IEEE Open Access Journal of
Power and Energy

Received 14 July 2023; revised 31 October 2023; accepted 6 November 2023.
Date of publication 13 November 2023; date of current version 29 November 2023.

Digital Object Identifier 10.1109/OAJPE.2023.3331884

Real-Time Excitation Control-Based Voltage Regulation Using DDPG Considering System Dynamic Performance

YULING WANG (Student Member, IEEE) AND VIJAY VITTAL (Life Fellow, IEEE)

School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA
CORRESPONDING AUTHOR: V. VITTAL (vijay.vittal@asu.edu)

ABSTRACT In recent years, there has been an increasing need for effective voltage control methods in power systems due to the growing complexity and dynamic nature of practical power grid operations. This paper proposes a real-time voltage control method based on deep reinforcement learning (DRL) that continuously regulates the excitation system in response to system disturbances. Dynamic performance is considered during control by incorporating the voltage dynamics data that influence the practical power grid operation. The proposed approach utilizes the deep deterministic policy gradient (DDPG) algorithm, capable of handling continuous action spaces, to adjust the voltage reference of the generator excitation system in real time. To analyze the power system dynamic process, a versatile transmission-level power system dynamic training and simulation platform is developed by integrating the power system simulation software PSS/E and a user-written DRL agent code developed in Python. The platform facilitates the training and testing of various power system algorithms and power grids in dynamic simulations. The efficacy of the proposed method is evaluated based on the developed platform through extensive case studies on the IEEE 9-bus system and the Texas 2000-bus system. The results validate the effectiveness of the approach, highlighting its promising performance in real-time control with respect to dynamic processes.

INDEX TERMS Voltage control, deep reinforcement learning, DDPG, power system dynamic control, real-time, excitation control.

I. INTRODUCTION

POWER system voltage stability is critical to the reliable operation of the system. With the increasing integration of utility-scale renewable energy and distributed energy resources, the power system variability has further increased due to the nonlinearity and unpredictable consumer patterns of these new types of resources and loads. These factors enhance the chances of power system dynamic instability and pose severe challenges to real-time voltage control [1].

Excitation system control is of significant importance in maintaining generators' voltages and can impact power system dynamic stability directly [2]. Excitation control is considered to be one of the most economical and effective methods for maintaining voltage and dynamic performance enhancement [3]. Numerous excitation control methods have been conducted in terms of voltage regulation considering system dynamic stability after disturbances. A decentralized nonlinear voltage controller was proposed in [4] to achieve both voltage regulation and system stability improvement. Global control(GC) where a stable controller is used for the fault period and a voltage controller is activated for voltage level regulation in [5]. Different controllers need to be switched at different operating stages to guarantee a satisfactory voltage level and system dynamic performance. Lyapunov-function-based methods can achieve voltage regulation and dynamic stability control simultaneously by designing the excitation control without switching [6]. A Lyapunov-based decentralized control (LBC) was proposed in [7] to enhance power system dynamic performance by simultaneously controlling the excitation and governor systems. The time-derivative of the Lyapunov function is designed by the feedback control of synchronous generators, and voltage deviation is considered as the feedback variable to realize voltage regulation as well as dynamic performance improvement. The majority of these model-based methods have been claimed to achieve promising performance.

However, they rely heavily on accurate information of power system topology and parameters. Furthermore, power systems are experiencing uncertainties of load changes and contingencies and it is quite challenging to apply the above model-based methods. Therefore, a voltage regulation method that is flexible and scalable to the application and operational uncertainties needs to be developed.

Artificial intelligence (AI) techniques have matured and are now being applied to various power system applications [8], [9], [10], [11], [12], [13]. These data-driven, model-free methods [14] are particularly well-suited for highly nonlinear and high-dimensional power systems, especially with the availability of phasor measurement units (PMUs) that enable the synchronized transfer of dynamic data across the grid. Advanced control schemes for enhancing power system stability based on AI methods have been developed, and the recent success of reinforcement learning (RL) has shown promise in addressing various power system challenges. An RL agent can be trained to respond instantaneously to a range of system operating conditions based on knowledge obtained by interacting with the power system environment during the training process. Therefore, a real-time application based on RL is possible. Q-learning, a conventional RL method, has been utilized in [15] and [16] to learn a reactive power optimal control scheme and keep the voltage within the normal range. Q-learning was also adopted in [17] for optimal tap setting of on-load tap changers of step-down transformers (connecting electric distribution systems with the rest of the system) to control the distribution system side voltages under uncertain load dynamics. Reference [18] proposed a control scheme of active power generations to prevent system cascading failure based on Q-learning, the controller operates in the system's normal state and takes actions in the form of preventive control to make adjustments in case of cascading failure when the system suffers large disturbances. However, conventional RL methods only work in environments with discrete and finite state and action spaces and thus are not suitable for large, complex problems, such as real-time control problems for large-scale power systems. To overcome this disadvantage, deep reinforcement learning (DRL) has been developed by researchers, which utilizes powerful deep neural networks as function approximators that enable high-dimensional feature extraction. Reference [19] proposed a two-time-scale voltage control scheme, including fast inverter control and switching of shunt capacitors at a slower time control based on the Deep Q-Network (DQN) algorithm. Reference [14] applied DQN and Deep Deterministic Policy Gradient (DDPG) for subsystem voltage control and found that DDPG performed better with sufficient training scenarios. The voltage set point of a STATCOM is regulated using SARSA to facilitate discrete reactive power injection for voltage control in [20]. ESS, PV, and SVC output power levels are managed with the SAC algorithm to mitigate voltage violations in [21] where predefined discrete power levels are used for voltage control. Active power security correction control is implemented using TD3 in [22], while autonomous line flow control is achieved through PPO [23]. Reference [24] combined multiple types of equipment, including transformers and switched shunts, to realize voltage regulation based on the DDPG algorithm. Reference [9] adopted multi-agent deep deterministic policy gradient (MADDPG), which is a multi-agent continuous actor-critic-based algorithm, to realize voltage regulation among different regional zones based on power flow data. However, these works focused on the steady-state performance of the system, ignoring the influence of the dynamic behaviors in the transient process when subjected to a disturbance.

