approach is demonstrated on the IEEE 57-bus and IEEE 300-bus systems. The results show that the proposed multi-agent framework is effective for installation planning and controlling of shunts to enhance power system resilience.**

***Index Terms*—Deep reinforcement learning, hybrid soft actor critic, multi-agent framework, power system resilience, shunt reactive power compensators, voltage control.**

I. Introduction

EXTREME weather events and man-made attacks have severe impacts on power systems ranging from long outage duration to multiple equipment failures. This calls for developing appropriate countermeasures to improve power system resilience. Maintaining voltage stability during and after extreme events and multiple equipment failures is one of the key factors to improve power system stability, resilience, and ability to prevent cascading failures. Deployment of shunt resources can be a promising solution to maintain voltage constraints under N − k (k>1) contingencies through providing reactive power support. Therefore, developing a method to plan for deploying shunts to maintain voltage constraints during and after extreme events is indispensable to enhance the resilience of power grids.

Several power system resilience enhancement methods have been proposed in the literature to provide emergency response during multiple contingencies [1]–[14]. A mixed-integer linear programming (MILP)-based two-stage method has been proposed in [1] to enhance power system resilience through changing network topology, re-dispatching generators, and shedding loads. In [2], a two-stage control algorithm to dispatch unused capacitors of a multi-microgrid system has been proposed to enhance power system resilience. A MILP-based method to re-dispatch generation during ice storms has been proposed in [4] to enhance power system resilience. Also, resilience enhancement approaches have been proposed to provide emergency responses through switching topology [5], re-dispatching loads [6], [7], forming networked microgrid [8], and using energy storage devices [9]. Moreover, several preventive action-based strategies such as a multi-sensor prediction-based wide-area monitoring and control [11], a linear-programming-based optimal siting and sizing of energy storage devices [12], a Monte-Carlo simulation (MCs)-based proactive unit commitment framework [13], and an MCS-based crew preposition and network reconfiguration technique [14] have been proposed to enhance the resilience of power systems. Although the proposed methods in [1]–[14] are effective to enhance the resilience of power systems, these methods are not computationally flexible to use for large integrated power grids due to their dependency on accurate system information. Also, these methods do not have the flexibility to utilize the necessary resources to maintain system constraints with feeding the entire load demand during contingencies. Therefore, a resilience enhancement method that does not require accurate system knowledge and flexible to use expansively accessible resources needs to be developed to achieve resilient power grids.

Manuscript received September 16, 2020; revised January 27, 2021 and April 10, 2021; accepted May 4, 2021. Date of publication May 10, 2021; date of current version October 20, 2021. This work was supported by the U.S. National Science Foundation (NSF) under Grant NSF 1847578. Paper no. TPWRS-01593-2020.R2. *(Corresponding author: Jiajun Duan.)*

Md. Kamruzzaman and Mohammed Benidris are with the Department of Electrical and Biomedical Engineering, University of Nevada, Reno, NV 89557, USA (e-mail: mkamruzzaman@nevada.unr.edu; mbenidris@unr.edu).

Jiajun Duan is with Nextracker Inc., Fremont CA, 94555 USA (e-mail: jiajunduan.ee@gmail.com).

Di Shi is with AINERGY LLC., Santa Clara CA, 95050 USA (e-mail: sdxjtu@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPWRS.2021.3078446.

Digital Object Identifier 10.1109/TPWRS.2021.3078446

Although several model-based methods have been proposed in [15]–[18] to control voltages of power systems, these methods depend on accurate parameters and system knowledge, which is quite challenging for modern power grids with increasing dimensionality and complexity. Several deep-reinforcement-learning (DRL)-based voltage control methods have been proposed in the literature [19]–[23] to avoid the requirement of computationally expensive exact system models. A model free Q-learning-based voltage control algorithm has been introduced in [19] to provide optimal control settings for the constrained load flow problem. In [20], a Q-learning-based distributed voltage control method has been proposed to optimally dispatch reactive power. In [21], optimal tap setting policy for voltage regulation transformers has been determined using a DRL algorithm. A two-time scale voltage control algorithm that uses deep Q-network to determine optimal capacitor configuration in slow time scale has been proposed in [22]. A single-agent centralized automatic voltage control framework based on deep-Q-network and deep deterministic policy gradient (DDPG) method has been proposed in [23]. A centralized trained and decentralized executed DDPG-based multi-agent framework has been proposed in [24] to control voltages of large integrated power grids. The proposed DRL-based methods in [19]–[24] are effective to provide control actions without accurate system knowledge to maintain voltage constraints under $N-1$ contingency. However, these methods utilize existing system resources to provide a single type of corrective control actions. Therefore, developing a DRL-based computationally efficient multi-task method to control voltages against multiple contingencies is important to enhance resilience of power systems.

In this paper, a data-driven multi-agent framework using a DRL algorithm is proposed to plan for the deployment of shunts to enhance power system resilience against multiple line failures. Each agent of the multi-agent framework is constructed using a hybrid soft-actor-critic (HSAC) algorithm. The HSAC is developed using the fundamental SAC algorithms for continuous [25] and discrete [26] actions. To plan for the deployment of shunts using the proposed multi-agent framework, first, the entire power system is partitioned into several regions. Then, offline training of the multi-agent framework is performed in centralized fashion to determine locations and sizes of shunts to maintain voltage constraints against multiple line failures. The training algorithm is performed using historical data and fragility curves of transmission lines and is periodically updated to capture changes in system parameters. At each training period, a reward function is used to evaluate the effectiveness of actions—selected locations and sizes for shunts. Actions, rewards of the agents, and system states at each training period are stored in a replay buffer. A randomly sampled minibatch is used to periodically update the parameters of the multi-agent framework. Through the penalization mechanism designed in the reward function, the agents learn from experiences to avoid the execution of inaccurate actions and eventually get trained to provide accurate actions. Similar to the other DRL-based methods [20]–[21], the proposed multi-agent SAC (MASAC) framework can provide a near-optimal or occasionally optimal solution. Therefore, after installing the shunts, the trained agent can be used to provide support to grid operators to control dispatched power from shunts using local measurements of power grids. However, decisions from the agents can be firstly confirmed by system operators to avoid risks. Contributions of the proposed work in comparison with existing methods are summarized as follows.

- The proposed multi-agent MASAC framework can improve the scalability issues of existing DRL-based methods. In addition to the voltage control problem of power systems, the proposed algorithm is also flexible to extend and apply for other control problems.
- The agents of the proposed framework provide both continuous and discrete actions simultaneously, which provide the flexibility to determine locations and sizes for expansively accessible resources to maintain system security during contingencies.
- Policies of the proposed method to provide actions are trained to maximize the trade-off between entropy (randomness measure of the policy) and expected return, which is closely related to the exploitation and exploration performance. The increase in the entropy leads to more exploration that accelerates the learning rates of the agents. This also avoids premature convergence of the policies, which is important to obtain local optimum.
- All actor networks of the proposed MASAC framework share information with a centralized critic network directly and update their policies based on the provided rewards by the critic network and regional states (bus voltages of each region). This eliminates the requirement of coordinators to share information among the actor networks. This also provides flexibility to apply the proposed MASAC framework for large-scale integrated power systems with a low computational burden.

The rest of the paper is arranged as follows. Section II describes problem formulation to control voltages under extreme events using the MASAC framework. Section III explains the formulation of the HSAC. Section IV describes proposed power system resilience enhancement algorithm. Section V provides the training and execution algorithm of the multi-agentframework. Section VI demonstrates the proposed solution. Section VII provides several concluding remarks.

II. Problem Formulation

To achieve the goal of enhancing power system resilience using the developed MASAC framework, first, the entire power system is divided into several small regions based on the geographic locations/electrical distances of buses. Then, each agent of the MASAC framework is assigned to control voltages of a region. It should be noted that the number of agents of the MASAC framework is equal to the number of regions. Finally, the voltage control problem under extreme events is formulated for the MASAC framework. In the MASAC framework, the agents provide actions based on system states, which are implemented in the environment (power system) to control voltages. These actions are also transmitted to the critic network using a centralized communication network during the training process. The critic network provides a reward to agents, which is used to update policies of actor networks. Well-trained agents of

TABLE I
CALCULATION OF REWARDS FOR POWER SYSTEM BUSES

Operation Zone	Bus voltage (V_k^t)	r_k^t
Normal	$[V_{ref}, V^{ub}]$	$\frac{V^{ub} - V_k^t}{V^{ub} - V_{ref}}$
Normal	$[V^{lb}, V_{ref}]$	$\frac{V_k^t - V^{lb}}{V_{ref} - V^{lb}}$
Violation	$[V^{ub}, 1.25]$	$\frac{V_k^t - V_{ref}}{V_{ref} - 1.25}$
Violation	$[0.8, V^{lb}]$	$\frac{V_{ref} - V_k^t}{0.8 - V_{ref}}$
Diverge	$[0.0, 0.8]$	-5
Diverge	$[1.25, \infty]$	-5

the MASAC framework use only the states to provide control commands during testing/execution without a communication network. Therefore, we need to properly design states, observations, actions, and rewards to formulate the voltage control problem for the MASAC framework. This section describes states, observations, actions, and rewards in the context of power system resilience enhancements.

