Virtual Synchronous Generator Control Using Twin Delayed Deep Deterministic Policy Gradient Method

Oroghene Oboreh-Snapps, *Graduate Student Member, IEEE*, Buxin She, *Graduate Student Member, IEEE*, Shah Fahad, Haotian Chen, *Graduate Student Member, IEEE*, Jonathan Kimball, *Senior Member, IEEE*, Fangxing Li, *Fellow, IEEE*, Hantao Cui, *Senior Member, IEEE*, and Rui Bo, *Senior Member, IEEE*

***Abstract*—This article presents a data-driven approach that adaptively tunes the parameters of a virtual synchronous generator to achieve optimal frequency response against disturbances. In the proposed approach, the control variables, namely, the virtual moment of inertia and damping factor, are transformed into actions of a reinforcement learning agent. Different from the state-of-the-art methods, the proposed study introduces the settling time parameter as one of the observations in addition to the frequency and rate of change of frequency (RoCoF). In the reward function, preset indices are considered to simultaneously ensure bounded frequency deviation, low RoCoF, fast response, and quick settling time. To maximize the reward, this study employs the Twin-Delayed Deep Deterministic Policy Gradient (TD3) algorithm. TD3 has an exceptional capacity for learning optimal policies and is free of overestimation bias, which may lead to suboptimal policies. Finally, numerical validation in MATLAB/Simulink and real-time simulation using RTDS confirm the superiority of the proposed method over other adaptive tuning methods.**

***Index Terms*—Deep reinforcement learning, frequency response, MATLAB/SIMULINK, microgrid, RTDS, virtual damping, virtual inertia, virtual synchronous generator.**

I. Introduction

RENEWABLE energy sources, such as wind and solar power, rely on power electronic interfaces to connect to the grid and are thus known as inverter-based resources (IBRs). Unlike conventional synchronous generators (SGs) that possess inherit inertia and damping characteristics, IBRs has no rotating masses to provide physical inertia. This poses significant challenges for grid operation, stability and security, particularly in low-inertia power networks that are dominated by IBRs [1], [2], [3], [4].

Manuscript received 11 December 2022; revised 23 April 2023 and 4 July 2023; accepted 4 August 2023. Date of publication 30 August 2023; date of current version 21 February 2024. This work was supported by the US DOD ESTCP Program under Grant EW20-5331 to complete this research work. Paper no. TEC-01286-2022. *(Corresponding author: Rui Bo.)*

Oroghene Oboreh-Snapps, Shah Fahad, Haotian Chen, Jonathan Kimball, and Rui Bo are with the Department of ECE, Missouri University of Science and Technology, Rolla, MO 65409 USA (e-mail: oogdq@mst.edu; shah.fahad@mst.edu; hc8xv@mst.edu; kimballjw@mst.edu; rbo@mst.edu).

Buxin She and Fangxing Li are with the Department of EECS, University of Tennessee, Knoxville, TN 37996 USA (e-mail: bshe@vols.utk.edu; fli6@utk.edu).

Hantao Cui is with the ECE, Oklahoma State University, Stillwater, OK 74075 USA (e-mail: h.cui@okstate.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TEC.2023.3309955.

Digital Object Identifier 10.1109/TEC.2023.3309955

To address this concern, the concept of virtual synchronous generator (VSG) is proposed for controlling grid-tied inverters. This form of control aims to mimic the inertia and damping characteristics of SGs [5], [6]. This control strategy offers numerous benefits to the power system, as it can work both as a grid-following and grid forming inverter when required [7]. The VSG control typically consists of two main loops: the active and reactive power loops. The emulation of inertia is carried out by utilizing the swing equation of a SG in the active power loop (APL) of the VSG. A well-designed APL parameter can guarantee suitable frequency response during disturbances such as load change or loss of generation, which ultimately would improve frequency stability in power systems [8]. An added advantage of the VSG when compared to the SG is its flexibility. In SGs, since the rotor mass is fixed, the available inertia is also a fixed characteristic. However, because a VSG is just a control algorithm, which is executed by a software, its parameters can be made adaptive to respond to disturbances as they occur in the system [9], [10]. While the APL loop parameters can provide improved frequency response (both frequency nadir and RoCoF), it is not beneficial to excessively increase its parameters as this could result in poor frequency response (prolonged settling time) [11]. Therefore, the study of an optimal-adaptive VSG becomes necessary to ensure high-quality power delivery, desirable frequency response and safe operation of the power system.

Improving the dynamic response of the VSG has been an active area of research in recent years. The current body of work can be categorized into two main groups; model-based and model-free methods. The model based methods depend on an accurate system model to develop the adaptive tuning law. The concept of Linear Quadratic Regulator (LQR) control was introduced in [11] to determine the optimal value of the virtual inertia constant for an individual VSG. Subsequently, this approach was extended to accommodate multiple VSGs in [12] Since the LQR controller falls in the category of optimal controllers, a trade-off was established between the microgrid frequency response and control cost. In [13], the voltage angle deviation stability of a microgrid consisting of multiple VSGs was analyzed. Afterwards, particle swarm optimization (PSO) was employed to tune the parameters of the VSG such that it guaranteed a smooth transition after a disturbance and maintained the voltage angle deviation within special limits. While these methods do improve the dynamic response of the

VSG, the adaptation mechanism is built on small signal analysis which involves utilizing a simplified and linearized mathematical system model. It is also important to highlight that the VSG performance is highly dependent on other system parameters such as the line parameters and grid strength [14]. This presents a more daunting task when considering the exact mathematical relationship that describes the interaction of VSG with the rest of the grid. Moreover, if such exact models can be developed, power-systems have different structures, which implies that the control strategy designed for a particular system structure may not work well within a structure with different system models (generators and loads) and parameters.

Due to the limitations of the model-based methods, model-free methods have become an attractive alternative approach. Model-free methods rely on data measurements hereby addressing most of the difficulties and flexibility surrounding model-based approaches [15]. For example in [16], a model free dynamic controller is designed for controlling the voltage of a DER in microgrids, while both [17] and [18] provides insights as to how model free methods can be adopted in stability analysis and dealing with parameter uncertainty respectively. Typically, model-free approaches can be further categorized into two sub-groups: rule-based, and reinforcement learning (RL).The rule-based methods predict the VSG APL parameters by using a set of predefined rules. For example, in [19], an adaptive-gain inertia control was used to improve frequency nadir while guaranteeing a stable power system operation. In order to establish good rules that will govern the adaptive response of the APL parameters and ultimately result in improved frequency response, the key parameters of a frequency response, frequency nadir/zenith and its RoCoF should be critically studied under different conditions [20]. In [21], an adaptive virtual control strategy based on bang-bang strategy is developed by evaluating both the change in frequency and its derivative in order to select suitable virtual inertia (J) and damping (D) parameters. Fuzzy logic controllers are also particularly useful in the domain of rule-based control. This class of controllers has three layers; Fuzzification, Inference and Defuzzification layer, to provide control parameters based on input data [22]. In [23], this class of controller was used for tuning the virtual inertia parameter in the VSG APL. This approach eliminates the need for an accurate mathematical model, as it adjusts the VSG APL parameters by relying on good expert knowledge rules and relevant system data measurements. However, the authors only considered the virtual inertia parameter of the VSG APL which mainly influences the RoCoF. In a recent study [24], fuzzy logic control was incorporated with the (VSG) to enhance damping during transient events by increasing the system's inertia. This was accomplished by introducing a correction term to the governor output power, effectively boosting the inertia of the system during transients. However, given that the APL of the VSG is designed to mimic the swing equation of a synchronous generator (SG), it would be more advantageous to investigate the impact of virtual inertia and damping on system performance. In [25], by using both frequency deviation and its derivative, the fuzzy logic controller modified the both virtual inertia and damping parameters which ultimately improves the frequency response of the microgrid. However, the absence of a secondary level controller in islanded mode encourages high selection of J/D parameters which slows down the VSG response.