References [25] and [26] addressed transient stability issues to keep the system in synchronism by controlling power system components, such as wind turbines and generators. Approximate Dynamic Programming (ADP) is used in [25] to optimize the closed-loop performance of a wind-integrated power grid by providing supplementary damping control. Another study [26] proposed a wide-area control architecture that includes a local supervised PSS control and an RL-based global wide-area control, which locally damps and inter-area oscillations while prioritizing local signals. In [27], the authors used DRL methods to implement dynamic braking and under-voltage load shedding for power system emergency control. While these methods have been tested on the IEEE 39-bus system or the 68-bus system, practical regional power grids are larger and more complex, which need significant information exchange between RL agents and the power grid environment, especially considering the dynamic performance and real-time control application. To address these challenges, this paper proposes a real-time voltage control method that continuously regulates the excitation system based on DRL. The dynamic performance attributes are considered to include dynamic stability factors that may influence power system operation in practical power grids. The voltage control function is achieved by adjusting the generators' excitation system under system disturbances. The DDPG algorithm, which deals with continuous action spaces, is used in this paper to continuously control the voltage reference of the generator excitation system. To focus on the dynamic process, a transmission-level dynamic power system training and simulation platform is built based on the commercial power system software package PSS/E and user-written code in Python. By using DRL, this paper proposes a controller that allows generators to change their reactive power output within specified limits in real time, enabling the system to satisfy operational requirements and provide voltage support in response to disturbances or load changes. The main contribution of this paper includes:

1) A novel real-time voltage control method based on DRL is proposed, which not only regulates and controls the voltage but also considers the dynamic performance of the power system after the control implementation. By leveraging DRL algorithms, the proposed method achieves improved dynamic performance, addressing the challenges of practical power grids characterized

by large size, complexity, and real-time control requirements.

2) A transmission level power system dynamic training and testing platform is built in this study using a combination of a commercial power system software package PSS/E and a user-written DRL agent code developed in Python. This platform provides a versatile environment that enables the training and testing of various power system algorithms in different power grid environments. The platform supports different scenarios that enable the simulation of various system conditions.
3) A large-scale power system is tested and verified based on the dynamic training and testing platform to investigate the control performance for large power grids. The platform's ability to handle large and complex dynamic power system environments further ensures the practicality and effectiveness of the tested methods in real-world scenarios.

The paper is organized as follows: Section II formulates the problem and introduces the relevant analytical background. In Section III, the proposed DRL-based dynamic voltage control method is described in detail. The design of different reward functions is discussed. The power system dynamic training and simulation platform is discussed in Section IV. Section V demonstrates the simulation results for the IEEE 9-bus system and the 2000-bus Texas synthetic test system. Finally, concluding remarks are provided in Section VI.

II. PROBLEM STATEMENT AND DEEP REINFORCEMENT LEARNING

A. POWER SYSTEM DYNAMIC OPERATION WITH EXCITATION SYSTEM

A dynamical system is a complex system where the behavior evolves over time, and the power system is an example of such a system. It involves interactions between subsystems with an enormous number of variables that are constantly changing during operation. Thus, the dynamic process of power grid operation possesses a highly non-linear characteristic, which is essentially a process of sequential decision-making. In the event of a disturbance, it becomes essential to take appropriate control measures to ensure optimal control while considering power system stability, control cost, and variation of the dynamic variables of the power grid. This decision-making process can be described as a Markov decision process (MDP) [27] and solved by DRL algorithms, which will be discussed in more detail in Section III.

As for the action for the control of the excitation system, numerous parameter setting methods have been discussed by researchers. However, the parameters are usually set as a constant before the generators are put into operation, which results in inflexibility and underutilization of reactive power [28]. To address this issue, DRL can be implemented to continuously optimize the excitation system parameters in real-time during system operation. This allows the DRL algorithm to interact with the power system environment, exchange information, and learn the control policy of highly non-linear power systems without requiring detailed power grid model information.

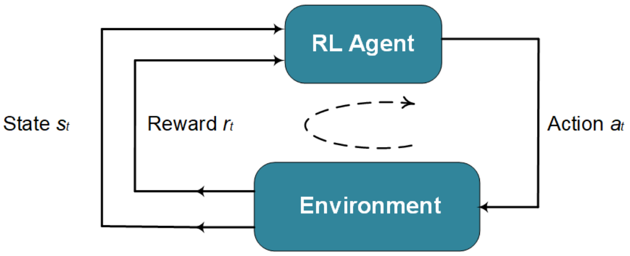

FIGURE 1. **Interaction between RL agent and environment.**

B. REINFORCEMENT LEARNING

RL agent learns by interacting with the environment and making sequential decisions through a trial-and-error process. During the training process, the learned policy is continuously evaluated to guide the agent toward adjusting its control policy in the right direction. The RL agent aims to maximize the value of a reward function that is carefully designed to capture the objectives of the task. The agent explores different actions and extracts information about the state representations of the environment in real-time or through simulation to achieve this goal. If an action results in an increase in the reward value, the agent reinforces the trend of the action; otherwise, the action is attenuated. By adding various event scenarios to the data set, the RL agent can be fully trained to learn a behavior that yields maximum rewards.

The environment follows the Markov Decision Process (MDP). The formulation is defined as a finite MDP [29], M:

$$M \in (S, A, P, R, \gamma) \tag{1}$$

which includes a continuous or discrete state space S and action space A. The environment transition probability P maps a state-action pair at time t to a probability distribution over possible next states. A reward R is given for each state-action pair and a discount factor $\gamma \in [0, 1]$ is used to balance immediate and future rewards.

Figure 1 illustrates the interaction between the RL agent and the environment. At each step t, the agent observes the current state s_t from the environment and selects an action a_t based on its current policy. The agent obtains a reward r_t based on its action and state, and the environment transitions to a new state s_{t+1}. This process is repeated iteratively with the agent continuously updating its policy based on the observed states, actions, and rewards until a preset number of episodes is reached to end the training.

The agent aims to choose the optimal action given the current state to achieve the maximum accumulated discounted reward R_t over time:

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r_i \tag{2}$$

where T is the time step. The key concept in searching for the optimal policy is evaluating the state-value function V and the action-value function Q, which is also known as the Q-function. The state-value function evaluates the goodness of a state for an agent under policy π, as shown in (3)

$$V^{\pi}(s) = E[R_t|s_t = s] \tag{3}$$

The Q-function $Q(s, a)$ represents the expected cumulative future discounted reward for an agent under policy π and estimates the value of performing a certain action a_t in a given state s_t:

$$Q^{\pi}(s_t, a_t) = E[R_t|s_t = s, a_t = a] \tag{4}$$

The Q-function is updated by the recursive relationship in the Bellman equation [30]:

$$Q_{t+1}(s, a) = E[R + \gamma max_{a'} Q_t(s', a')|s, a] \tag{5}$$

The Bellman equation will eventually converge to the optimal solution $Q^*(s, a)$ as the iterations proceed if the states follow the Markov property.

C. DEEP DETERMINISTIC POLICY GRADIENT ALGORITHM

DDPG is a reinforcement learning algorithm that is well-suited for continuous action spaces. It uses an actor-critic structure that concurrently learns a Q-function (modeled by the critic neural network) and a policy (modeled by the actor neural network). To improve the stability of the approach, DDPG utilizes a copied actor neural network and a critic neural network to calculate the target values, which are periodically updated with the weights from the main neural networks to ensure consistency. In total, DDPG includes four networks to estimate the policy and value function: actor, target-actor, critic, and target-critic. Equation (6) is used to update the critic $Q(s, a)$ value.

$$Q_{j+1}^{(s,a)} = Q_j^{(s,a)} + \alpha[R_j + \gamma max Q_j^{(s',a')} - Q_j^{(s,a)}] \tag{6}$$

where α is the learning rate, γ is the discount rate, and $Q_j^{(s',a')}$ represents the target critic neural network.