1) States, Observations, and Actions: Various parameters can be used to represent system states [24], [27], [28]; however, for reactive power control studies, voltage magnitudes have been widely used. It indicates the effectiveness of DRL algorithms in streaming valuable information using partial states, which provides flexibility to save a large number of measurements and communication [24]. In this study, voltage states are divided into three zones as shown in Table I.

We assume that each agent can observe and manage voltage magnitudes of its own region. The control actions are defined as a vector of locations and amount of output reactive power of shunts. Each element of this vector is updated continuously by adjusting both the locations and amount of output reactive power of shunts. The amount of output reactive power from shunts are adjusted within a predefined range of minimum and maximum values. The discrete actions taken by the agents are used to determine locations for shunts. Each or a selected number of buses of a region can be used as candidate buses in the discrete action space to select locations by an agent.

1) Calculation of Rewards: We design the reward function through a hierarchical assumption to evaluate the effectiveness of the actions taken by the MASAC. The first objective to design the reward, r_k^t, is to encourage each agent to decrease bus voltage magnitude deviation during contingencies from a predefined reference value, $V_{ref} = 1.0$ p.u. The definitions of rewards, $r_{i,k}^t$, are provided in Table I.

From Table I, it can be seen that buses with small deviations are awarded large rewards. If the values of all bus voltages remain in normal or violation zones after dispatching shunts then, the total reward is calculated as follows,

$$R^t = \frac{\sum_{k=1}^{N^b} r_k^t}{N^b} \tag{1}$$

where R^t is the total reward at time step t; r_k^t is the reward for k^th bus at time step t; V_k^t is the voltage magnitude of bus k at time t; and N^b is the total number of buses.

If the voltage magnitude of a bus is in the divergence zone, then a relatively large penalty is assigned.

III. PROPOSED MULTI-AGENT FRAMEWORK

Each agent of the proposed MASAC framework provides both continuous and discrete actions using two separate policy functions. The fundamental SAC algorithms provided in [25], [26] are adapted in this study to construct policy functions and training algorithms of the proposed framework.

A. Policies for Actor-Networks of the Proposed Framework

Similar to the single agent SAC algorithm for either continues or discrete action, in the proposed multi-agent SAC (MASAC) framework, the continuous actor-network of each agent is developed based on a squashed Gaussian distribution function and the discrete actor-network is developed based on a Gumbel soft-max distribution function. The policies for the actor-networks can be expressed as,

$$a_t^{ci} \sim \pi_{\phi^{ci}}(a_t^{ci}|o_t^i), \tag{2}$$

$$a_t^{di} \sim \pi_{\phi^{di}}(a_t^{di}|o_t^i), \tag{3}$$

where o_t^i is the observation vector at time t for agent i; a_t^{ci} is the action provided by the continuous actor-network of agent i; a_t^{di} is the provided action by discrete actor-network of agent i; ϕ^{ci} is the parameter for continuous policy network of agent i; ϕ^{di} is the parameter for discrete policy network of agent i; $\pi_{\phi^{ci}}(a_t^{ci}|o_t^i)$ is an unbounded Gaussian policy; and $\pi_{\phi^{di}}(a_t^{di}|o_t^i)$ is the discrete policy. The continuous actions need to be bounded to a finite value in practice and, therefore, a squashing function is applied to the Gaussian samples.

B. Policy Training Algorithm for Continuous Actors

Following the same convention of fundamental SAC algorithms, policies of the MASAC are updated in each iteration to maximize the trade-off between entropy (randomness measure of the policy) and expected return. The policy used in the fundamental SAC algorithm to maximize entropy is modified as follows.

$$\pi_{\phi^{ci}}^* = \arg\max_{\pi_{\phi^{ci}}} \sum_{t=0}^{T} \mathbb{E}_{(s_t, a_t^{ci}) \sim \tau_{\pi_{\phi^{ci}}}} \left[\gamma_t R(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di}) + \alpha^{ci} H\pi_{\phi^{ci}}(.|s_t)\right] \tag{4}$$

where $\pi_{\phi^{ci}}^*$ is the optimal policy for agent i; T and R represent the number of time steps and reward function; $\gamma \in [0, 1]$ is a discount factor; a_t^{-ci} is the selected actions by continuous actors of all other agents; a_t^{-di} is the selected actions by discrete actors of all other agents; α^{ci} is a temperature parameter which determines the relative importance between entropy and reward of agent i; s_t is the set of system states; $\tau_{\pi_{\phi^{ci}}}$ is the induced trajectories by the policy $\pi_{\phi^{ci}}$; and $H\pi_{\phi^{ci}}(.|s_t)$ is the entropy of the policy $\pi_{\phi^{ci}}$ for state s_t, which is calculated as $H\pi_{\phi^{ci}}(.|s_t) = -log\phi^{ci}(.|s_t)$.

As the continuous policy needs to be bounded in practice, an approximation to soft policy iteration needs to be derived. Similar to the fundamental SAC algorithm, an alternative method

between the policy evaluation and improvement is used to maximize the entropy. The policy evaluation requires the calculation of value of the policy. The soft value function, $V_{\psi^i}^{ci}(o_t^i)$, that is used to measure the value of continuous policy for agent, i, is expressed as,

$$V_{\psi^i}^{ci}(o_t^i) = \mathbb{E}_{a_t^{ci}\sim\pi_{\phi^{ci}}}\left[Q_\theta(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di}) - \alpha^{ci}\log\left(\pi_{\phi^{ci}}(a_t^{ci}|o_t^i)\right)\right] \quad (5)$$

where ψ^i is the parameter for the value function network of agent i; θ is the parameter for Q value function; and $Q_\theta(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di})$ is a centralized soft policy evaluation or critic function for the continuous actors.

The expression to minimize the squared residual error of a soft Bellman function to train the soft value functions of continuous actors of the MASAC framework is expressed as,

$$J_v^{ci}(\psi^i) = \mathbb{E}_{s_t^{ci}\sim\mathcal{D}}\left[\frac{1}{2}V_{\psi^i}^{ci}(o_t^i) - \left[Q_\theta(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di}) - \alpha^{ci}\log\left(\pi_{\phi^{ci}}(a_t^{ci}|o_t^i)\right)\right]^2\right] \quad (6)$$

where $\mathcal{D}$ is a replay buffer to store previous experiences.

The expression to determine the gradient of (6) based on an unbiased estimator is as follows.

$$\hat{\nabla}_{\psi^i} J_v^{ci}(\psi^i) = \nabla_{\psi^i} V_{\psi^i}^{ci}(o_t^i)\left(V_{\psi^i}^{ci}(o_t^i) - Q_\theta(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di}) + \alpha^{ci}\log\left(\pi_{\phi^{ci}}(a_t^{ci}|o_t^i)\right)\right) \quad (7)$$

The actions in (7) are sampled from the current policy. The expression to train the soft-Q parameters for single-agent continuous actor provided in [25] is modified for the MASAC framework as,

$$J_{Q_\theta}^{ci}(\theta^i) = \mathbb{E}_{(s_t^{ci}, a_t^{ci})\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\theta(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di}) - \hat{Q}(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di})\right)^2\right] \quad (8)$$

with

$$\hat{Q}(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di}) = r(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di}) + \gamma\mathbb{E}_{s_{t+1}\sim p}[V_{\bar{\psi}^i}^{ci}(o_{t+1}^i)] \quad (9)$$

where $\bar{\psi}^i$ is an exponentially moving average of the value network weights for agent i.