The major drawback of fuzzy logic controllers and other rule-based techniques is their dependence on expert knowledge to establish the rule set. Thanks to the recent rapid improvements in artificial intelligence (AI), this dependency can be addressed by implementing a reinforcement learning (RL) agent. An RL agent interacts with the environment by taking an action based on received state information and obtaining a reward which indicates how good or bad the action taken was. The end goal is to learn the optimal policy that maximizes the expected cumulative reward [26], [27]. In [28], Q-learning was adopted to adjust the VSG controller parameters during a frequency event. However, Q-learning is a discrete state, discrete action algorithm which depends on a lookup table (Q-table) to store Q-value for each state-action pair. Therefore, as the state and action pair increases, Q-learning performance tends to degrade. A solution to this problem is to replace the Q-table with a neural network as discussed in [29]. Most recently, actor-critic methods have also been investigated and applied in VSG control. In [30], DDPG was applied to solve the optimal and adaptive problem of the VSG. In this work, the DDPG algorithm was tasked with finding parameters for the VSG that satisfied multiple performance indices. However, the VSG was assumed to be operating without a secondary level controller and the reward function was designed without considering the response speed for the VSG. These factors when considered could encourage the agent to select high virtual inertia and damping factor parameters as this strategy guarantees improved frequency response. In addition, no pictorial presentation of the actions are given in [30]. Also, in [31], DDPG has been adopted to tackle the optimal tuning of the VSG APL. However, the lack of an explicit reward function makes it hard to draw a fair comparison with the recent work. However, it is well-known that DDPG suffers from overestimation bias which could lead to poor control policy implementation. With the increased integration of DRL methods with DERs control, few literature have delved into the performing stability analysis for these methods [32], [33]. Since this work only focuses on the integration of DRL for controlling VSG inverters, details regarding DRL stability analysis are not presented in this article.

Based on the above discussion, this article presents an adaptive control design of a VSG for which no knowledge of the model is required. The Twin Delayed Deep Deterministic Policy Gradient (TD3) method is adopted to find the optimal policy. TD3, being an upgraded version of DDPG is immune to overestimation bias that results in suboptimal policies. In addition to frequency deviation and RoCoF, settling time is also included in state information and reward function design, which ensures the actions predicted by the agent are properly optimized. Then, an RL-VSG controller is configurated to improve the microgrid response. It interacts with the microgrid till finding the optimal policy based on a well-designed reward function. The major contributions of this work are summarized below:

- The optimal and adaptive VSG control problem is formulated as an RL problem hereby alleviating the need for complex mathematical models. This is achieved by

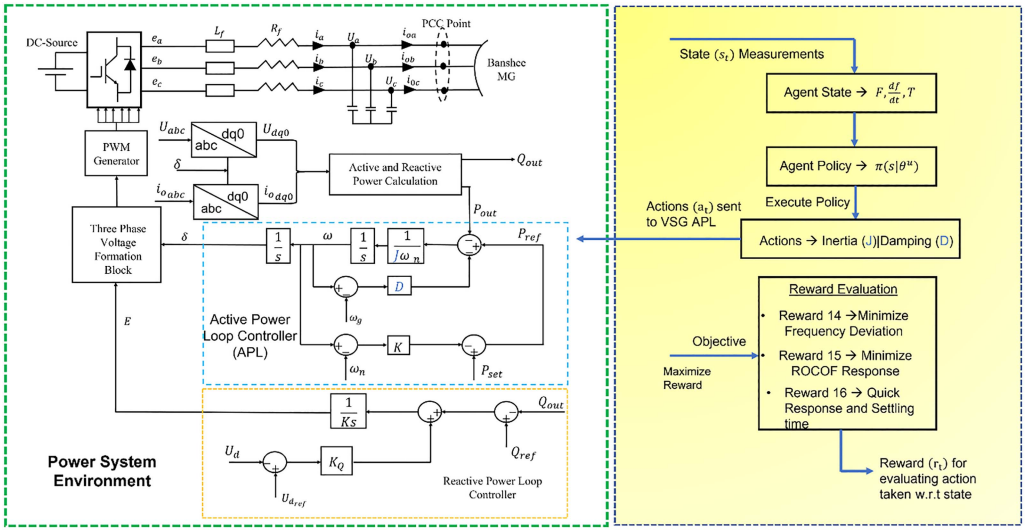


Fig. 1. VSG control with reinforcement learning framework.

adopting the state-of-the-art TD3 algorithm to obtain optimal control parameters.

- The unique inclusion of settling time in the reward function motivates the agent to further optimize its policy to ensure fast and constrained frequency response. In addition, a detailed sensitivity analysis of the reward function is presented to demonstrate the impact of intensifying the parameters of the reward function. This analysis aids in selecting the intensity of the parameters and further highlights the superiority of the proposed reward function.
- Furthermore, real time control and evaluation of the proposed TD3-VSG is performed using the Real-Time-Digital Simulator.

The rest of this article is organized as follows: Section II performs the model-based analysis, which specifically includes the modeling and small signal analysis of VSG. Section III formulates the control problem as an RL problem that aims at providing optimal actions while also satisfying different performance constraints. Section IV verifies the proposed method through numerical simulation in MATLAB/SIMULINK, while Section V presents the validation of the proposed controller using the real-time-digital-simulator (RTDS). Lastly, Section VI concludes the article with some recommendations for future work.

II. Model-Based Analysis for Virtual Synchronous Generator Control

A. Modeling of VSG

The prime objective of introducing VSG control in grid-tied inverter control is to emulate the inertia and damping properties of SGs. The schematic of a VSG-controlled inverter is shown in Fig. 1. As shown, the APL works based on the principle of the SG swing equation. Whereas, for controlling the flow of reactive power, a PI controller or droop control can be adopted. In this work, the primary focus involves improving the dynamic response of the APL. Hence, particular attention is given to two crucial parameters: virtual inertia and damping factor.

The swing equation [34] employed in the APL is expressed as:

$$P_{ref} - K_p(\omega - \omega_n) - D(\omega - \omega_g) - P_{out} = J\omega\dot{\omega} \quad (1)$$

From (1), P_{ref}, P_{out}, J, D, and K_p represents the active power reference, output active power, virtual inertia, virtual damping, and active power droop gain respectively. While ω, ω_n, and ω_g represent the speed of the virtual rotor, nominal angular speed, and the grid angular speed which is obtained through a PLL while the inverter is connected to the grid or the reference angular velocity while the inverter works in a standalone mode.

Based on power theory, P_{out} is given as;

$$P_{out} = \frac{3EU\sin\delta}{2X_{eq}} \quad (2)$$

With $X_{eq} = X_{line} + X_{filter}$ being the effective reactance. The inverter voltage magnitude E is the control output of the reactive power loop while the control output of the APL is the inverters' load angle δ which is given by:

$$\delta = \int (\omega - \omega_g)\, dt \quad (3)$$

In a traditional SG, the control of reactive power flow is realized by manipulating the field excitation. This concept is replicated in a VSG-controlled inverter by using a voltage droop control mechanism, as illustrated in Fig. 1. As the primary focus of this study is on the APL, we do not delve into an extensive discussion regarding the RPL. In a SG, the rotor size determines the available inertia. On the contrary the VSG is implemented as a software-based control algorithm which allows adaptive adjustment of the virtual inertia to minimize the impact of disturbances, i.e., ensuring improved frequency response. Before applying an adaptive control law using DRL, it is essential to understand the influence of VSG parameters on frequency

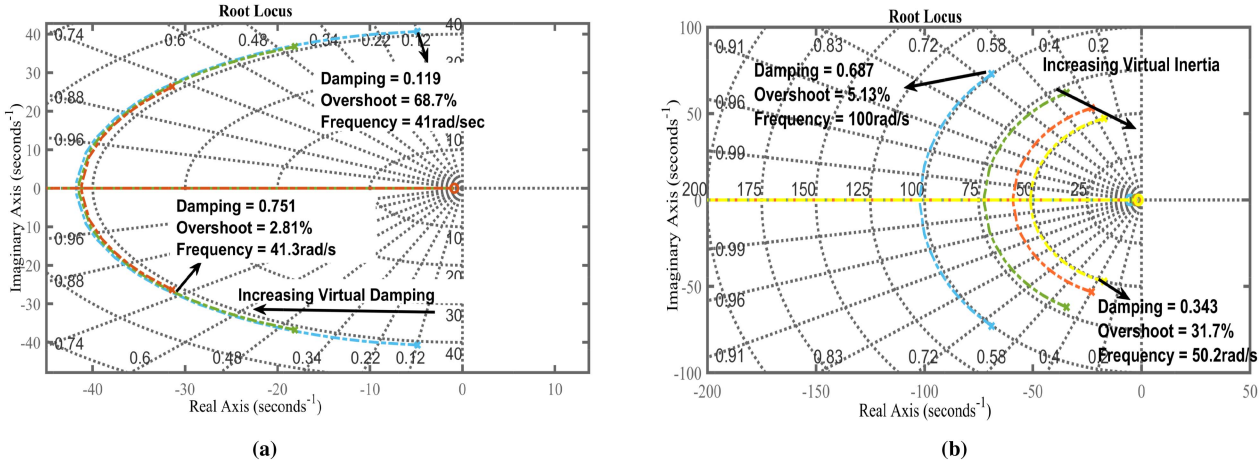


Fig. 2. Root-loci plots for (a) virtual damping (b) virtual inertia.

response. A small signal analysis is performed to examine this impact. Subsequently, the TD3 method is utilized to enhance the VSG controller's frequency response by dynamically adjusting the virtual inertia and damping parameters in the presence of disturbances.