The control action is obtained from the actor neural network, which enables DDPG to handle a continuous action space in a practical large-scale system. The actor neural network uses a parameterized actor function to determine a deterministic action based on the system states. During training, the policy π is updated in the direction suggested by the critic neural network to maximize the expected reward by taking steps in the direction of $\nabla_{\theta\mu} J$ with respect to the actor parameters. It is formulated as:

$$\nabla_{\theta\mu} J = \frac{1}{N} \sum \nabla_a Q(s, a)|_{s=s_j, a=\mu_{(s_j)}} \nabla_{\theta\mu} \mu(s|\theta^{\mu})|_{s=s_j} \tag{7}$$

where J is the starting distribution, $\mu(s|\theta^{\mu})$ is the parameterized actor function, and θ^{μ} is the policy neural network parameter.

The weights of the target neural networks are periodically updated using a soft update method: $\theta' \leftarrow \rho\theta + (1 - \rho)\theta'$, where ρ is a fraction weight that lies between 0 and 1.

During the action exploration, a decaying noise is added to the policy to improve the agent's ability to explore the range of actions available to solve the environment:

$$\mu'(s_j) = \mu(s_j|\theta_j^{\mu}) + \xi_j \tag{8}$$

where $\xi_{j+1} = r_d * \xi_j$ and r_d is the decay rate.

Both the critic and actor are approximated with parameterized neural networks. The details of the DDPG algorithm can be found in Algorithm 1 [30].

Algorithm 1 Deep Deterministic Policy Gradient algorithm for Real-time Dynamic Voltage Control

```
input  : power system environment states
output: control action applied to the power system
        environment
Initialize the critic network $Q$, $Q'$ and actor network
  $\mu$, $\mu'$ with random weights $\theta, \theta' \leftarrow \theta$ and $\phi$,
  $\phi' \leftarrow \phi$.;
Initialize the experience replay buffer $D$.;
for episode 1 to $M$, do
    Initialize the environment and obtain initial state
      $S_0$;
    Initialize a random process $N$ for action
      exploration;
    for step 1 to $T$, do
        Select action $a_t = \mu(s_t|\theta + N_t)$ according to
          the current policy and exploration noise;
        Execute action $a_t$, observe $r_t$ and next state
          $s_{t+1}$ ;
        Store transition ( $s_t, a_t, r_t, s_{t+1}$) in $D$;
        Sample a random minibatch of $B$ transition (
          $s_j, a_j, r_j, s_{j+1}$) from $D$;
        Compute the critic target:
              $y_j = R_j + \gamma Q'(s_{j+1}, \mu'(s_{j+1}|\theta^{\mu'})|\theta')$
        Update the critic Q-function by gradient
          descent using:
              $L = 1/N \sum_j (y_j - Q(s_j, (a_j|\theta^Q))^2$
        Update the target networks as:
        $\nabla_{\theta\mu} J =$
          $\frac{1}{N} \sum \nabla_a Q(s, a)|_{s=s_j, a=\mu_{(s_j)}} \nabla_{\theta\mu} \mu(s|\theta^{\mu})|_{s=s_j}$
        Update the network parameters:
              $\theta' \leftarrow \rho\theta + (1 - \rho)\theta'$,
              $\phi' \leftarrow \rho\phi + (1 - \rho)\phi'$ ;
```

III. REINFORCEMENT LEARNING-BASED VOLTAGE CONTROL

A. DEFINITION OF ACTION, STATE AND OBSERVATION

Voltage magnitudes are commonly used to represent the operating condition of a power system in reactive power and voltage control problems, since other electrical statuses in system operation can be appropriately reflected in the voltage change [9], [19], [31]. Partial states in DRL algorithms can still work well for streaming valuable information, allowing for flexibility in data measurement and communication [9]. Thus, this paper adopts bus voltage magnitudes as the observation states in the Markov decision process.

The control actions are defined as a vector of excitation system voltage reference values of the controlled generators. Each element of this vector is updated continuously. The value of the voltage references of the excitation system is adjusted within a predefined range of minimum and maximum values in considering the reactive power regulation capacity of each generator.

B. DEFINITION OF REWARD

1) CONSIDER VOLTAGE MAGNITUDE DEVIATION AND REGULATION COST

The reward function r_t is designed to evaluate the effectiveness of the control actions at each training step. To restore the voltage level under the control of the DRL agent, the reward is designed to motivate the agent to reduce the deviation of the observed bus voltage magnitude from the reference value V_{ref}. As shown in (9), if the system diverges after applying the control action, a significant negative reward will be imposed. Otherwise, with less bus voltage deviation, a smaller negative value will be added to the reward at each training step according to the first term of (9) in the case of system convergence. This results in a larger accumulated reward after each training episode composed by a predefined amount of steps. The reward function will gradually guide the agent to regulate its actions to reach better states. To enhance the efficiency of the learning process, the second term associated with the control action is added and aims to direct the agent in generating excitation control commands around the reference value a_{ref}, especially when the initial random exploratory actions deviate too far from a_{ref} during training, as such impractical deviations can result in inefficiency. c_1 and c_2 are the weights of these two parts, and they are chosen based on the expert knowledge of the system as well as trial and error selection [27]. The definition of Δv and Δa can be seen in (10)-(11).

$$r_t = \begin{cases} \textit{Huge penalty}, & \textit{power system diverges} \\ -c_1 * \sum_i \Delta v_i(t) - c_2 * \sum_j \Delta a_j(t), & \textit{otherwise} \end{cases} \tag{9}$$

$$\Delta v_i(t) = \left| v_i(t) - V_{ref} \right| \tag{10}$$

$$\Delta a_j(t) = \left| a_j(t) - a_{ref} \right| \tag{11}$$

2) CONSIDER VOLTAGE MAGNITUDE DEVIATION, REGULATION COST AND HISTORICAL VOLTAGE DATA

Power systems possess significant inertia. The dynamic process of system operation is sequential, which means the current state of the system is affected by both the control actions as well as the previous system states. Significant information lies in the massive historical state data for an operating power grid or a given simulation. For voltage control problems, historical information can be provided by observing the history of bus voltage magnitudes. Therefore, the historical voltage magnitude data is added to the input to help the DRL agent learn a more accurate policy to cope with system disturbances. The reward function considering the historical data is formulated as (12):

$$r_t = \begin{cases} \textit{Huge penalty}, \quad \textit{power system diverges} \\ -c_1 * \sum_i \Delta v_i(t) - c_2 * \sum_j \Delta a_j(t) - \\ c_3 * \sum_{t-c_t}^{t} \sum_i \Delta v_{h-i}(t), \quad \textit{otherwise} \end{cases} \tag{12}$$

$$\Delta v_{h-i}(t) = \left| v_{h-i} - V_{ref} \right| \tag{13}$$

where Δv_{h-i} is the historical voltage magnitude difference of bus i with bus reference value V_{ref}, c_t is the historical time range considered for a certain past time during system operation, and c_3 is the weight related to the historical data in the reward function.