The gradient to optimize the soft Q-function of (8) is as follows.

$$\hat{\nabla}_{\theta^i} J_{Q_\theta}^{ci}(\theta^i) = \nabla_{\theta^i} Q_\theta(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di}) \times \left(Q_\theta(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di}) - r(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di}) - \gamma V_{\bar{\psi}^i}^{ci}(o_{t+1}^i)\right) \quad (10)$$

To improve the policy, the policy needs to be updated in such a way that it will maximize the rewards. In [25], the authors have used the soft Q-value during policy evaluation to guide the policy update. In actual scenario, the policy update has been directed toward exponential of new soft Q-function to make the policy tractable. Also, the authors restricted the possible policies to a parameterized distributions family (e.g., Gaussian). Following the same convention, the policy update expression of single agent SAC in terms of Kullback-Leibler divergence is modified for the MASAC as,

$$\pi_{\phi^{ci}}^{new} = \arg\min D_{KL}\left(\pi_{\phi^{ci}}(.|o_t^i)\middle\|\frac{Q_\theta(s_t, .)}{Z_\theta(s_t)}\right) \quad (11)$$

where $Z_\theta(s_t)$ is an intractable partition function which does not contribute to the gradient with respect to the new policy.

The policy $\pi_{\phi^{ci}}(.|o_t^i)$ is parameterized for continuous action setting using the continuous policy network of agent, i, with parameter ϕ^{ci}. Finally, the expected KL-divergence of (11) is multiplied by α^{ci}, and then, minimized ignoring $Z_\theta(s_t)$ to train the policy parameters of agent, i, as follows.

$$J_{\pi_{\phi^{ci}}}^{ci}(\phi^{ci}) = \mathbb{E}_{s_t^{ci}\sim\mathcal{D}}\left[\mathbb{E}_{a_t^{ci}\sim\pi_{\phi^{ci}}}\left[\alpha^{ci}\log\left(\pi_{\phi^{ci}}(a_t^{ci}|o_t^i)\right) - Q_\theta(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di}, a_t^{-di})\right]\right] \quad (12)$$

Although several options are available to minimize the objective function $J_{\pi_{\phi^{ci}}}^{ci}(\phi^{ci})$, the reparameterization technique has been applied in [29] to achieve target density which is Q-function. The modified expression to reparametrize the policy of agent, i, is as follows.

$$a_t^{ci} = f_{\phi^{ci}}(\epsilon_t^{ci}; o_t^i) \quad (13)$$

where ϵ_t^{ci} is a noise vector that is sampled using a spherical Gaussian distribution.

Thus, the new continuous policy objective for agent, i, is as follows.

$$J_{\pi_{\phi^{ci}}}^{ci}(\phi^{ci}) = \mathbb{E}_{s_t^{ci}\sim\mathcal{D}, \epsilon_t^{ci}\sim\mathcal{N}}\left[\alpha^{ci}\log\left(\pi_{\phi^{ci}}(f_{\phi^{ci}}(\epsilon_t^{ci}; o_t^i)|o_t^i)\right) - Q_\theta(s_t, f_{\phi^{ci}}(\epsilon_t^{ci}; o_t^i), f_{\phi^{-ci}}(\epsilon_t^{-ci}; s_t^{-i}), a_t^{di}, a_t^{-di})\right] \quad (14)$$

where $f_\phi^{-ci}(\epsilon_t^{-ci}; s_t^{-i})$ is the parameterized policies of other continuous actors.

In [30], an alternative approach has been provided to learn the temperature parameters of SAC algorithm for continuous actions without the need of setting hyper-parameter. Instead of reproducing rigorous formulation of obtaining the temperature parameter learning objective function, the provided temperature objective function in [30] is modified for continuous actors of the proposed MASAC framework as follows.

$$J^{ci}(\alpha^{ci}) = \mathbb{E}_{a_t^{ci}\sim\pi_{\phi^{ci}}}\left[-\alpha^{ci}\left(\log\left(\pi_{\phi^{ci}}(a_t^{ci}|o_t^i) + \bar{H}\right)\right)\right] \quad (15)$$

where $\bar{H}$ is an equivalent constant vector of the hyper-parameter to represent target entropy. The (15) cannot be minimized directly due to the expectation operator. Therefore, it is minimized using a Monte-Carlo estimator after sampling experiences from replay buffer based on the procedure from [30]. In the proposed MASAC, two soft Q-networks for all continuous agents are trained, and then, the minimum value among the outputs of two Q-networks is used in the objective function of (15). We do this because it is beneficial to combat state-value overestimation [31].

C. Policy Training Algorithm for Discrete Actors

The developed policy update procedure in [26] is modified in this study to develop the policy update procedure for discrete actor-networks of the proposed MASAC framework. The steps involved in deriving objectives for the continuous actor-networks of each agent hold for the discrete actor-networks. The only change that happens for the discrete actor-networks is that the policy of agent i, $\pi_{\phi_{di}}(a_t^{di}|o_t^i)$ provides a probability instead of a density. Therefore, the objective functions for the continuous actor-networks described in (10), (12), and (15) remain same for the discrete actor-networks. However, the following changes are necessary to optimize these objective functions.

- Using output value (Q-value) of the soft-Q function for each possible action is more efficient than providing only actions as input to train discrete actors. Therefore, the Q-function for discrete actors changes from $Q_\theta(s_t, a_t^{ci}, a_t^{-ci}, a_t^{di})$ to $Q_\theta(s_t)$, which is not possible for infinitely many possible actions of continuous actors.
- The discrete policy can directly provide action distribution without the need of calculating the mean and covariance of action distribution. Therefore, application of a softmax function in the output layer can provide a valid probability distribution of discrete actions.
- The estimation of the soft state value in (5) requires taking an expectation over the distribution of continuous actions. Therefore, sampled actions from replay buffer needs to be used to minimize the cost of the soft Q-function, $J_{Q_\theta}^{ci}(\theta^i)$, in (8) using the Monte Carlo estimate of the soft state-value function (5). On the other hand, as the action space is discrete for the discrete actors, actions can be recovered fully, which removes the requirement of using a Monte Carlo estimator for expectation calculation. This reduces the variance involved in estimating the objective, $J_{Q_\theta}^{ci}(\theta^i)$, (8).

The soft-value function, $V_{\psi^i}^{di}(o_t^i)$, for the discrete actor-network of agent, i, is as follows.

$$V_{\psi^i}^{di}(o_t^i) = \pi_{\phi^{di}}(s_t)^T \left[Q_\theta(s_t) - \alpha^{dis} \log\left(\pi_{\phi^{di}}(o_t^i)\right)\right] \quad (16)$$

where θ is the parameter for Q value function; $Q_\theta(s_t)$ is a centralized soft policy evaluation or critic function for the discrete actors; and α^{dis} is a temperature parameter that determines the relative importance between entropy and reward for agent i.

It can be noticed that the centralized Q-function for discrete actors receives observation from all agents. Thus, the continuous policy objective function in (14) is modified for the discrete actors as follows.

$$J_{\pi_{\phi^{di}}}^{di}(\phi^{di}) = \mathbb{E}_{s_t^{di}\sim\mathcal{D}}\left[\pi_{\phi^{di}}(s_t)^T \left[\alpha^{di} \log\left(\pi_{\phi^{di}}(o_t^i)\right) - Q_\theta(s_t)\right]\right] \quad (17)$$

The temperature function described in (15) is modified for the discrete actors as follows.

$$J^{di}(\alpha^{di}) = \pi_{\phi^{di}}(s_t)^T \left[-\alpha^{di}\left(\log\left(\pi_{\phi^{di}}(o_t^i)\right) + \bar{H}\right)\right] \quad (18)$$

A hypothetical architecture of the proposed MASAC framework based on the above descriptions is shown in Fig. 1.

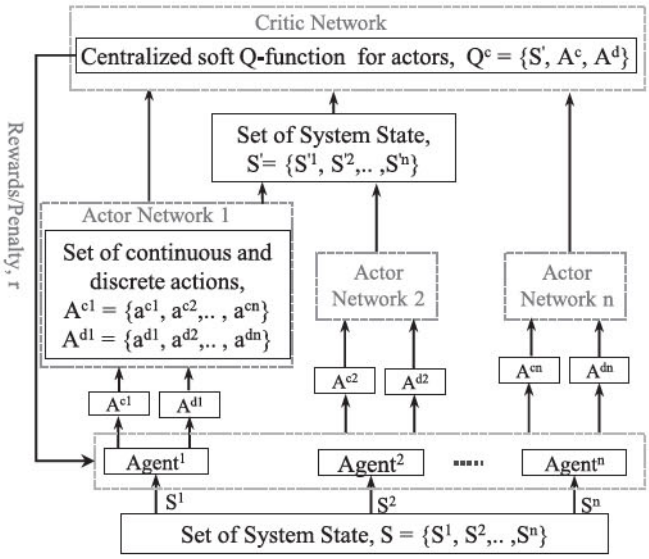

Fig. 1. Architecture of the proposed MASAC framework.