B. Small Signal Analysis

Based on the approach given in [9], the VSG-APL transfer function can be derived by developing a small signal model. To achieve this, (1), (2), and (3), can be linearized as:

$$-K_p\Delta\omega - D(\Delta\omega - \Delta\omega_g) - \Delta P = J\omega s\Delta\omega \tag{4}$$

$$\Delta P = \frac{3EU cos\delta}{2X_{eq}}\Delta\delta + \frac{3U sin\delta}{2X_{eq}}\Delta E + \frac{3E sin\delta}{2X_{eq}}\Delta U \tag{5}$$

Generally, the inverter load angle δ is small. Consequently, $\sin\delta$ and $\cos\delta$ are approximately 0 and 1. The simplified form of (5) can be expressed as

$$\Delta P = \frac{3EU}{2X_{eq}}\Delta\delta \tag{6}$$

From (3) we get

$$\Delta\delta = \frac{\Delta\omega - \Delta\omega_g}{s} \tag{7}$$

It is obvious that, when a change in active power such as load disturbance occurs, there is a corresponding effect to the sensed frequency. Hence, the transfer function for the APL can be expressed as $G(s) = \frac{\Delta P}{\Delta\omega}$

From (4), it can be deduced that

$$\Delta\omega = \frac{D\Delta\omega_g - \Delta P}{J\omega s + K_p + D} \tag{8}$$

Also, ΔP can be expressed as,

$$\Delta P = \frac{3EU\Delta\omega - 3EU\Delta\omega_g}{2X_{eq}s} \tag{9}$$

Equation (9) can also be expressed as;

$$\Delta\omega = \frac{2X_{eq}\Delta P + 3EU\Delta\omega_g}{3EU} \tag{10}$$

Hence, solving for $G(s)$ we get:

$$G(s) = \frac{-3EU}{2X_{eq}} \frac{s + \frac{K_p}{J\omega}}{s^2 + \frac{Kp+D}{J\omega}s + \frac{3EU}{2X_{eq}J\omega}} \tag{11}$$

The transfer function derived in (11) can be used to plot the root-locus of the VSG APL model. Fig. 2 shows the root-locus plot for the APL parameters as they are varied.

Based on Fig. 2(a), the consequence of increased virtual damping increases the system damping ratio which correspondingly leads to reduced overshoot and damped oscillations which could force the system into an over-damped region thereby increasing the settling time. However, based on (1), high virtual damping directly impacts the frequency deviation and thus improves the nadir/zenith. In contrast, by increasing the virtual inertia as shown in Fig. 2(b), the system damping reduce which causes slow periodic oscillations into the system, increases overshoots, and the system takes a longer time period to settle at steady-state. However, from (1) the RoCoF is improved with increased virtual inertia.

As per the above analysis, the motivation of the proposed study is to design an adaptive VSG control that can change its parameters dynamically and optimally in such a manner that the frequency and RoCoF response are kept within specified limits while guaranteeing a quick VSG response. This objective can be actualized through data-driven RL methods, as specified in Section III.

III. Data-Driven Implementation for Virtual Synchronous Generator Control

This section designs TD3-based VSG with a detailed formulation of the reward function to guide the agents' learning.

A. Introduction to TD3 Algorithm

The TD3 algorithm is a model-free RL algorithm suitable for continuous control, as visualized in Fig. 3. It is revealed in [35] that the critic network of DDPG tends to overestimate the value function and leads to suboptimal policy and unstable training.

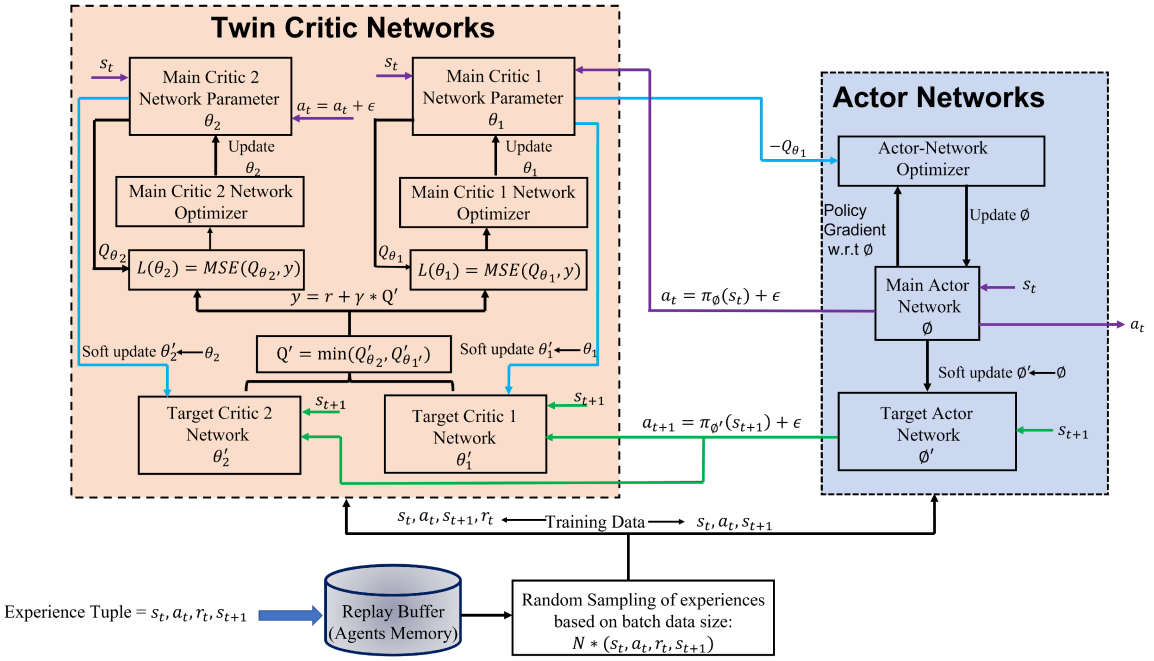


Fig. 3. Detailed TD3 agent structure.

Hence, TD3 addresses it by using delayed actor network updates, twin critic networks, and target policy smoothing regularization.

1) Actor and Critic Network: TD3 utilizes six neural networks shown in Fig. 3: 1 actor and 1 target actor network which are parameterized by ϕ and ϕ', respectively; 2 twin critic networks parameterized by θ_1 and θ_2; and 2 target twin critic networks parameterized by θ'_1 and θ'_2. At the start of training, the parameters of these networks are randomly initialized, alongside an empty finite buffer serving as a storage cache for the agent.

The actor-network works as the policy $\pi(s_t|a_t)$, which dictates how the actor-network should act (a_t) given a state(s_t). On the other hand, the goal of the twin critic network is to evaluate the action value function $Q_i(s_t, a_t|\theta_i)$ which is dependent on the action from the actor-network and the state information from the environment. Target networks, which are frozen copies of the actual networks, are popular tools used in DRL to achieve stability in training. Since DRL networks require multiple gradient updates to converge, target networks provide a stable objective during training which enables a greater coverage of the training data [35], [36].

2) Training Process: Algorithm 1 shows how the TD3 agent is trained. The agent takes a user-defined number of training steps T_s in each episode. It has no experience in how to act in the environment at the start of training. To encourage exploration, a decaying noise bounded within the maximum and minimum allowable action is added to the actions predicted by the actor network. The predicted actions a_t based on the current state s_t are applied to the environment and then the agent transitions to a new state s_{t+1}. The consequence of taking action a_t in state s_t is a reward r_t which is a measure of how good the action taken was. This sequence of events, denoted as s_t, a_t, r_t, s_{t+1}, creates a transition tuple saved in the buffer $\boldsymbol{B}$. Experiences stored in this buffer are randomly sampled and used to train the networks. Since buffer $\boldsymbol{B}$ is a finite cache, older experiences are removed to make space for newer experiences when it is filled. This facilitates the convergence because actions taken at the earlier stage of training have a high chance of being poor actions and could mislead the agents.