3) CONSIDER VOLTAGE MAGNITUDE DEVIATION, REGULATION COST, HISTORICAL VOLTAGE DATA, AND VOLTAGE RATE OF CHANGE

During the system's dynamic evolution and control implementation after a disturbance or load change, the dynamic performance is also of significant importance. In order to avoid system oscillations and voltage fluctuations so as to facilitate the system voltage recovery in a more stable fashion, both the rates of voltage changes and their historical values are considered in the reward function (14) to guide the agent to generate a control policy that is able to aid in the recovery of the system voltage with more desirable dynamic performance. The reward function considering both voltage historical data and voltage rate of change is shown as (14):

$$r_t = \begin{cases} \textit{Huge penalty}, \quad \textit{power flow diverges} \\ -c_1 * \sum_i \Delta v_i(t) - c_2 * \sum_j \Delta a_j(t) - \\ c_3 * \sum_{t-c_t}^{t} \sum_i \Delta v_{h-i}(t) - \\ c_4 * \sum_{t-c_t}^{t-\Delta t} \sum_i \frac{v_{h-i}(t) - v_{h-i}(t-\Delta t)}{\Delta t}, \quad \textit{otherwise} \end{cases} \tag{14}$$

where c_4 is the weight related to the rate of voltage change in the reward function, Δt is the time interval of every learning step in the training process. When applied to a practical power system, Δt could be the data sampling time step of the measurement device.

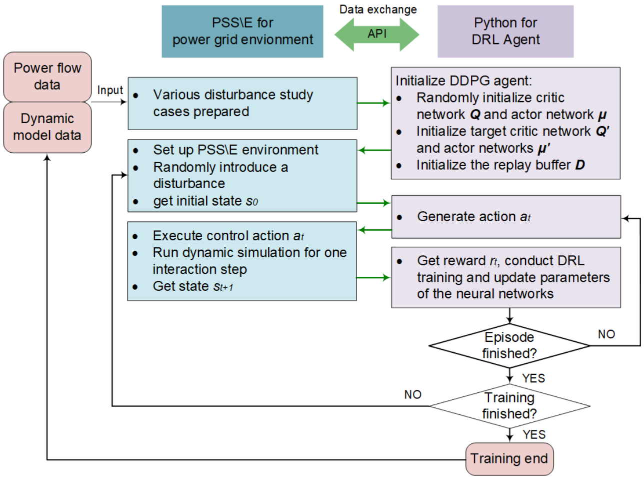
PSS\E for
power grid envionment
Data exchange
API
Python for
DRL Agent
Power flow data
Dynamic model data
Input
Various disturbance study cases prepared
Initialize DDPG agent:
Randomly initialize critic network *Q* and actor network *μ*
Initialize target critic network *Q′* and actor networks *μ′*
Initialize the replay buffer *D*
Set up PSS\E environment
Randomly introduce a disturbance
get initial state s_0
Generate action a_t
Execute control action a_t
Run dynamic simulation for one interaction step
Get state s_{t+1}
Get reward r_t, conduct DRL training and update parameters of the neural networks
Episode finished?
NO
YES
Training finished?
NO
YES
Training end

FIGURE 2. Simulation platform for training DRL algorithm in power system environment.

IV. SIMULATION PLATFORM DEVELOPMENT AND IMPLEMENTATION

A transmission-level power system dynamic simulation and training platform is developed for the training and implementation of the DRL algorithm in the power system dynamic simulation environment. The time-domain simulation software Siemens-PTI PSS/E is used as the power system simulator to conduct power system dynamic simulations and emulate the power grid environment. PSS/E provides application programming interfaces (APIs) with Python, which can communicate the power system simulation environment with the DRL agent in real time to exchange information.

Figure 2 shows the training procedure and the data interaction between the power system simulator and DRL agent in the training platform. The blue and purple blocks represent the actions conducted in PSS/E and Python, respectively. The two software elements constantly exchange information using API in training. The green arrows show the interaction data flow between them. Power flow and dynamic model files are prepared to perform power system dynamic simulation. At the start of the training process, four neural networks with different sets of random weights and the replay buffer size are initialized. For each episode, the power flow is solved, and dynamic simulation is initialized based on the selected study case. The disturbance is randomly introduced, and the initial states are obtained for each training episode, in which one round of dynamic simulation begins. A loop for a predefined number of steps per episode starts with the action generated by the DRL agent. The action then will be sent to the power system simulator and implemented in PSS/E by adjusting the voltage reference input of the excitation system. Then, the dynamic simulation will be run for one training step interval to update the states of the power system environment, and the most updated states are sent back to the DRL agent. The reward will be calculated based on the system observation to evaluate the performance of the learned policy. The data will be collected and stored after each round of interaction between the power system simulator with the objective of further training. The DRL agent will then learn and update the parameters of the neural networks based on the observation data. Another round of learning begins until reaching the predefined number of steps, and then another episode is initiated. The agent learns from this repetitive process and keeps updating the parameters of the critic and actor neural networks by maximizing the accumulated reward that was designed to adjust the policy of the action generation until the maximum limit on the episodes is reached.

For each training time step interval in the platform, the dynamic simulation will run for one time step to update the system states, and there will be one round of interaction between the power system simulator (PSS/E) and DRL agent (Python), during which data exchange happens.

This training platform is based on power system dynamic simulation (both power flow data and dynamic data are required) and is used for the emulation of real-time power system operation environment. Dynamic characteristics of systems can be observed by continued interaction and data exchange during detailed time-domain simulations. Different power system control problems can be addressed by applying and testing various state-of-the-art DRL algorithms based on this platform across a range of power grid simulations varying in scale.

V. SIMULATIONS AND RESULTS

The IEEE 9-bus system [32] and the 2000-bus Texas synthetic grid systems [33], [34], [35] are used as the test systems, based on which time-domain simulations are conducted and interfaced with the DDPG controller. All the case studies, including training and testing, were performed in the simulation environment based on the platform described in Section IV.

A. SIMULATION PARAMETERS

The training hyperparameters parameters are crucial to the efficiency of the algorithm. With careful tuning by trial and error, the learning rates for both the actor and critic are set as 0.001 with a 0.9 discount rate. The batch size, which indicates the number of sampled training data utilized from the reply buffer in one iteration, is set as 128 in considering the number of states and actions space in this study. A value of 10,000 memory capacity is adopted to adapt for a 128-batch size. Exploration noise is set as 1.6 to introduce explorations that can enrich the data set.

Both actor and critic neural networks have two hidden layers, which are connected with activation functions. The actor neural networks adopt Relu and Tanh activation functions, and critic networks adopt Relu as the activation function. Each layer includes 32 units to store and update the data.