IV. Implementation of the MASAC Framework to Enhance Power System Resilience

To implement the proposed approach for power system resilience enhancement, we need to construct scenarios under extreme events to train and execute the MASAC framework. This section describes implementation procedures of the proposed resilience enhancement algorithm.

A. Resilience Enhancement

An approach to construct power system component failure scenarios and a resilience index are necessary to demonstrate power system resilience enhancement using the proposed approach. As the main focus of this paper is to develop a resilience enhancement strategy, instead of proposing new resilience indices and approaches to construct scenarios, an existing scenario construction approach from [32] and a resilience index from [33] are used in this study. The adapted scenario construction approach and resilience index are described as follows.

1) Line Outage Scenarios Based on Fragility Curves: Although the proposed approach is applicable to enhance the resilience of power systems under any extreme event, in this work, we use line failures due to high wind speed as an example of extreme events. A fragility curve from [32] is used to construct line failure scenarios based on wind speed.

The failure probability of a transmission line based on the adapted fragility curve can be expressed as follows.

$$P_l(\omega) = \begin{cases} \bar{P}_l, & \text{if } \omega_s < \omega_s^{critical} \\ P_{lh}, & \text{if } \omega_s^{critical} \le \omega_s \le \omega_s^{collapse} \\ 1, & \text{otherwise} \end{cases} \quad (19)$$

where $P_l(\omega)$ represents a function to determine failure probabilities of lines in terms of wind speed; $\bar{P}_l$ is the failure probability of a line under normal wind speed that is assumed as 1×10^{-2} [32]; and the failure probability between wind speeds of $\omega_s^{critical}$ and

$\omega_s^{collapse}$ follows a linear relationship. The value of $\omega_s^{critical}$ and $\omega_s^{collapse}$ are assumed as 30 m/sec and 55 m/sec, respectively.

2) Resilience Index: A resilience index "Survivability" from [33] is used in this work to measure the resilience. Survivability can be defined as the ability of a power system to supply the maximum amount of loads without compromising the most critical loads during contingencies. As the main focus of this paper is to develop a method to avoid load shedding during multiple contingencies, the "Survivability" is represented by a binary string—'1' indicates the power system is capable to feed the entire load demand and '0' indicates that the system fails to satisfy its entire demand.

3) Training and Execution Algorithms for the MASAC Framework: To train and execute the MASAC framework for resilience enhancement, first, the power grid is partitioned into several regions based on the electrical distance where each region is controlled by an agent. The users/power system operators have the privilege to select the number of agents for the MASAC framework depending on system sizes. The voltage of a region during contingencies is regulated within predefined limits by control actions of an agent. The control actions of an agent based on the provided input states of its region can be expressed as follows.

$$a_t^{ci} = \begin{cases} \pi_{\phi^{ci}}(a_t^{ci}|o_t^i), & \text{if } \mid \Lambda_t^i \mid > 0 \\ a_{t-1}^{ci}, & \text{if } \mid \Lambda_t^i \mid = 0 \end{cases} \tag{20}$$

and,

$$a_t^{di} = \begin{cases} \pi_{\phi^{ci}}(a_t^{di}|o_t^i), & \text{if } \mid \Lambda_t^i \mid > 0 \\ a_{t-1}^{di}, & \text{if } \mid \Lambda_t^i \mid = 0 \end{cases} \tag{21}$$

where $\mid \Lambda_t^i \mid$ represents the number of violated bus that exist in the respective region of agent i; the continuous action, a_t^{ci}, represents the amount of the dispatched shunt reactive power for agent i; and the discrete action, a_t^{di}, represents the dispatch Status of the shunts for agent i. A replay buffer is used to train all the agents, which is expressed as follows.

$$\mathcal{D} \leftarrow \left(s_t, o_t^i, a_t^{ci}, a_t^{di}, a_t^{-ci}, a_t^{-di}, r^t, s_{t+1}, o_{t+1}^i, a_{t+1}^{ci}, a_{t+1}^{di}, a_{t+1}^{-ci}, a_{t+1}^{-di}\right) \tag{22}$$

The training and testing/execution algorithms for the MASAC framework to enhance the resilience under multiple line failures due to high wind speeds are summarized in algorithm 1 and algorithm 2, respectively.

Algorithm 1: Training Algorithm for the MASAC Framework.

1: **for** episode $= 1$ to M **do**
2: Construct a lines outage scenario using a fragility curve.
3: Perform initial power flow for the constructed scenario and send o_i^t and s_t to each agent.
4: count $\mid \Lambda_t^i \mid$.
5: **while** voltages violate and step $< N$ **do**
6: Calculate both continuous and discrete actions, a_t^{ci} and a_t^{di} for each agent using (20) and (21).
7: Execute actions a_t^{ci} and a_t^{di} in environment using a power flow solver (e.g., Pypower).
8: Observe s_{t+1} and r_t to check terminal conditions.
9: Store $(s_t, o_t^i, a_t^{ci}, a_t^{di}, a_t^{-ci}, a_t^{-di}, r_t, s_{t+1})$ in replay buffer $\mathcal{D}_\rangle$ based on (22).
10: If s_{t+1} is terminal reset the environment.
11: Update weights of continuous and discrete policies of each agent using (14) and (17), respectively.
12: Update the Q-function parameters of local and target networks of each agent using (10).
13: Update temperature of actor-networks of each agent using (15) and (18), respectively.
14: Update target networks weights of each agent using $\bar{Q}_m \leftarrow \tau Q_m + (1-\tau)\bar{Q}$, where, $m \in \{1, 2\}$ and $m \ll 1$.
15: **end while**
16: **end for**

Algorithm 2 Testing Algorithm for the MASAC Framework

1: **for** episode $= 1$ to M **do**
2: Construct a lines trip scenario using the fragility curve.
3: Perform initial power flow for the constructed scenario and send o_i^t and s_t to each agent and count $\mid \Lambda_t^i \mid$.
4: **while** voltages violate and step $< N$ **do**
5: Calculate both continuous and discrete actions, a_t^{ci} and a_t^{di} for each agent using (20) and (21).
6: Execute actions a_t^{ci} and a_t^{di} in environment using the power flow solver.
7: Observe s_{t+1} and r_t to check terminal conditions.
8: **end while**
9: **end for**

V. Training and Execution of the MASAC Framework

To train and execute the proposed MASAC framework, contingency scenarios for a standard test system (environment) are constructed and power flow for each scenario is performed using the Pypower [34]. Each agent contains continuous and discrete actors, and all the agents have a centralized critic. At the beginning of a scenario, discrete actors select locations whereas continuous actors select the output power simultaneously. Then, the discrete (locations) and continuous (output power) actions are provided to the critic network to update both actions based on agents' policies and termination criterion of the algorithm. Also, a centralized replay buffer is used to store the information of all agents. In algorithm 1 and algorithm 2, M and N represent sizes of the training data set (total number of episodes) and the maximum number of iterations in each episode, respectively. The size of the training dataset (total number of scenarios) needs to be large enough to capture extensive operation Status of power systems. On the other hand, the maximum number of iterations should not be too large to avoid the negative impacts on training due to consequential transitions with ineffective actions. The

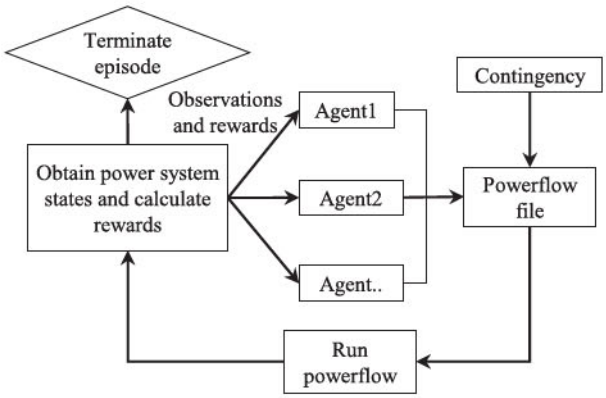

Fig. 2. The flow of information during training of the agents.

process to flow information during training of the agents of the MASAC is shown in Fig. 2.