Algorithm 1: TD3 Alogrithm.

Initialize critic networks
$Q_{\theta 1}, Q_{\theta 2}$, and the actor network $\pi_\phi$with random parameters θ_1, θ_2, ϕ
Initialize target networks $\theta'_1 \leftarrow \theta_1, \theta'_2 \leftarrow \theta_2, \phi'_1 \leftarrow \phi_1$
Initialize replay buffer $\boldsymbol{B}$
for $t = 1$ to T_s **do**
  select action with noise
  $a \rightarrow clip(\pi_\phi(s) + \epsilon, a_{min}, a_{max})$, where $\epsilon \rightarrow$ ***Noise***
  Execute a in the environment
  Observe reward *r* and next state *s'*
  Store transition tuple *(s,a,r,s')* in $\boldsymbol{B}$
  Sample mini batch *N* transitions *(s,a,r,s')* from $\boldsymbol{B}$
  $\tilde{a} \leftarrow \pi'_\phi(s) + \epsilon, \epsilon clip\left((0, \sigma), c_{min}, c_{max}\right)$
  $y \leftarrow r + \gamma min_{i=1,2} Q_\theta{}'i(s', \tilde{a})$
  Update critics $\theta_i \leftarrow argmin_{\theta i} N^{-1} \sum (y - Q_{\theta i}(s, a))^2$
  if t mod d **then**
    Update ϕ by the deterministic policy gradient
    $\nabla_\phi J(\phi) = N^{-1} \sum \nabla_a Q_{\theta_1}(s, a)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)$
    Soft update for target networks:
    $\theta'_i \leftarrow \tau\theta_i + (1 - \tau)\theta'_i$
    $\phi'_i \leftarrow \tau\phi_i + (1 - \tau)\phi'_i$
  end if
end for

3) Upgrading From DDPG to TD3: Twin critic networks have been applied in 1) to prevent the overestimation of the Q-value. The following illustrates the other two upgrades of TD3 from DDPG.

i) *Target policy smoothening:* As discussed in [35], [37], there exists a tendency for deterministic policies to overfit the Q-value, which can cause increased variance in the target network. This implies that similar actions sometimes produce different value estimates which has a negative impact on the agent's learning. To curb this issue, the inclusion of a small amount of random noise in the target actor's actions improves the agent's exploration during training. This can enforce similar actions to have similar values and thus improves the learned policy of agents.
The target actor receives the new state s_{t+1} and estimates the target actions $\tilde{a}$. Both the target actions and the new state are then passed on to the twin target critic networks to compute their respective next action value Q'_{θ_i}. Next, to compute the target action value function y according to Algorithm 1, the minimum of the twin target critic network Q-function must be evaluated. By computing the minimum Q-value of the two target networks, the overestimation bias which causes the agent to learn sub-optimal action values can be avoided. The computation of the target-Q value y is dependent on the reward r, a discount factor γ, ranging from 0 to 1, and the minimum Q-value obtained from the target critic networks. The discount factor parameter enables the agent to balance a trade-off between immediate rewards (γ=0) and long-term rewards(γ=1). Next, the mean squared error between the target Q-value and each critic network Q value is computed independently to obtain each critic network loss value. To update the actor and critic networks, the gradients of the critic loss with respect to the weights of each critic network are computed. The gradients of the critic networks are then used to update the weights of each network using an optimizer such as Adam.

ii) *Delayed actor network updates:* The update of the actor network is delayed by the modulus of training step t and a hyperparameter called actor update frequency d. According to algorithm [1] J is the loss function of the actor network with respect to its network parameters ϕ. The gradient of this loss function is given by the inverse of the number of training batch samples N multiplied by the mean or the sum (Σ) of the gradient of the first critic network with respect to the state and action pair $\nabla_a Q_{\theta_1}(s,a)$ which is then multiplied with the gradient of the policy network $\nabla_\phi \pi_\phi(s)$. For simplicity, this means that to update the actor-network, the gradient of the Q value of the critic network with respect to the actor-network parameter ϕ is computed using the gradient of the value function of the first critic network. The actor loss is then computed based on the negative mean of the Q-values obtained from the previous steps. The gradients of the actor loss with respect to the network parameters are computed and used in updating the actor-network parameters.

Lastly, the target networks(critic and actor) are updated periodically by copying the parameters from the main networks via using the soft update rule with respect to a learning rate parameter τ as shown in Algorithm 1.

B. TD3-Based VSG Control

This subsection highlights the integration of the TD3 algorithm with the VSG controller. As discussed in Section III-A, the agent requires relevant state information to arrive at the desired control outputs (action). To this end, the following are defined explicitly.

- TD3 Agent and Network Structure: the APL of the VSG that mimics the swing equation is the controller of interest as it is responsible for frequency response improvement. Hence, the TD3-Agent is tasked with finding suitable control parameters that would satisfy certain performance indexes in the reward definition. Fig. 4 depicts the structure of the actor and twin critic networks. Their respective target networks are replicas of the main network structure.
- State Inputs: To achieve desirable performance, the RL agent has to interact with the power system environment by observing state input measurements s_t. Since this work primarily focuses on frequency response improvement, the corresponding state inputs to the agent are defined as;

$$s_t = \{F, df/dt, T\} \tag{12}$$

  where F is the measured frequency, df/dt is the RoCoF, T is the agent timesteps which provide the agent with relevant information about the frequency and RoCoF values at each time point. This is later used to access the frequency settling time.
- Action Outputs: The goal for the agent is to find a control policy that would maximize its cumulative reward. The control policy dictates what actions the agent should select actions based on state input. In this regard, to improve the frequency response, the agent is tasked with providing two action outputs, virtual inertia (J) and virtual damping (D) as shown below;

$$a_t = \{J_t, D_t\} \tag{13}$$

At each time instant t during the training, the TD3 agent interacts with the power system environment by receiving the state information, taking the corresponding action, and receiving a reward as shown in Fig. 5.

C. Reward Function Design

The goal of an RL agent is to maximize the expected discounted future rewards by optimizing its policy. In this work, the focus is primarily on the APL of the VSG. In [30] the exclusion of settling time could make the agent go for high J/D values hereby degrading the frequency response. However, in real-world application, it is desired that the agent made a trade-off between frequency and RoCoF improvements with respect to settling time. Hence, this article integrates the performance indices, i.e., frequency nadir, RoCoF, and settling time, into the reward function design.

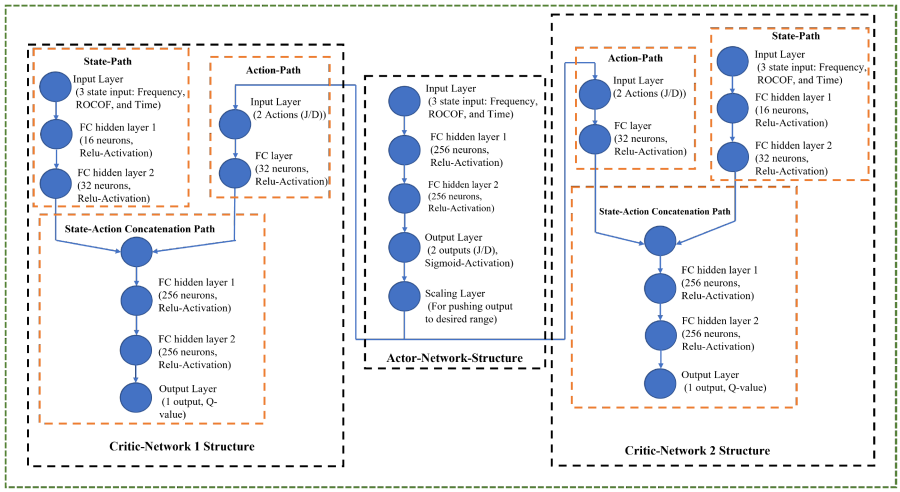


Fig. 4. TD3 actors-critic networks structure.

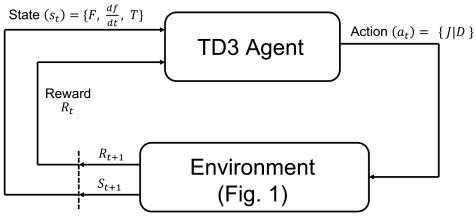


Fig. 5. TD3-agent and environment interaction.