The training step interval is set as 1 second, which means the power grid environment will exchange information with the DDPG agent, send current states, and get action commands every 1 second. The maximum step number, which

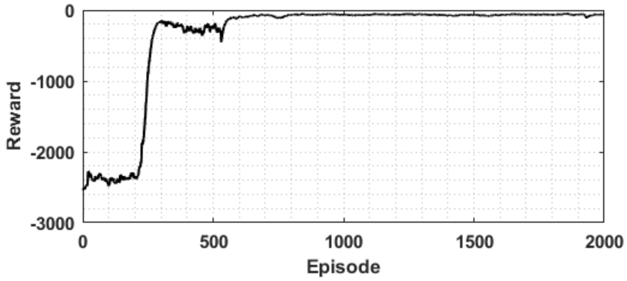

FIGURE 3. **Case 1 of IEEE 9-bus system: Average reward.**

indicates the maximum iteration count of each episode during training, is set to 50. The dynamic simulation will first run for 5 seconds to provide the initial states to start the training in each episode. Then, the disturbance is added at 5s. Therefore, each round of dynamic simulation will run for 55 seconds in total in every episode.

B. IEEE 9-BUS SYSTEM

The IEEE 9-bus system includes three generators and nine buses. Generator 1, a hydraulic unit with the salient-pole generator model GENSAL, is connected to slack bus 1. Generators 2 and 3 are steam turbines with the round-rotor synchronous generator model GENROU, they are controlled by the DDPG agent to participate in voltage control. All three generators are equipped with an IEEE type 1 excitation system model (IEEET1, the detailed model can be found in [36] of type DC1A excitation system model) and an IEEE standard governor model (IEESGO). The maximum action output is set as 1.3. The system loads include an active power component of constant current load and a reactive power component of constant impedance load. The reactive power load is randomly perturbed as the disturbance, which results in around 3% - 5% voltage fluctuations. The desired voltage normal range is conservatively considered as 0.98-1.02pu in this study, so the voltage reference in (10) and (13) is set as 1.00 pu to guide the DRL agent to control the voltage within the set range. Three cases with different training reward functions are discussed and analyzed below.

1) CONSIDERING VOLTAGE MAGNITUDE DEVIATION

The agent is trained with the reward function of (9) that considers bus voltage magnitude deviation. Figure 3 shows the moving average reward finally reaches a satisfactory level after 2000 episodes of training. The DDPG agent is applied to the system after being well-trained for testing by adding load disturbance at 5s to induce voltage changes. The test results, depicting the response to a 90 MVar reactive power load increase, are illustrated in Figure 4 and Figure 5.

Under generator control with constant exciter parameters, the system bus voltage magnitudes are significantly impacted and keep decreasing after the disturbance, which puts the system at high risk of losing stability. With the DDPG agent participating in the voltage control, bus voltages can be regulated to normal levels. The change of the excitation system voltage reference value of the two controlled generators can be seen in Figure 5. Generator 3 provides full voltage support after detecting the disturbance, and generator 2 is responsible for the voltage regulation in real time according to the system operating. The two generators cooperate under the control of the DDPG agent to help the system restore voltage.

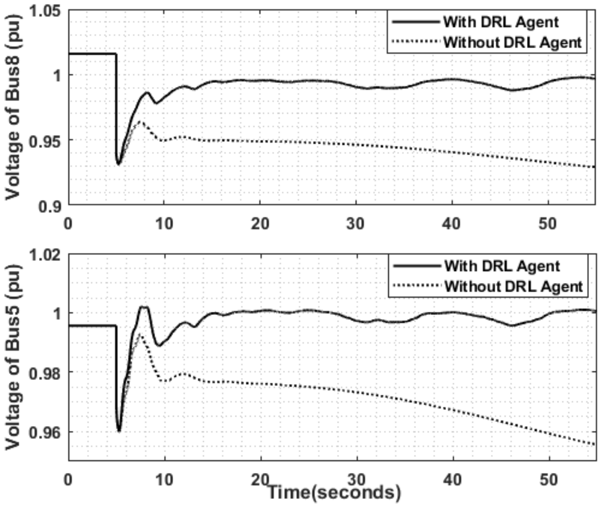

FIGURE 4. **Case 1 of IEEE 9-bus system: Voltage of bus 8 and bus 5 with and without DRL agent.**

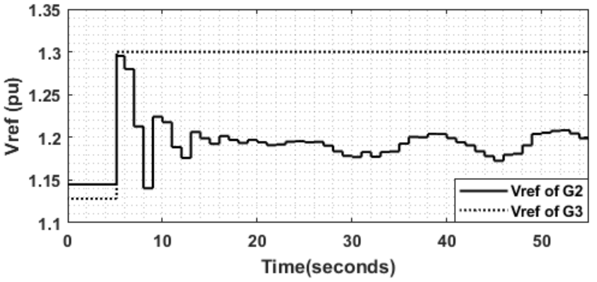

FIGURE 5. **Case 1 of IEEE 9-bus system: Generator voltage reference commands from DRL agent.**

2) CONSIDER VOLTAGE DEVIATION AND HISTORICAL VOLTAGE DATA

To further analyze the impact of historical data on agent control performance, we trained the DDPG agent with the reward function (12) that considers historical voltage data and bus voltage magnitude. c_t in (12) is set as 5, meaning the past 5 seconds of data are considered. After 2000 episodes of training, the moving average reward shown in Figure 6 reached and maintained a high level. After the training converges, the DDPG controller is implemented in the dynamic simulation of the system. This test simulation involves introducing the same 90 MVar reactive load change, enabling a comparison with case 1. The results of Figure 7 show that the DDPG agent's control policy considering historical voltage data can provide support to the system, helping it recover

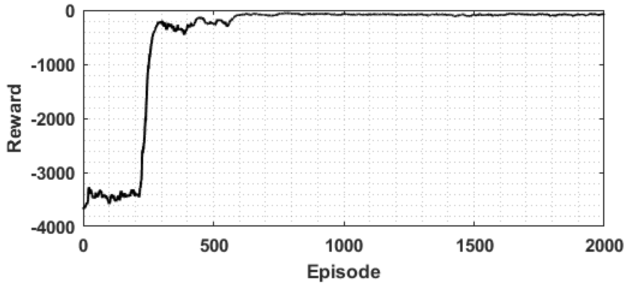

FIGURE 6. Case 2 of IEEE 9-bus system: Average reward.

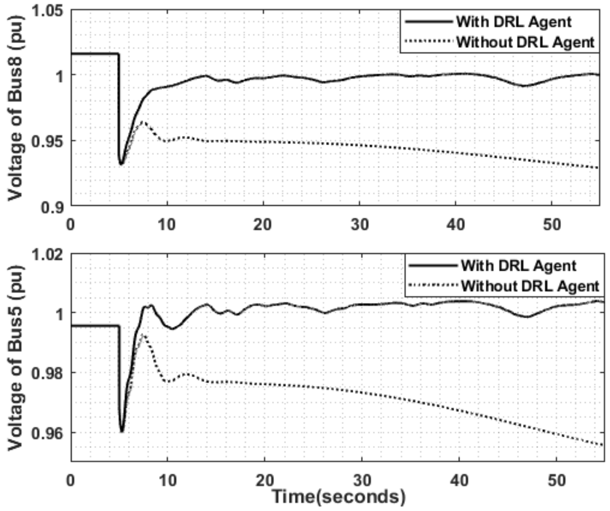

FIGURE 7. Case 2 of IEEE 9-bus system: Voltage of bus 8 and bus 5 with and without DRL agent.

to a normal voltage level. It's worth noting that in case 2, the bus voltage recovered faster with fewer oscillations, which demonstrated better dynamic performance compared to case 1. This provides evidence that historical data can provide valuable information to the DDPG agent, improving its policy accuracy in managing voltage oscillations and fluctuations during system operation.