The detailed training and execution process of the proposed MASAC to enhance resilience is as follows.

- **Step 1.** Power flow is solved for the environment (failure scenario of power system) at the beginning of each episode to obtain initial grid states (bus voltage magnitudes). Then, grid states are divided based on the predefined regions. After this point, the states of a region is fed as input to the assigned agent for the respective region. If an agent detects voltage violations in its region, then the observations are extracted. Otherwise, move to the next episode (i.e., redo **step 1**).
- **Step 2.** If an agent does not find voltage violations, then that agent maintains original actions in the respective region. If an agent detects a violation, then the agent executes new actions in the respective region using (20) and (21). Then, power flow is performed for the modified environment to obtain new system states. According to the obtained new states, the reward and new observations of each agent are calculated and extracted, respectively.
- **Step 3.** Each agent stores the transitions in the centralized replay buffer. Periodically, the actor and critic networks are updated in turn with a randomly sampled minibatch.
- **Step 4.** Along with the training, each agent keeps reducing the noise to decrease the exploration probability. If one of the episode termination conditions is satisfied, store the information and go to the next episode (i.e., redo Step 1).

The above closed-loop process continues for all of the episodes in the training dataset. For each episode, the training process terminates when one of two conditions is satisfied: i) violation cleared; ii) the maximum number of iterations reached. It does not matter whether voltage violation still exists if the episode is terminated under condition ii). Through the penalization mechanism designed in the reward function, the agents can learn from the experience to avoid providing inaccurate actions. It is worth mentioning here that the proposed MASAC framework requires a centralized communication network to provide actions of all agents to the critic network during training. After receiving these actions, the critic network provides rewards to agents, which are used to update policies of the actor networks. It should be noted that this process can be executed offline without real-time interaction with the system. During testing/execution, well-trained agents of the MASAC framework use only local measurements to provide control commands, which can be checked by grid operators before execution. This decentralized execution regulates the regional voltage without any communication.

VI. Numerical Examples

The proposed approach is demonstrated on the IEEE 57-bus and IEEE 300-bus systems to analyze its effectiveness on different system sizes. The training scenarios are constructed using a adapted fragility curve from [32]. Also, training data for the constructed scenarios are synthetically generated from a feasible power flow solution. The specific generation process for the training data is summarized as follows. First, one or multiple lines are tripped based on the fragility curve to construct contingency scenarios. Variations in the wind velocity are considered between 30–60 m/s to capture both minor (failure of a small number of lines) and major (failure of a large number of lines) contingency scenarios. Then, the power flow solver (Pypower) is used to check whether the given contingency case is solvable or not. Finally, feasible power flow cases are stored as training data. During training, the locations and sizes for shunts are determined using actions of the agents. The case studies for both the IEEE 57-bus and IEEE 300-bus system are as follows.

Case I—Training the MASAC for the IEEE 57-bus system. The IEEE 57-bus system has 57 buses and 80 branches. It is worth mentioning here that the proposed MASAC framework can determine appropriate sizes and locations for shunts based on selected potential locations and minimum and maximum limits of shunts during training. Also, it should be noted that power system planners have the privilege to select potential locations, and minimum and maximum sizes of shunts. In this study, we demonstrated the effectiveness of the proposed framework through providing a snapshot, which is as follows.

To train and execute the MASAC framework, the IEEE 57-bus system is partitioned into six regions. The MASAC is constructed using six agents to provide actions for the six regions. The allocation of bus numbers to different regions are shown in Fig. 3.

We assume that a windstorm passes through the entire system. The selected vulnerable candidate lines are (1–2), (9–12), (18–19), (31–32), (38–44), (50–51), and (56–41). At each iteration of an episode, an agent identifies both size and location (a candidate bus) from its own region to install a shunt. We assume that output reactive power of shunts varies between 0–6 MVAr. Variations in the shunt output power should not be too high to avoid the negative impacts on training due to exploration of a very large action space. Total 20 000 episodes are used to train the MASAC using Algorithm 1. The maximum number of iterations used for each episode is 30. Fig. 4(a) and 4(b) show the number of iterations and reward amounts, respectively, for training episodes.

From Fig. 4(a) and Fig. 4(b), we can see that the action time (number of iterations) decreases while the reward amounts

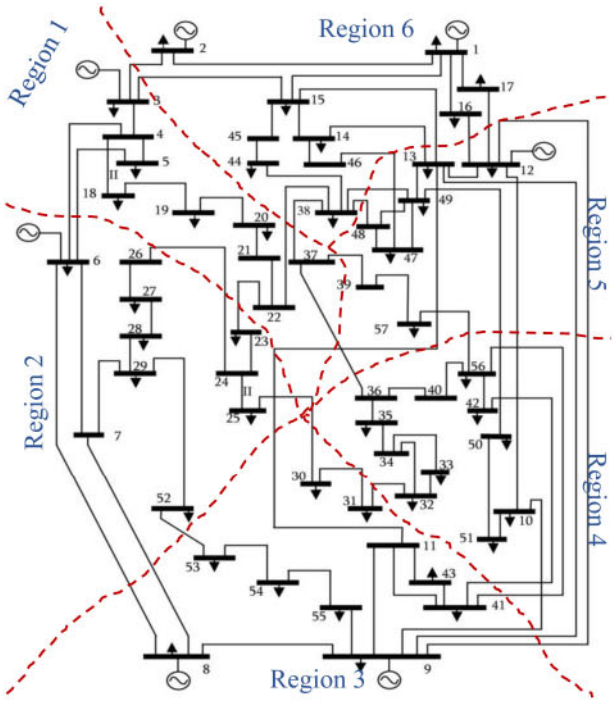

Fig. 3. Allocation of buses to agents for IEEE 57-bus system.

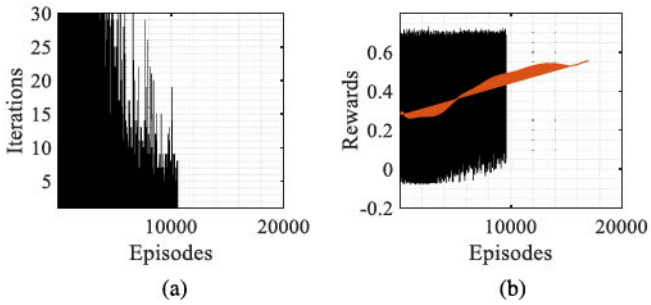

Fig. 4. (a) Required number of iterations (b) amount of rewards of training episodes for the IEEE 57-bus system.

increase with the increase of training episodes. This indicates that the agents are learning from previous experiences to avoid inaccurate actions. For instance, from 1–13 000 episodes, the agents took bad actions and failed to resolve the impacts of contingencies on voltage constraints for a large number of episodes. Reward amounts during this period are also low for a large number of episodes. However, when the agents get trained, impacts of all contingencies on voltage constraints are resolved very quickly (within few iterations) and the reward amount increased significantly.

Fig. 5(a) shows the critic losses for the continuous and discrete actors, which fluctuate at the beginning, and finally converge to equilibrium. Thus, we can say that the MASAC is getting trained to provide actions for removing the negative impacts of multiple line failures on voltage constraints.

Case II—Determining locations and sizes of the shunts using the trained MASAC for the IEEE 57-bus system. In order to determine candidate locations and sizes to install shunts for maintaining voltage limits under multiple line failures, the

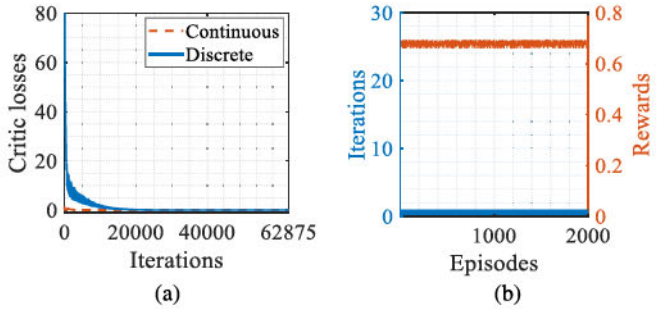

Fig. 5. (a) Losses of critics during training and (b) Required number of iterations and amount of rewards of testing episodes for the IEEE 57-bus system.