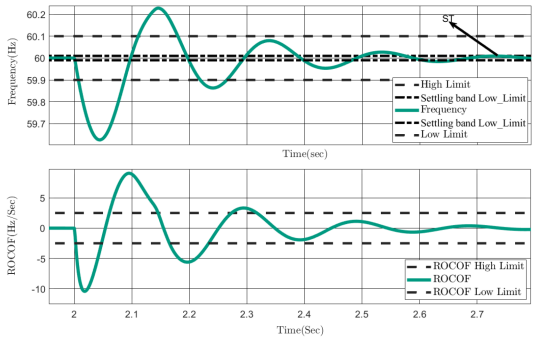


Fig. 6. Reward description.

1) Part I. Frequency Deviation: With regards to the frequency deviation, a limit is set to both the upward deviation (generation increase or load decrease) and downward deviation response (load increase or generation loss) of the frequency to guarantee that the agent learns to select optimal parameters that keep the frequency deviation above these critical points (see Fig. 6). Then, the following characteristic for the frequency deviation can be defined:

$$r(f_{div}) = -\alpha(F - F_{ref})^2 \tag{14}$$

where F, F_{ref} and f_{div} correspond to the measured frequency, reference frequency (60 Hz), and the deviation in frequency caused by the disturbance. The parameter α serves as a penalty factor such that if $F < F_{low-limit}$ or if $F > F_{high-limit}$ then α is a big penalty value, indicating to the agent that this constraint has been violated and receives a bad reward. Otherwise, if the frequency is within the tolerance limit, the agent is still motivated to further optimize its policy towards improving the frequency deviation since α becomes a small value in such a scenario.

2) Part II. RoCoF: The second performance index of interest is the RoCoF. The RoCoF of the system can be measured using a PMU or a filtered derivative. This index is important in microgrid operation because an aggressive RoCoF could trigger protection devices to false trip even under normal conditions. As a result, we design this index to be as follows;

$$r(df/dt) = -\beta(df/dt)^2 \tag{15}$$

where dfdt serves as the real time RoCoF measurement from the microgrid and β is the penalty function associated with this measurement. That is if $df/dt < df/dt_{low-limit}$ or if $df/dt > df/dt_{high-limit}$ then β is a large value, which would give the agent an unsatisfactory reward. However, just like in the frequency deviation component, even if there is no violation, it is still desired that the RoCoF is as minimal as possible during the disturbance, thus β is a small number outside the violation window.

3) Part III. Frequency Settling Time: While in [20] the DDPG algorithm has been considered for a similar problem, the exclusion of settling time from the reward function encourages the agent to select high values for the VSG-APL to improve the RoCoF and frequency response. However, as discussed in Section III, high APL parameters could harm the system stability and response time. To this end, the last component of the reward function, settling time, is designed to ensure the agent further optimizes its action selection to yield fast response time both in grid-connected and islanded mode. Hence, the characteristic of settling time is evaluated after the disturbance fades away and the frequency returns to the nominal value. To guide the agent, the settling time is defined as the time the frequency enters the settling band limit and stays in the band limit as shown in Fig. 6. The reward function for settling time is displayed below:

$$r(ST) = -\zeta(ST)^2 \tag{16}$$

where ST is the settling time and ζ is the penalty factor which is a fixed value in this case.

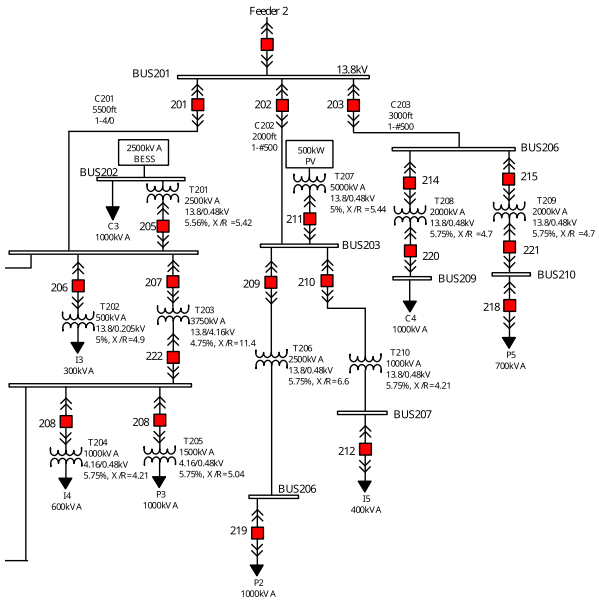


Fig. 7. Modified feeder 2 banshee microgrid.

4) Unified Reward Function: Putting all the reward components together, the final reward function is expressed as:

$$reward = r(f_{div}) + r(df/dt) + r(ST) \tag{17}$$

Based on (17), the agent aims at maximizing the discounted sum of future rewards by interacting with the environment and finding the optimal action(s). Fig. 6 shows a pictorial representation of the factors when designing the reward function. A disturbance in the form of active power reference change is introduced at 2 seconds. Then, a poorly designed VSG APL parameter would have a response curve that violates both the frequency limits and RoCoF limits. Translating this to the reward function would imply the agent obtains a bad reward because of large α, β parameters, and a huge penalty for settling time introduced by ζ.

It should be noted, the reward function described in (17) can give different results depending on the ascribed weights of penalty factors. For example, setting a higher priority to settling time as compared to the RoCoF and frequency nadir would imply that the agent would select actions that result in quick settling time which might impact the agent's ability to satisfy the frequency and RoCoF limits. In our work, since the frequency deviation and RoCoF are more important features of frequency events the penalty factor associated with them (α and β) are given more priority as compared to settling time penalty ζ.

IV. Case Studies

This section verifies the proposed TD3-VSG controller through numerical simulation.

A. Case Overview

A modified version of feeder 2 of the Banshee microgrid is adopted for training and testing. As shown in Fig. 7, it consists of two renewable energy sources (PV and BESS) located at BUS 203 and BUS 202, respectively. In islanded mode, the BESS is equipped with the proposed TD3-VSG controller which forms the microgrid voltage and frequency to guarantee stability when in islanded mode. In addition, a secondary controller is employed to guarantee frequency recovery.

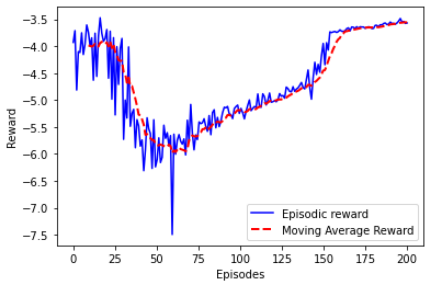
B. Training in Numerical Simulator

The actor and critic networks shown in Fig. 4 are designed in Python software and merged with the modified Banshee microgrid modeled in MATLAB/SIMULINK. The TD3 agent interacts with the environment by receiving observations of frequency, RoCoF, and time. The RoCoF is obtained by applying a filtered derivative to reduce the measurement noise. In addition, the state information is passed from MATLAB/SIMULINK to the TD3 agent in Python software, where the actor-network computes the actions at each time instant. The generated actions are then passed back and executed in the power system environment in MATLAB/SIMULINK. During the training, various disturbances such as active power reference change and load change are implemented. The agent is also restricted to making a maximum of 100 steps per episode. This is done to ensure the agent receives as much information regarding the state of the microgrid in order to achieve an optimal policy. Also, as shown in Fig. 8, the agent is trained for a maximum of 200 episodes which takes roughly 2 and a half hours when using ACER AV15-51 Laptop which has 16 GB RAM and a base clock of 2.92 GHz. This is considerably less when compared to the Lenovo IdeaPad 5 which has 8 GB RAM with a base clock of 2.40 GHz, which takes 5 hours for training. However, these are significantly less when compared to [30], which takes 31 h & 23mins. Although the available computational power, the number of training steps, and the complexity of the environment can affect the training time, the existing computational time is shorter enough to show the superior design over that in [30].

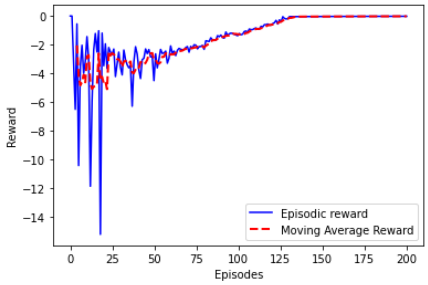
C. Comparison of Different Reward Functions

As discussed in Section III(b), the reward is designed to ensure the frequency and RoCoF response stays within a specific limit while ensuring quick settling time. The performance of all DRL methods is highly dependent on the reward function. Hence, it is important to have a reward function that accurately captures the desired behavior. This section presents the comparison results of different reward functions to show the impact of the punishment factors on the performance of the TD3-VSG controller. Three cases are selected and presented below.