3) CONSIDER VOLTAGE DEVIATION, HISTORICAL VOLTAGE DATA AND RATE OF CHANGE OF VOLTAGE

To explore the impact of the rate of voltage change data on the DDPG agent's performance, further training using the reward function in (14) is conducted. This function considers the rate of voltage change in historical data, which is calculated using the previous and present voltage values. For the preceding 5 seconds of historical data, there are four rates of voltage change data for each controlled bus. Following the completion of training, which is shown in Figure 8, the DDPG controller is tested with a reactive power load increase of 90 MVar as well. As shown in Figure 9, the results demonstrate that the agent can effectively support the system voltage recovery to the desired range in a more stable manner.

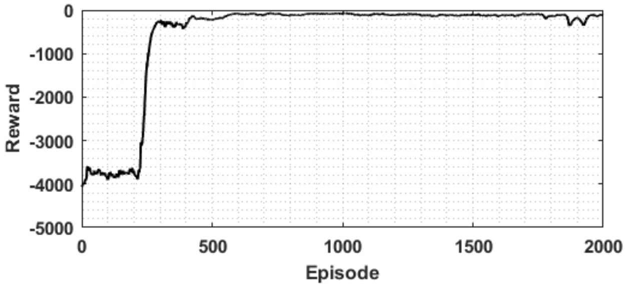

FIGURE 8. Case 3 of IEEE 9-bus system: Average reward.

In the analysis of the DDPG agent's dynamic control performance, different types of information in the reward functions are analyzed in Case 1 through Case 3. Figure 10 shows a comparison of the voltage control performance when the DDPG agent is tested with the same disturbance. The solid curve of Case 3, which considers both historical voltage data and voltage rate of changes, exhibits the smoothest voltage curve with the least fluctuation under the control of the DDPG controller. Additionally, Case 3 is capable of regulating and recovering the voltage faster due to the controller's ability to more accurately predict voltage changes based on dynamic features learned during training. The agent provided with extra information on the rate of voltage change can generate more effective actions to not only control the voltage level but also achieve better dynamic control performance.

4) WITH TIME-VARYING LOAD CHANGES

For a more comprehensive assessment of control performance under time-varying load fluctuations, varying load changes are introduced subsequent to the initial disturbance during the dynamic simulation. The reward function of (14) is used in this scenario, which incorporates the information on voltage deviation, the historical voltage data, and the voltage rate of changes. Specifically, after the initial 50MVar load change, an additional 30MVar load change at 20s and a 20MVar load change at 35s are introduced. The test results for bus voltage are depicted in Figure 11. Compared to bus voltage without the DRL controller, the voltage levels show significant improvement and ultimately return to approximately 1 pu. The smooth and satisfactory voltage recovery process underscores the effectiveness of the proposed control method in successfully addressing time-varying disturbances and effectively managing cascading failures.

C. TEXAS 2000-BUS SYNTHETIC TEST SYSTEM

To evaluate the effectiveness of the proposed DRL-based dynamic voltage control method on a more realistic system, simulations are conducted on the Texas 2000-bus synthetic power system, which is a large-scale representation of an actual power grid. This serves as a crucial step to test the proposed control method and the training platform.

In Figure 12, the structure of the Texas 2000-bus synthetic power system is depicted, where disturbances are introduced

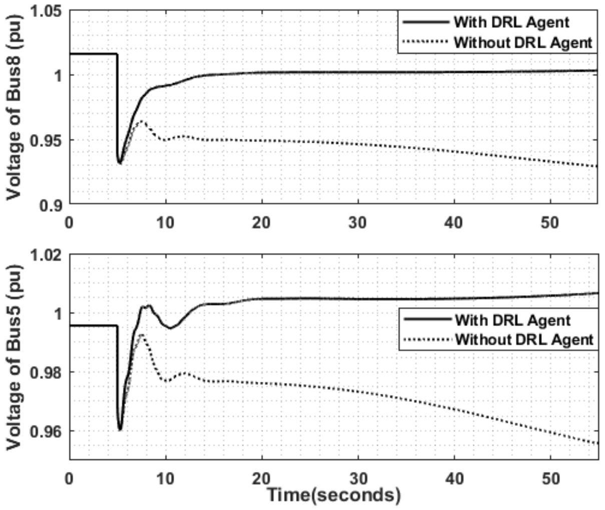

FIGURE 9. Case 3 of IEEE 9-bus system: Voltage of bus 8 and bus 5 with and without DRL agent.

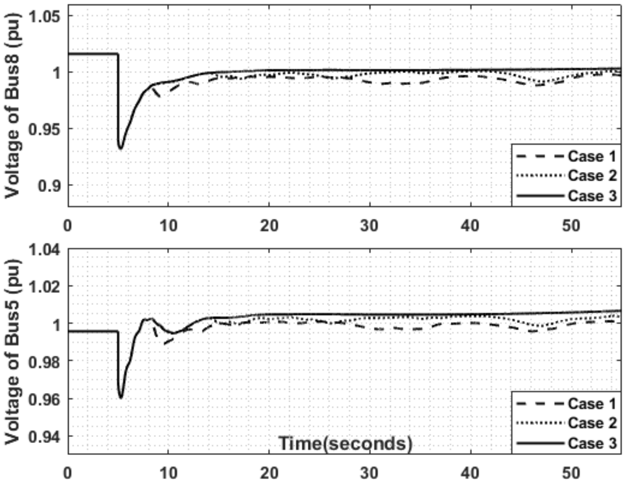

FIGURE 10. Case 1 to Case 3 comparison of IEEE 9-bus system: Voltage of bus 8 and bus 5.

in the heavily loaded Houston area (highlighted in red) to simulate scenarios with voltage issues. Among the generators, 7098 and 7099 are well-suited as controlled generators due to their large capacity and ample reactive power capability. As generator 7098 is connected to the swing bus of the system, generator 7099 is selected as the controlled generator, along with generator 7310, which is located at a short electrical distance from the Houston area, these two generators are chosen as the controlled generators for this case study. Generator 7099 and 7310 are both represented with the GENROU generator model. Generator 7099 employs an IEEET1 exciter model and IEEE type 1 speed-governing model (IEEEG1). While generator 7310 utilizes the ESST4B exciter model(the detailed model can be found in [36] of type ST4B excitation system model) and a general turbine-governor model(GGOV1). The system includes the same load model as the 9-bus system. To train and evaluate the controller's response to voltage changes, system disturbances are induced by altering the reactive power loads.