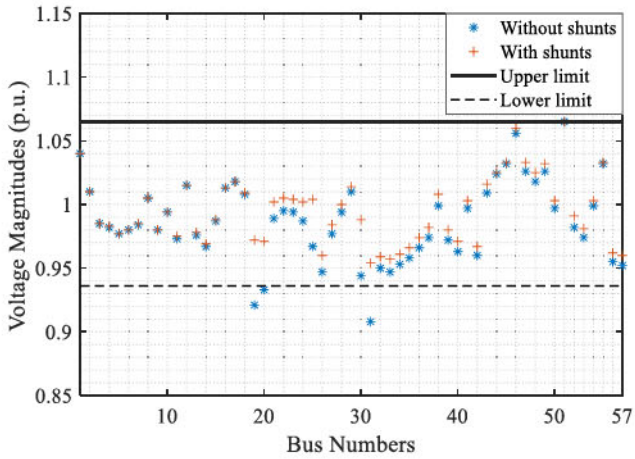

Fig. 6. Status of the bus voltages of IEEE 57-bus system before and after dispatching shunts for outage of all candidate lines.

trained network is tested for 2000 episodes using algorithm 2. The testing scenarios are also constructed using the fragility curve. Fig. 5(b) shows the number of iterations and reward amounts, respectively, for testing episodes. From Fig. 5(b), we can see that the contingencies for the testing episodes are solved within one iteration and the agents get the maximum rewards for each testing episode. This indicates that the trained MASAC framework can remove voltage violations. The calculated sizes of the shunt reactive power compensators to maintain voltage constraints under failures of the selected lines are approximately 3.10, 3.10, 3.10, 0.00, 3.10, and 3.10 MVArs. The determined locations for these shunts are buses number 4, 18, 29, 42, and 51. The shunt reactive power compensators are installed in the IEEE 57-bus system based on the calculated locations and sizes. The accuracy of the calculated locations and sizes are checked for two scenarios–*scenario I (major outage):* outage of all the candidate lines and *scenario II (minor outage):* outage of one candidate line (selected randomly). The status of the bus voltage magnitudes before and after dispatching shunts for both scenarios are shown in Fig. 6 and Fig. 7.

The upper and lower voltage limits of the IEEE 57-bus system during normal operating conditions are 1.065 p.u. and 0.935 p.u., respectively.

From Fig. 6 and Fig. 7, we can see that voltages of three buses are violated for the outages of all candidate lines, whereas voltage of one bus is violated for the outage of one line (31–32).

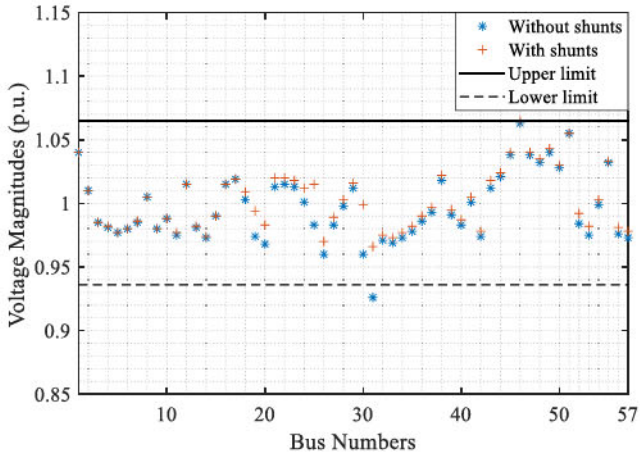

Fig. 7. Bus voltages of the IEEE 57-bus system before and after dispatching shunts for outage of line 31–32.

Fig. 6 and Fig. 7 also show that the voltage violations of both scenarios are resolved successfully using the installed shunts based on the proposed approach. Thus, the proposed approach can be used to determine locations and sizes of shunts to achieve '*Survivability*' of '1' during the failure of multiple lines for the IEEE 57-bus system.

Case III—Training of the MASAC for the IEEE 300-bus System. The IEEE 300-bus system has 300 buses and 411 branches. Similar to *Case I*, the entire system is partitioned into six regions with each region is controlled by an agent. In this case, equal number of buses are assigned chronologically to different agents. In other words, agents 1–6 control voltages at buses 1–50, 51–100, 101–151, 151–200, 201–251, and 251-300, respectively. The 300-bus system is comparatively a large test system, and we assume that windstorm passes through one region. Although the windstorm passes through one region, some lines that connect buses of the affected region with other regions may be tripped. Also, as all regions are interconnected, line failures in one region may affect voltage constraints of other regions. In this study, we assume that the windstorm passes through region1, and ten candidate lines are selected, which may trip depending on their respective failure probabilities and and velocity of wind. The selected candidate lines for this case are (1–5), (3–7), (8–11), (12–21), (14–15), (19–87), (33–40), (35–72), (45–60), and (47–113). Also, shunt values (continuous action space) for different agents are varied as follows: agent1: 0–8 MVAr, agent2: 0–7 MVAr, agent3: 0–6 MVAr, agent4: 0–5 MVAr, agent5: 0–4 MVAr, and agent6: 0–3 MVAr. Each agent identifies a bus from its own region at each iteration to dispatch a shunt. The total number of episodes used to train the MASAC for this case using Algorithm 1 is 15 000. The maximum number of iterations used for each episode is 30. Fig. 8(a) and 8(b) show the number of iterations and total reward amounts, respectively, for each training episode.

Fig. 8(a) and Fig. 8(b) show that the action time decreases while the amount of rewards increases with the increase of episodes. This validates efficient learning of the agents from the previous experiences.

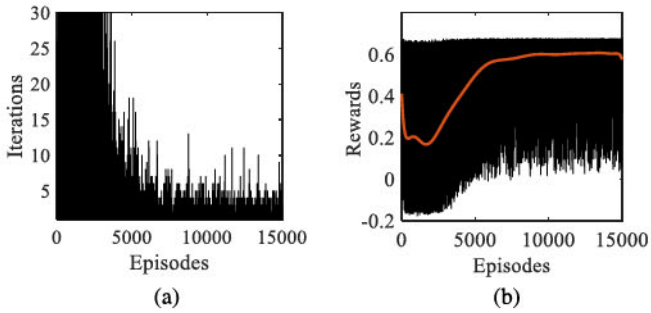

Fig. 8. (a) Required number of iterations and (b) amount of rewards of training episodes for the IEEE 300-bus system.

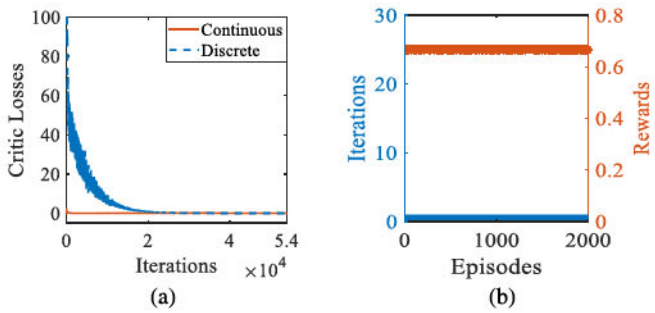

Fig. 9. (a) Losses of critics during training and (b) Required number of iterations and amount of rewards of testing episodes for the IEEE 300-bus system.

Fig. 9(a) shows that the actor and critic losses for the continuous and discrete actors fluctuate at the beginning, and finally converge to equilibrium. This indicates that the agents are learning to provide effective control actions.

Case IV—Determining locations and sizes of shunts using the trained MASAC for the IEEE 300-bus system. Similar to Case II, the trained network for the IEEE 300-bus system is tested for 2000 episodes using algorithm 2. Fig. 9(b) shows the number of iterations and reward amounts, respectively, for testing episodes. From Fig. 9(b), it can be seen that contingencies for the testing episodes are solved within one iteration and the agents get the maximum rewards. Therefore, it can be concluded that the trained MASAC framework can solve voltage violations. The calculated sizes of the shunt reactive power compensators to maintain voltage limits under failures of selected lines are approximately 5.10, 4.10, 3.10, 2.10, 2.00, and 2.10 MVArs. The planned locations for these shunts are buses 30, 83, 125, 185, 203, and 268.

Shunt reactive power compensators are installed in the IEEE 300-bus system based on the calculated locations and sizes. Similar to case II, the accuracy of the calculated locations and sizes are checked for two scenarios: major outage (outage of all the candidate lines) and minor outage (outage of three randomly selected lines) scenarios. The status of the bus voltage magnitudes before and after dispatching shunts for both scenarios are shown in Fig. 10 and Fig. 11. The upper and lower voltage limits of the IEEE 300-bus system during normal operating condition are 1.074 p.u. and 0.929 p.u., respectively.