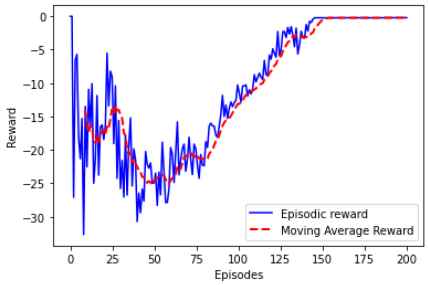
1) Case A1. Reward function with settling time as priority: The reward function of Case A1 is characterized by (17) and Fig. 8(a) shows the converged reward curve. Here, all components of the frequency response are captured in the reward, however, the settling time penalty (ζ) is weighted much more than the other penalty factors. This strategy implies that the agent is encouraged to select actions from the action space that guarantees a quick settling time which could cause violations in other components of the reward definition. In this case, the active power reference is changed from 1.5 MW to 0.5 MW at 1 s. The reward curve indicated by Fig. 8(a) shows that the agent



(a) Reward Curve for Case A1



(b) Reward Curve for Case A2



(c) Reward Curve for Case A3

Fig. 8. TD3 training curves with different reward functions.

learns a strategy that fits this scenario. However, as shown in Fig. 9(b) and (c), while this reward setting does have the best settling time response, its frequency nadir violates the threshold limit, and its RoCoF response is worse compared to the other cases. This is caused by the agent selecting significantly lesser values of virtual inertia and damping factor parameters when compared to the other case scenarios as illustrated in Fig. 9(d) and (e)

2) Case A2. Reward function with priority on frequency deviation and RoCoF component: The reward function for Case A2prioritizes frequency deviation and RoCoF more significantly as compared to settling time. Similarly, the VSG active power reference is reduced from 1.5 MW to 0.5 MW. The training curve for this case is shown in Fig. 8(b) which indicates the agent learns a policy that suits the modified reward definition.

As expected, the agent is motivated to select high values of virtual inertia (J) and virtual damping (D) as shown in Fig. 9(d) and (e) while neglecting the requirement for settling time. Such reward function implies that there is only one unique strategy (high control gains) that can guide the agent to achieve a maximum reward. This is not desirable as in the case of most control problems. A consequence of this sort of reward definition is a higher frequency nadir and improved RoCoF response, along with a prolonged settling time. Then, the trade-off between the three frequency indices is not considered.

3) Case A3. Reward function with balanced performance index: Case A3 balances all the frequency response requirements in the reward function. A desired frequency response should not violate either the frequency high or low limits and the RoCoF should not violate its limits either. To this end, the penalty factor for the frequency deviation is selected to have a higher weight as compared to the other elements. This would typically result in higher values of virtual damping and slightly higher values of virtual inertia. Since both virtual inertia and damping have correlated effects on the frequency nadir and RoCoF, a lower penalty factor can be defined for the RoCoF penalty. The settling time penalty has a slightly lower penalty to prevent the agent from selecting remarkably high values. In this way, the agent is motivated to improve its response time to achieve a better reward when the frequency deviation and RoCoF constraints are satisfied.

Similar to other cases, the active power reference is reduced from 1.5 MW to 0.5 MW at 1 s. The reward curve is shown in Fig. 8(c), from which it is evident that the agent learns an optimal policy over time. Also, as shown in Fig. 9(b) and (c), Case A3 enables the agent to find optimal parameters that ensure the frequency and RoCoF stay within their limits and recover to nominal value as quickly as possible. In Fig. 9(d) and (e), the agent takes full advantage of the action space by increasing these values at the start of the disturbance to arrest the frequency nadir and RoCoF, while also reducing its control gains to achieve quick settling time.

Based on the results in Fig. 8, it is clear that every DRL agent finds the optimal policy that maximizes the cumulative reward. In essence, this implies the importance of defining proper reward functions to capture the true control tasks. Fig. 9 shows the microgrid active power, frequency, and RoCoF response for all 3 cases discussed thus far. In this work, the objective is to keep frequency and RoCoF within special bounds while also ensuring a quick response. Hence, all three kinds of frequency indices should be properly scaled to achieve the desired goal, as validated in Case A3.

D. Intentional Islanding With a Secondary Controller

This subsection verifies the performance of the TD3-VSG controller under intentional islanding. According to [25], the modified feeder 2 of the Banshee microgrid consists of only RES. The BESS is designed to have more capacity than the PV, with which the voltage and frequency of the whole microgrid are maintained in islanded mode. Then, comparisons between a fixed VSG, a Fuzzy-VSG, DDPG-VSG based on the reward in [30], TD3-VSG are conducted to demonstrate the advantages of the proposed method

With regards to the fuzzy logic control, 5 membership functions are defined to classify the measured states and the corresponding fuzzy outputs. As for the DDPG-VSG, [30] aimed to maintain deviations in frequency within special limits, preserve well-damped oscillations, and obtain slow frequency drop in the transient process. The reward function only includes the

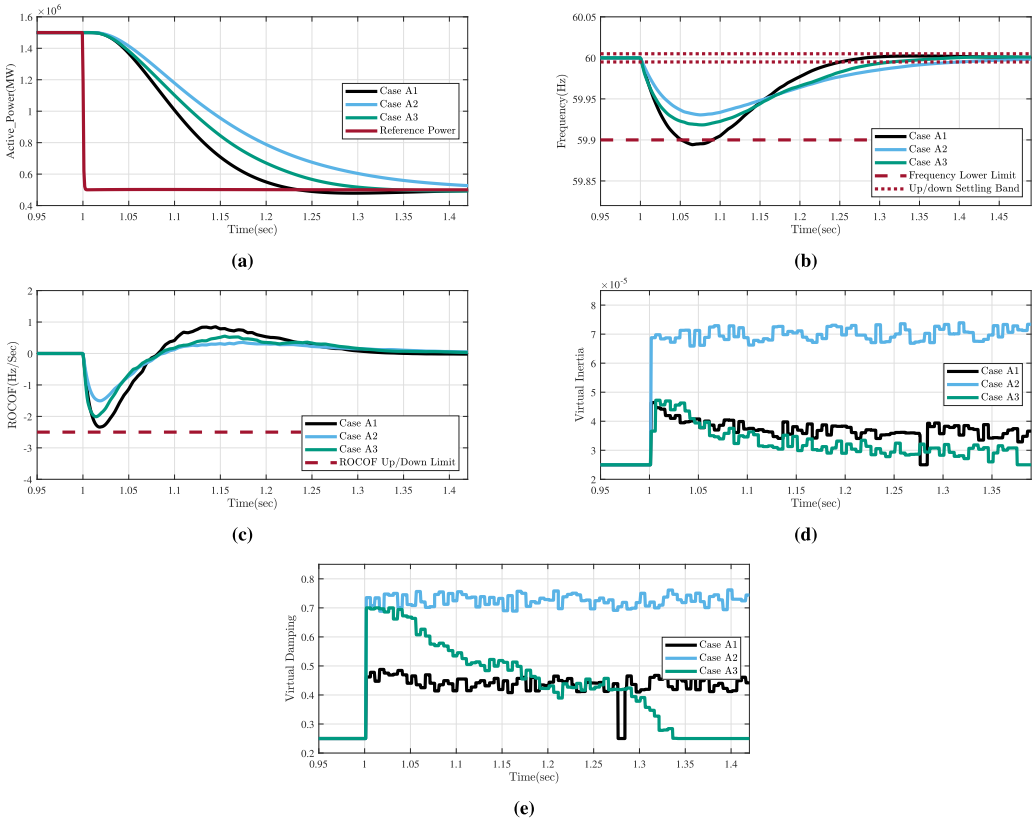


Fig. 9. System responses to active power reference change with different reward functions. (a) Active Power Response. (b) Frequency Response. (c) RoCoF Response. (d) Virtual Inertia. (e) Virtual Damping.