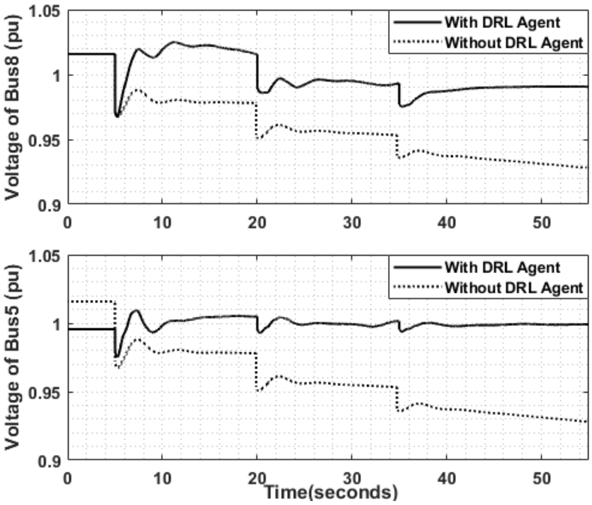

FIGURE 11. Case 4 of IEEE 9-bus system: Voltage of bus 8 and bus 5 with and without DRL agent.

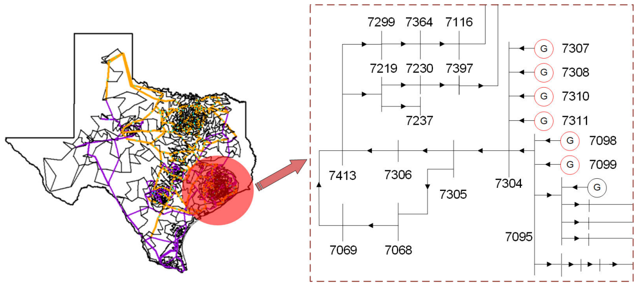

FIGURE 12. Diagram of disturbance area of 2000-bus system.

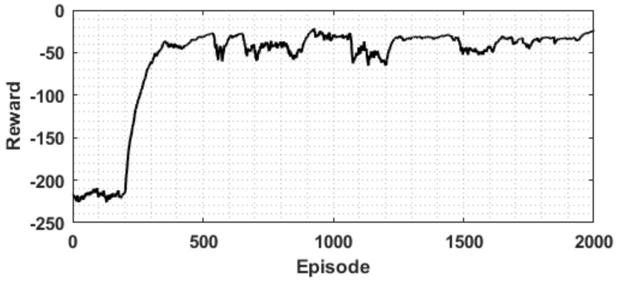

FIGURE 13. Case 1 of 2000-bus system: Average reward.

1) CONSIDERING VOLTAGE MAGNITUDE DEVIATION AND REGULATION COST

The simulation begins with the base case that utilizes equation (9) as the reward function, which considers voltage magnitude and regulation cost. After training the DDPG agent, as shown in Figure 13, where the reward reaches a high level, the agent is tested and the results are shown in Figure 14 to Figure 15. The results indicate that the agent can improve the

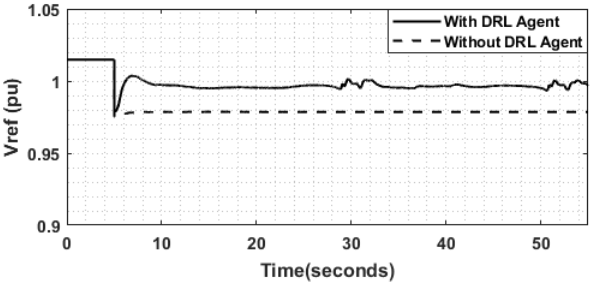

FIGURE 14. Case 1 of 2000-bus system: Voltage of bus 7068 with and without DRL agent.

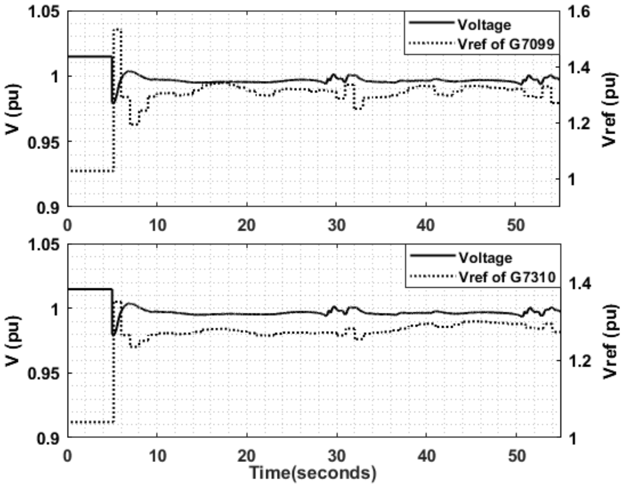

FIGURE 15. Case 1 of 2000-bus system: Generator voltage reference commands from DRL agent.

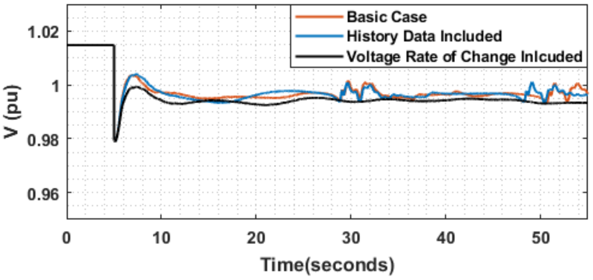

FIGURE 16. Comparison of 2000-bus system test results.

voltage to a satisfactory level compared to the conventional control mode. When a 230 MVar reactive load increases at 5s, the agent can detect the voltage change and generate commands to improve the generators' output immediately. The voltage of bus 7068 is shown in Figure 14 and Figure 15 as the voltage level representative for analysis. The voltage is restored to a normal level in about 2 seconds after the disturbance, and the generators can continuously regulate the excitation systems to achieve real-time voltage control in the recovery process. The two generators can respond quickly to voltage fluctuations under the control of the DDPG agent, which performs well in both situations of quick voltage control during sudden disturbances and minor voltage regulation in the process of system recovery.

2) CONSIDER VOLTAGE DEVIATION, REGULATION COST, HISTORICAL VOLTAGE DATA AND RATE OF CHANGE OF VOLTAGE

Various reward functions are employed for the DDPG controller in the Texas 2000-bus system, including considering historical voltage deviation and adding voltage rate of change in addition to the base case. The simulation results with the same load disturbance as case 1 are presented in Figure 16. The addition of voltage rate of change in the reward function leads to voltage recovery with a smoother curve, compared to the basic case and the case that includes historical voltage deviation. These two cases exhibit minor voltage oscillations and deviations, which do not exhibit satisfactory dynamic performance, though the voltage level has recovered to the normal level. The reward function which includes the voltage rate of change can guide the agent to achieve a maximum reward value and mitigate the oscillations, which improves the system's dynamic performance during control.