TABLE II
COMPARISON ANALYSIS OF THE PROPOSED METHOD WITH THE PREVIOUSLY PROPOSED RL-BASED METHOD

Method	Provided Actions Types	Solution Methodology	Control Performance
[23]	Continuous or discrete	1. DDPG based agent: actor + critic + replay buffer; 2. Deterministic policy	1. With random load fluctuations and contingencies applied to the operation data, the DRL agent can fix the voltage violation issues for small systems; 2. Accuracy is approximately 80%.
[24]	Continuous or discrete	1. DDPG based agent: actor + critic + coordinator + independent replay buffer; 2. Operation rule based deterministic policy	1. With random load fluctuations and contingencies applied to the operation data while considering communication limits, the DRL agent can solve voltage violations for maximum 200-bus system. 2. Accuracy is approximately 87%.
Proposed method	Both Continuous & **discrete**	1. SAC-based agent: actor + critic + independent replay buffer; 2. Entropy regularization-based deterministic policy	1. With random load fluctuations and contingencies applied to the operation data while considering communication limits, DRL agents solve voltage violations with improved scalability and regional controllability 2. Accuracy is approximately 92%.

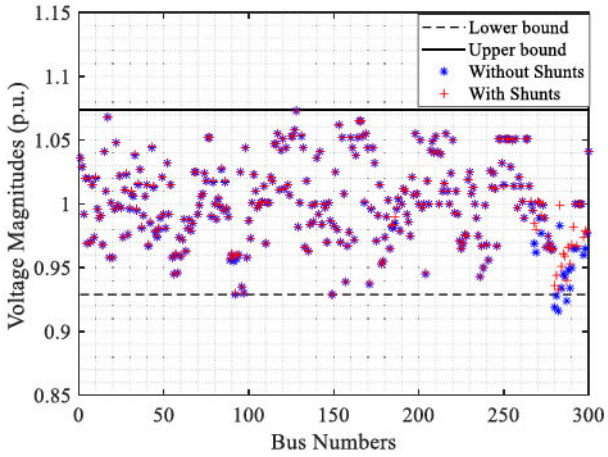

Fig. 10. Bus voltages of the IEEE 300-bus system before and after dispatching shunts for the outage of all candidate lines.

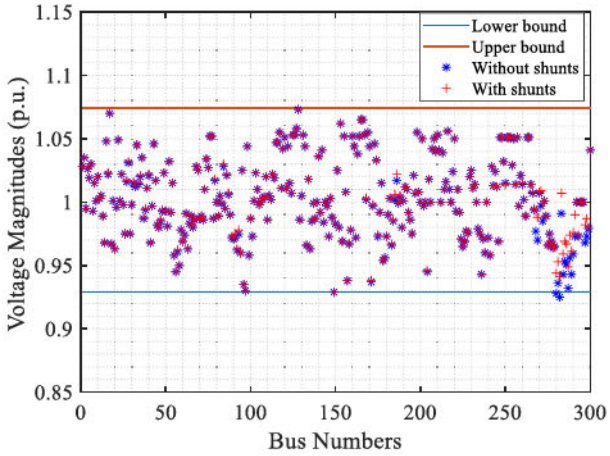

Fig. 11. Bus voltages of the IEEE 300-bus system before and after dispatching shunts for outages of line 45–60, 14–15, and 12–21.

From Fig. 10 and Fig. 11, we can see that the four buses have voltage violations for the major outage, whereas two buses have voltage violations for the minor outage. Fig. 10 and Fig. 11 also show that voltage violations of both scenarios are resolved successfully using the installed shunts based on the proposed approach. Thus, the proposed approach can be used to determine locations and sizes of shunts to achieve '*Survivability*' of '1' during the failure of multiple lines for the IEEE 300-bus system.

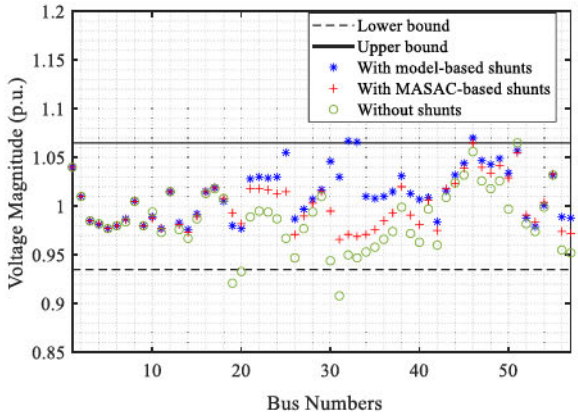

Fig. 12. Comparison between the MASAC method and model-based method for outage of line 31–32 of IEEE 57-bus system.

Case V—Comparison Analysis. The proposed method is compared with a conventional method for shunt planting based on the operating rules (normalized voltage magnitudes of the buses) of the power grids without the need for an accurate model [35]. The simulation is performed on the IEEE 300-bus system. In [35], the candidate nodes to install the shunts are selected based on the normalized voltage magnitudes of the buses. The calculated locations based on the described method in [35] are bus number 26, 30, 31, 32, and 33. We select same shunt sizes for both the methods. Fig. 12 shows the comparison between the obtained results using the MASAC method and conventional method. From Fig. 12, we can see that all the voltages are within limits for the proposed method, while the voltages of three buses violate for the conventional method.

The main drawback of model-based methods is that they cannot provide solutions without accurate system information. Also, it should be noted that the model-based methods lose their effectiveness immediately once tested systems are changed. Under certain extreme conditions, model-based methods may result in a large deviation from the optimal point. For instance, the proposed DRL-based method can easily handle outages of multiple lines, which is quite challenging for model-based methods if line capacities or all other system information are not immediately available. Also, it is not straightforward to model a large number of system components such as nonlinear power electronic devices and renewable energy sources, which significantly limits the applicability of model-based methods

for large systems. Therefore, it is challenging or even impossible to obtain the solution using model-based methods for the dynamically changing systems under extreme events (outages of multiple lines). From this perspective, data-driven methods can be a promising option to solve these issues.

The advantage of the proposed work over the previously conducted DRL-based works to control power system voltages is also demonstrated through a comparison analysis in Table II. From Table II, it can be seen that the proposed method has an improved scalability over existing methods. Also, the actor network of each agent of previous methods provide only continuous actions, whereas the actor networks of the MASAC framework provide both the continuous and discrete actions simultaneously. Moreover, as the sizes of power grids increase, the existing works do not have potential to handle high dimensional input-output space for the actor network. On the other hand, the proposed method can solve the curse of dimensionality because each agent of the MASAC framework controls only local region/sub-system.

VII. Conclusion

This paper has proposed a multi-agent distributed computation and implementation framework using a DRL algorithm to explore the benifits of shunt reactive power compensators for power system resilience enhancement against extreme events and multi-component failures. The agents were constructed using a hybrid SAC (HSAC) algorithms. The HSAC was formulated using the fundamental SAC algorithms for both the continuous and discrete actions. To implement the proposed MASAC framework for power system resilience enhancement, a power system was partitioned into several regions where each region is controlled by an agent. Then, the proposed MASAC framework was trained using historical data and fragility curves of transmission lines and was periodically updated to capture changes in system parameters. The trained MASAC framework provided locations (candidate buses) and sizes of the shunts to enhance resilience of power systems under multiple line failures. The proposed approach was demonstrated on the IEEE 57-bus and IEEE 300-bus systems through numerical examples. The results showed that the proposed algorithm is effective to plan for the deployment of shunt reactive power compensators to enhance resilience of power grids under multiple line failures.

References

[1] G. Huang, J. Wang, C. Chen, J. Qi, and C. Guo, "Integration of preventive and emergency responses for power grid resilience enhancement," *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4451–4463, Nov. 2017.

[2] H. Farzin, M. Fotuhi-Firuzabad, and M. Moeini-Aghtaie, "Enhancing power system resilience through hierarchical outage management in multi-microgrids," *IEEE Trans. Smart Grid*, vol. 7, no. 6, pp. 2869–2879, Nov. 2016.

[3] K. P. Schneider *et al.*, "A distributed power system control architecture for improved distribution system resiliency," *IEEE Access*, vol. 7, pp. 9957–9970, 2019.