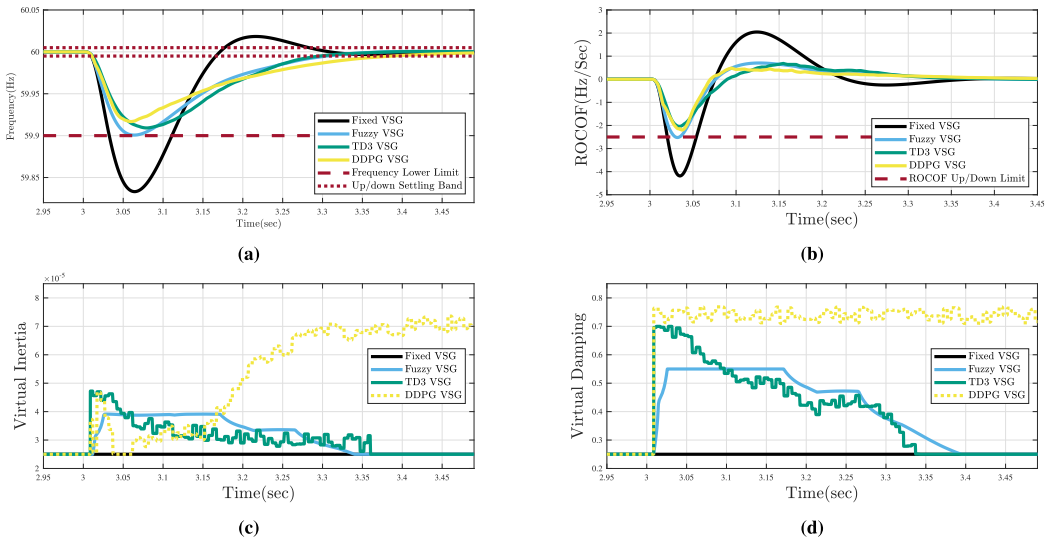


Fig. 10. Intentional islanding response (a) frequency response (b) rocof response (c) virtual inertia (d) virtual damping.

frequency deviation and RoCoF, which results in the high values of J and D at the same time.

At 3 s, the breaker connecting the feeder to the main grid is open and the modifed Banshee microgrid works in islanded mode with the BESS equipped with a VSG controller. For a fair comparison, all methods are initialized at the same minimum value in the action search space available to the DRL agent. As shown in Fig. 10(a) and (b), the fixed VSG method tends to have the worst response as its J and D values are not adaptive. To have a good response with such an approach, great effort must be made to arrive at a single fixed value that satisfies all performance indexes. On the other hand, since the Fuzzy-VSG, DDPG-VSG, and TD3-VSG can respond to sensed disturbance, improved responses can be observed with these methods. The

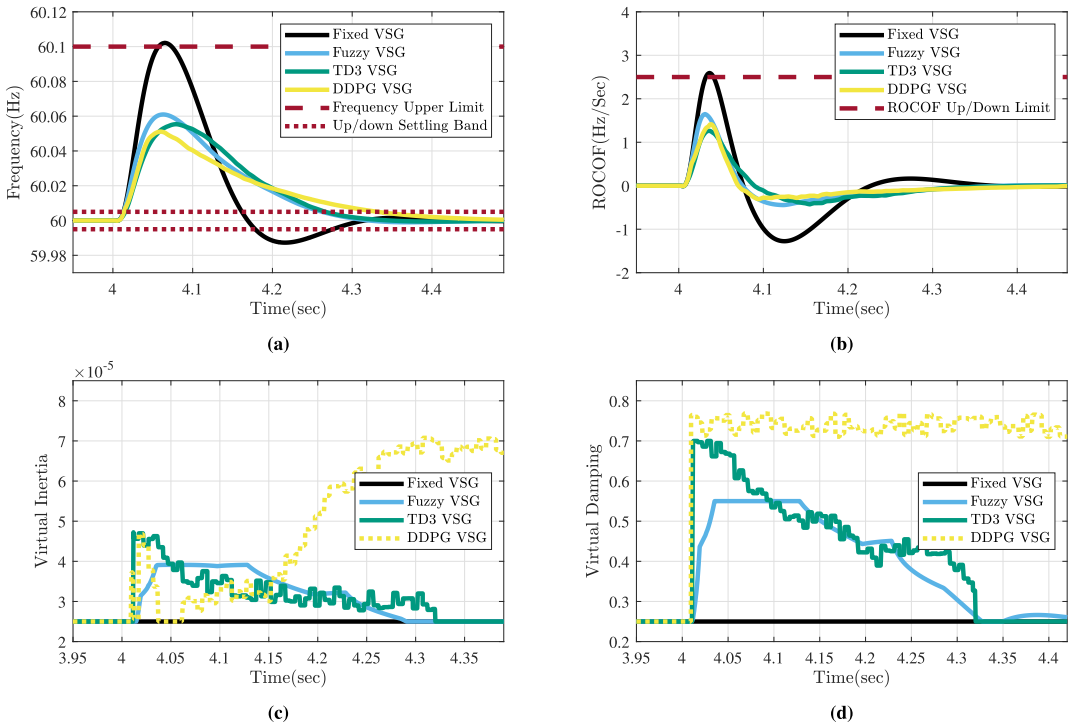


Fig. 11. Performance comparison of different VSG controllers under Load Change. (a) Frequency Response. (b) RoCoF Response. (c) Virtual Inertia. (d) Virtual Damping.

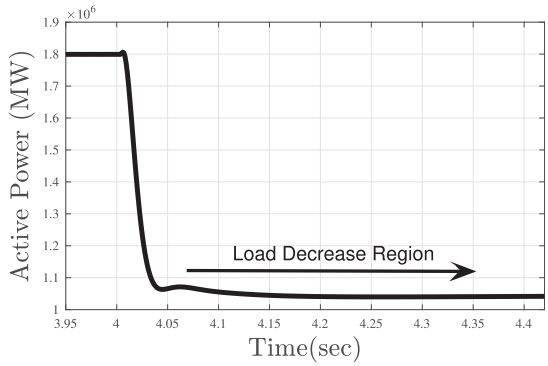


Fig. 12. Active power response to load change.

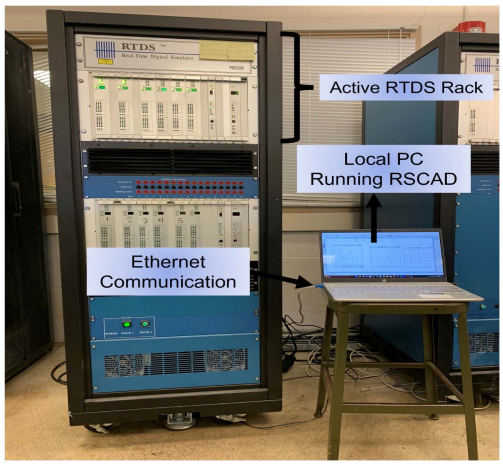


Fig. 13. RTDS/RSCAD setup.

control outputs of Fuzzy-VSG rely on expert knowledge in designing the rules. This heavy reliance on expert knowledge implies that the controller can only be as good as the defined rules and membership function. As shown in Fig. 10(c) and (d), the Fuzzy-VSG increases its control actions when the disturbance is sensed and slowly reduces it as the disturbance fades off. With regards to the DDPG-VSG in [30], the agents are motivated to increase the control parameters as high as possible and retain them there. This implies that the desired trade-off between improving frequency nadir and RoCoF is neglected and depending on the action space limits, the agent selecting actions at the maximum range would result in slow VSG response and prolonged settling time. But for TD3-VSG, it made the best trade-off between the three frequency indices.

E. Load Change in Islanded Mode

This subsection further compares the performance of different VSG controllers after load change in islanded mode. As shown in Fig. 12, the BESS supplies 1.8 MW of active power to meet the load demand before 4 s. Assume a load decrease happens at 4 s and causes a rise in the microgrid frequency.

Fig. 11(a) and (b) show the frequency and RoCoF response while Fig. 11(c) and (d) show the control actions. Due to the inability of the fixed-VSG to adaptively respond to the disturbance, both its frequency and RoCoF violate their respective limits. On the other hand, while the fuzzy VSG, DDPG-VSG, and TD3-VSG do constrain the frequency and RoCoF responses, their predicted actions are dependent on the fuzzy logic rule or the reward function. Both TD3- and DDPG-VSG have slightly better frequency zenith and RoCoF as compared to Fuzzy VSG.