VI. CONCLUSION AND FUTURE WORK

This paper proposes a DRL-based data-driven excitation control scheme to realize real-time voltage regulations. The voltage control problem is formulated as a Markov Decision Process considers historical voltage data and the voltage rate of change information besides the voltage deviation and regulation cost, which leads to better dynamic performance during voltage recovery after disturbances. The development of a dynamic simulation training and test platform provides a reliable environment for the training and testing of different scales of systems regarding various control problems based on DRL algorithms. The results show that the proposed DRL-based dynamic voltage control method outperforms conventional voltage control methods in terms of faster and more accurate voltage control without relying on complex system models. The method demonstrates promising dynamic performance and can be readily generalized to large-scale power systems, which has the potential to be applied in practical power systems for real-time voltage control.

DDPG exhibits strengths in managing continuous action spaces, promoting sample efficiency, and demonstrating effectiveness in solving problems that require deep neural networks. Nevertheless, it does exhibit sensitivity to hyperparameter configurations and relies on a straightforward noise injection policy for exploration. It is worth considering more advanced algorithms that offer stable performance and can effectively address the uncertainties inherent in power systems.

REFERENCES

[1] M. Glavic, "(Deep) reinforcement learning for electric power system control and related problems: A short review and perspectives," *Annu. Rev. Control*, vol. 48, pp. 22–35, Jan. 2019.

[2] H. Lomei, D. Sutanto, K. M. Muttaqi, and A. Alfi, "An optimal robust excitation controller design considering the uncertainties in the exciter parameters," *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4171–4179, Nov. 2017.

[3] T. F. Orchi, T. K. Roy, M. A. Mahmud, and A. M. T. Oo, "Feedback linearizing model predictive excitation controller design for multimachine power systems," *IEEE Access*, vol. 6, pp. 2310–2319, 2018.

[4] C. Zhu, R. Zhou, and Y. Wang, "A new decentralized nonlinear voltage controller for multimachine power systems," *IEEE Trans. Power Syst.*, vol. 13, no. 1, pp. 211–216, Feb. 1998.

[5] Y. Guo, D. J. Hill, and Y. Wang, "Global transient stability and voltage regulation for power systems," *IEEE Trans. Power Syst.*, vol. 16, no. 4, pp. 678–688, Nov. 2001.

[6] H. Liu, Z. Hu, and Y. Song, "Lyapunov-based decentralized excitation control for global asymptotic stability and voltage regulation of multi-machine power systems," *IEEE Trans. Power Syst.*, vol. 27, no. 4, pp. 2262–2270, Nov. 2012.

[7] H. Liu, J. Su, J. Qi, N. Wang, and C. Li, "Decentralized voltage and power control of multi-machine power systems with global asymptotic stability," *IEEE Access*, vol. 7, pp. 14273–14282, 2019.

[8] Z. Zhang and M. Wu, "Predicting real-time locational marginal prices: A GAN-based approach," *IEEE Trans. Power Syst.*, vol. 37, no. 2, pp. 1286–1296, Mar. 2022.

[9] 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.

[10] Z. Zhang and M. Wu, "Real-time locational marginal price forecasting using generative adversarial network," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids*, Nov. 2020, pp. 1–6.

[11] 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.

[12] X. S. Zhang, T. Yu, Z. N. Pan, B. Yang, and T. Bao, "Lifelong learning for complementary generation control of interconnected power grids with high-penetration renewables and EVs," *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 4097–4110, Jul. 2018.

[13] Z. Zhang and M. Wu, "Locational marginal price forecasting using convolutional long-short term memory-based generative adversarial network," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2021, pp. 1–5.

[14] 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.

[15] 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.

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

[17] 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.

[18] S. Zarrabian, R. Belkacemi, and A. A. Babalola, "Reinforcement learning approach for congestion management and cascading failure prevention with experimental application," *Electric Power Syst. Res.*, vol. 141, pp. 179–190, Dec. 2016.

[19] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage regulation in distribution grids using deep reinforcement learning," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids (SmartGridComm)*, Beijing, China, 2019, pp. 1–6, doi: 10.1109/SmartGridComm.2019.8909764.

[20] M. R. Tousi, S. H. Hosseinian, A. H. Jadidinejad, and M. B. Menhaj, "Application of SARSA learning algorithm for reactive power control in power system," in *Proc. IEEE 2nd Int. Power Energy Conf.*, Dec. 2008, pp. 1198–1202.

[21] Y. Chen et al., "SAC-based voltage control in active distribution network with renewable energy resource," in *Proc. Panda Forum Power Energy (PandaFPE)*, Apr. 2023, pp. 2134–2138.

[22] Y. Wang et al., "Deep reinforcement learning based approach for active power security correction control of power system," in *Proc. 4th Asia Energy Electr. Eng. Symp. (AEEES)*, Mar. 2022, pp. 701–705.

[23] B. Zhang et al., "Real-time autonomous line flow control using proximal policy optimization," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Aug. 2020, pp. 1–5.

[24] Y. Wang et al., "Reinforcement learning based voltage control using multiple control devices," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2023, pp. 1–5.

[25] R. Yousefian, R. Bhattarai, and S. Kamalasadan, "Transient stability enhancement of power grid with integrated wide area control of wind farms and synchronous generators," *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4818–4831, Nov. 2017.

[26] R. Yousefian and S. Kamalasadan, "Energy function inspired value priority based global wide-area control of power grid," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 552–563, Mar. 2018.

[27] 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.

[28] W. Shao and Z. Xu, "Excitation system parameter setting for power system planning," in *Proc. IEEE Power Eng. Soc. Summer Meeting*, Oct. 2002, pp. 541–546.

[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[30] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[31] 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.*, vol. 36, no. 6, pp. 5525–5536, Nov. 2021.

[32] J. Conto. *IEEE9 Jconto*. Accessed: Jun. 20, 2023. [Online]. Available: https://drive.google.com/drive/folders/0B7uS9L2Woq_7fmd4YXVxMEZKT3dJV2FleGkzS2FzVmd1RHhBNVdUTGpvdldkMnl2bXRLM1k?resourcekey=0-nuCqXu2XJ0_fxBzwHcmCGg

[33] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3258–3265, Jul. 2017.

[34] A. B. Birchfield, K. M. Gegner, T. Xu, K. S. Shetye, and T. J. Overbye, "Statistical considerations in the creation of realistic synthetic power grids for geomagnetic disturbance studies," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1502–1510, Mar. 2017.

[35] K. M. Gegner, A. B. Birchfield, T. Xu, K. S. Shetye, and T. J. Overbye, "A methodology for the creation of geographically realistic synthetic power flow models," in *Proc. IEEE Power Energy Conf. (PECI)*, Feb. 2016, pp. 1–6.

[36] *IEEE Recommended Practice for Excitation System Models for Power System Stability Studies*, IEEE Standard 421.5-2016, 2016, pp. 1–207.