[4] M. Yan *et al.*, "Enhancing the transmission grid resilience in ice storms by optimal coordination of power system schedule with pre-positioning and routing of mobile DC DE-icing devices," *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 2663–2674, Jul. 2019.

[5] X. Zeng, Z. Liu, and Q. Hui, "Energy equipartition stabilization and cascading resilience optimization for geospatially distributed cyber-physical network systems," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 45, no. 1, pp. 25–43, Jan. 2015.

[6] Q. Wang, Z. Yu, R. Ye, Z. Lin, and Y. Tang, "An ordered curtailment strategy for offshore wind power under extreme weather conditions considering the resilience of the grid," *IEEE Access*, vol. 7, pp. 54 824–54 833, 2019.

[7] J. Li *et al.*, "Resilience control of dc shipboard power systems," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6675–6685, Nov. 2018.

[8] Z. Li, M. Shahidehpour, F. Aminifar, A. Alabdulwahab, and Y. Al-Turki, "Networked microgrids for enhancing the power system resilience," *Proc. IEEE*, vol. 105, no. 7, pp. 1289–1310, Jul. 2017.

[9] L. Sun, W. Liu, C. Y. Chung, M. Ding, R. Bi, and L. Wang, "Improving the restorability of bulk power systems with the implementation of a WF-BESS system," *IEEE Trans. Power Syst.*, vol. 34, no. 3, pp. 2366–2377, May 2019.

[10] C. Wang, Y. Hou, F. Qiu, S. Lei, and K. Liu, "Resilience enhancement with sequentially proactive operation strategies," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2847–2857, Jul. 2017.

[11] A. S. Musleh, H. M. Khalid, S. M. Muyeen, and A. Al-Durra, "A prediction algorithm to enhance grid resilience toward cyber attacks in WAMCS applications," *IEEE Syst. J.*, vol. 13, no. 1, pp. 710–719, Mar. 2019.

[12] M. Nazemi, M. Moeini-Aghtaie, M. Fotuhi-Firuzabad, and P. Dehghanian, "Energy storage planning for enhanced resilience of power distribution networks against earthquakes," *IEEE Trans. Sustain. Energy*, vol. 11, no. 2, pp. 795–806, Apr. 2020.

[13] Y. Wang, L. Huang, M. Shahidehpour, L. L. Lai, H. Yuan, and F. Y. Xu, "Resilience-constrained hourly unit commitment in electricity grids," *IEEE Trans. Power Syst*, vol. 33, no. 5, pp. 5604–5614, Sep. 2018.

[14] B. Taheri, A. Safdarian, M. Moeini-Aghtaie, and M. Lehtonen, "Enhancing resilience level of power distribution systems using proactive operational actions," *IEEE Access*, vol. 7, pp. 137 378–137 389, 2019.

[15] Q. Guo, H. Sun, M. Zhang, J. Tong, B. Zhang, and B. Wang, "Optimal voltage control of PJM smart transmission grid: Study, implementation, and evaluation," *IEEE Trans. Smart Grid*, vol. 4, no. 3, pp. 1665–1674, Sep. 2013.

[16] N. Qin, C. L. Bak, H. Abildgaard, and Z. Chen, "Multi-stage optimization-based automatic voltage control systems considering wind power forecasting errors," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1073–1088, Mar. 2017.

[17] H. J. Liu, W. Shi, and H. Zhu, "Distributed voltage control in distribution networks: Online and robust implementations," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6106–6117, Nov. 2018.

[18] H. Zhu and H. J. Liu, "Fast local voltage control under limited reactive power: Optimality and stability analysis," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3794–3803, Sep. 2016.

[19] J. G. Vlachogiannis and N. D. Hatziargyriou, "Reinforcement learning for reactive power control," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1317–1325, Aug. 2004.

[20] Y. Xu, W. Zhang, W. Liu, and F. Ferrese, "Multiagent-based reinforcement learning for optimal reactive power dispatch," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 42, no. 6, pp. 1742–1751, Nov. 2012.

[21] H. Xu, A. D. Domínguez-García, 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, May 2020.

[22] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2313–2323, May 2020.

[23] 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.

[24] 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.

[25] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor," CoRR, vol. abs/1801.01290, 2018. [Online]. Available: http://arxiv.org/abs/1801.01290

[26] P. Christodoulou, "Soft actor-critic for discrete action set-tings," *CoRR*, vol. abs/1910.07207, 2019. [Online]. Available: http://arxiv.org/abs/1910.07207

[27] A. Singhal, V. Ajjarapu, J. Fuller, and J. Hansen, "Real-time local volt/var control under external disturbances with high PV penetration," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3849–3859, Jul. 2019.
[28] G. Qu and N. Li, "Optimal distributed feedback voltage control under limited reactive power," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 315–331, Jan. 2020.
[29] Z. Fan, R. Su, W. Zhang, and Y. Yu, "Hybrid Actor-Critic Reinforcement Learning in Parameterized Action Space," CoRR, vol. abs/1903.01344, 2019. [Online]. Available: http://arxiv.org/abs/1903.01344
[30] T. Haarnoja *et al.*, "Soft actor-critic algorithms and applications," CoRR, vol. abs/1812.05905, 2018.
[31] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," CoRR, vol. abs/1802.09477, 2018.
[32] M. Panteli, C. Pickering, S. Wilkinson, R. Dawson, and P. Mancarella, "Power system resilience to extreme weather: Fragility modeling, probabilistic impact assessment, and adaptation measures," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3747–3757, Sep. 2017.
[33] A. Hussain, V. Bui, and H. Kim, "Resilience-oriented optimal operation of networked hybrid microgrids," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 204–215, Jan. 2019.
[34] R. Lincoln. Pypower. [Online]. Available: https://pypi.org/project /PYPOWER/
[35] P. V. Babu and S. P. Singh, "Capacitor allocation in radial distribution system for maximal energy savings," in Proc. *Nat. Power Syst. Conf.*, Dec. 2016, pp. 1–6.

Md. Kamruzzaman (Member, IEEE) received the B.Sc. degree in electrical and electronic engineering from Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh in 2011, the M.Sc. degree in electrical engineering from LAMAR University, Beaumont, TX, USA, in 2016, and the Ph.D. degree from the University of Nevada, Reno, NV, USA, in 2020. He is currently a Postdoctoral Research Associate with Rensselaer Polytechnic Institute, Troy, NY, USA. Prior to joining RPI, he was an Intern with GEIRI North America, San Jose, CA, USA and Electric Power Research Institute, Washington, DC, USA. His research interests include power system reliability, resiliency, and monitoring.

Jiajun Duan (Member, IEEE) received the B.S. degree in power system and its automation from Sichuan University, Chengdu, China, the M.S. and Ph.D. degrees in electrical engineering with Lehigh University, Bethlehem, PA, USA, in 2015 and 2018, respectively. From 2018 to 2020, he was a Research Engineer with GEIRINA Inc., San Jose, CA, USA. He is currently with Nextracker Inc., Fremont, CA, USA, as a Senior Power System Engineer. His research interest includes power system, power electronics, control systems, and artificial intelligence.

Di Shi (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Arizona State University, Tempe, AZ, USA, in 2012. He is the Founder of AINERGY LLC, Santa Clara, CA, USA. He was the Director of fundamental R&D Center and Department Head of AI & System Analytics, GEIRINA, San Jose, CA, USA, and a Research Staff Member with NEC Laboratories America, Princeton, NJ, USA. His research interests include data analytics, energy storage systems, and applications of AI and IoT in power systems. He is the Editor of the IEEE TRANSACTIONS ON SMART GRID and the IEEE POWER ENGINEERING LETTERS.

Mohammed Benidris (Senior Member, IEEE) received the BSc. and MSc. degrees in electrical engineering from the University of Benghazi, Benghazi, Libya, and the Ph.D. degree in electrical engineering from Michigan State University, East Lansing, MI, USA. He is currently an Assistant Professor of electrical engineering with the University of Nevada, Reno (UNR), Reno, NV, USA. Prior to joining the UNR, he was Research Associate and Visiting Lecturer with Michigan State University, an Assistant Lecturer with the Department of Electrical and Electronics Engineering, University of Benghazi, and an Engineer with General Electric Company-Libya, Tripoli, Libya. He has more than five years of industry and consulting experience ranging in power plants control and operation to hardware design and installation, and total of more than ten years of academic experience. His main research interests include power system reliability and stability, and resilience evaluation and enhancement of cyber-physical energy systems.