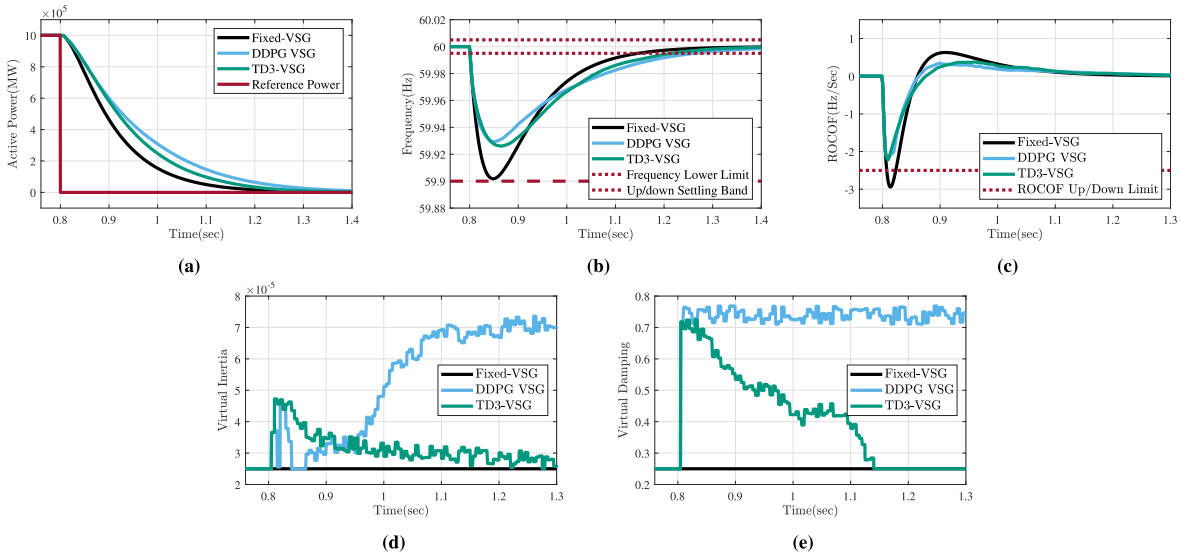


Fig. 14. RTDS response comparison with fixed VSG versus TD3-VSG for active power reference change (a) active power response (b) frequency response (c) RoCoF response (d) virtual inertia (e) virtual damping.

However, as stated previously, due to the characteristic of the reward function defined in [20], the DDPG-VSG eventually settles for higher control parameters resulting in a longer settling time as compared to TD3-VSG.

V. Validation Using Real-Time Digital Simulator

This section further validates the training results using a Real-Time-Digital-Simulator (RTDS) and its proprietary software RSCAD.

A. RTDS Configuration

The RTDS hardware is a powerful tool used for real-time simulations and analysis of power systems and microgrids. It is designed to accurately model and emulate the behavior of electrical grids in real-time. It can be utilized either in connection with a physical device or in a standalone mode. In this work, the standalone mode is utilized as only the controller performance is evaluated. Fig. 13 shows the setup between RTDS and RSCAD, where communication between the local PC and the RTDS rack is achieved through an Ethernet cable connection. The RTDS rack consists of 6 GPC cards, 1 GTWIF, and 1 GTNet card. The GPC cards handle the computation and execution of complex power system models and control algorithms. Each GPC card is equipped with multiple cores and provides significant processing power to handle the simulation workload. The main role of the GTWIF is to handle communication between the RTDS power system simulator and the host computer workstation. The GTNET card provides a real-time communication link to and from the simulator via Ethernet [38]. For testing the proposed control, the feeder 2 of the banshee microgrid is modeled in RSCAD and the trained TD3 agent is imported into the software for system validation.

TABLE I
Key System Parameters

PARAMETERS	VALUES	PARAMETERS	VALUES
BESS Inverter Rating	2500KVA	ζ	10
AC Bus Voltage (Bus 202)	480Volts	Actor-Network Structure	3-256-256-2
DC-Side Voltage	900Volts	Critic 1 Learning Rate	0.002
Fixed Virtual Inertia	0.00025pu	Critic 2 Learning Rate	0.002
Fixed Virtual Damping	0.25pu	Actor Learning Rate	0.001
Inverter Filter Resistance	1.9mΩ	Target Networks Learning Rate	0.005
Inverter Filter Inductance	0.05mH	Damping Action Search Space	[2.5 7.5] *10^-1
Microgrid Frequency	60Hz	Virtual Inertia Search Space	[2.5 7.5]*10^-5
α big	600	Frequency Minimum Limit	59.90Hz
α (small)	100	Frequency Maximum Limit	60.10Hz
β (small)	0.01	RoCoF Limit	2.5Hz/Sec
β (big)	2	Frequency BandLimit	0.005

B. Test Results

In Fig. 14, the active power reference is changed from 1 MW to 0 MW at 1 s leading to a frequency dip. While both the fixed VSG and TD3-VSG do not violate the frequency limit listed in Table I, the TD3-VSG still has a better frequency nadir than the fixed VSG and its RoCoF is above the specified limit. However, with the reward definition used in [30] for the DDPG-VSG case, a slightly improved frequency nadir is observed when compared with TD3-VSG. This is because the reward in [20] does not fully consider the trade-off between frequency and RoCoF versus settling time. Also, choosing high actions for J and D, and not reducing them after the disturbance fades off, could further elongate the system's settling time.

Next, in Islanded mode, the load is decreased from 2.3 MW to 1.48 MW at 1 s as indicated by Fig. 15. This disturbance results in reduced generation output by the VSG causing the frequency

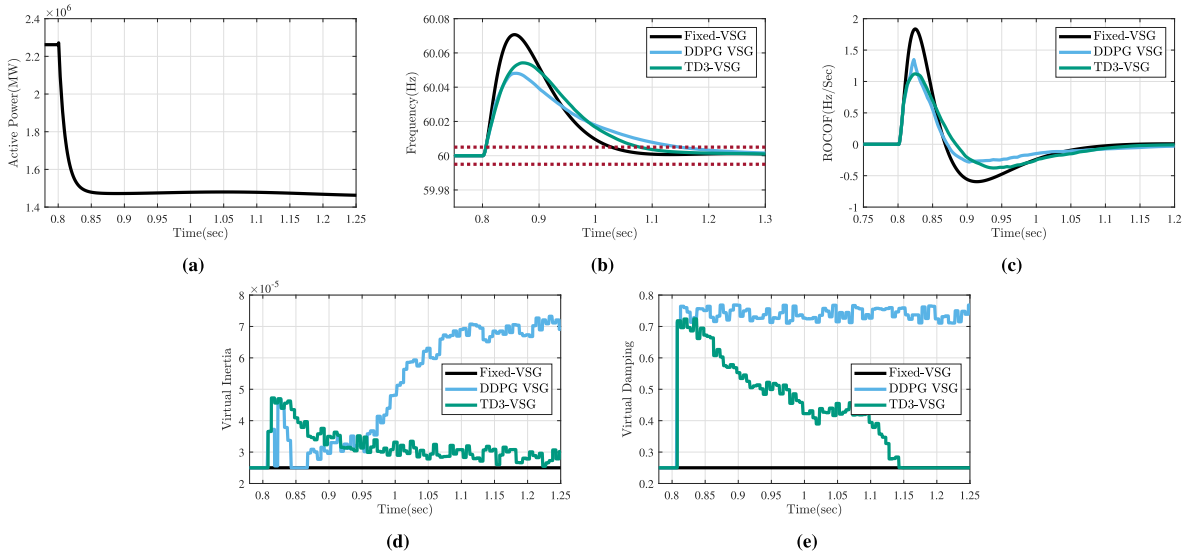


Fig. 15. RTDS response comparison with fixed VSG versus TD3-VSG for load change in islanded mode: (a) active power response (b) frequency response (c) RoCoF response (c) RoCoF response (d) virtual inertia (e) virtual damping.

to increase. Again, while both the Fixed and TD3-VSG do not violate the frequency limits imposed, the TD3-VSG still has an improved frequency zenith and its RoCoF does not violate the upper RoCoF limit described in Table I.

In summary, the controller real-time simulation experiments in RTDS validate the training results and the performance of the proposed TD3-VSG controller. It outperforms the Fuzzy-VSG and DDPG-VSG in [30].

VI. Conclusion

This article presents a TD3-based VSG controller for microgrids. To achieve the desired control performance for frequency response, multiple characteristic functions are investigated and used to design reward functions. The superiority of the proposed TD3-VSG controller is shown by comparing its performance with the proposed VSG controllers such as fuzzy-VSG and DDPG-VSG.

While fuzzy logic can provide a good frequency response by dynamically adjusting its actions, the actions depend heavily on a human expert in constructing the rules. On the other hand, DRL methods like DDPG and TD3 learn from interaction with an environment in a bid to maximize its reward. They require good reward design and proper penalty scaling to achieve desired responses. Via comparison, the reward proposed in this article shows the agent action satisfies the performance index without the need for excessively high control parameter selection as compared to that defined in [30]. In the future, this work would be extended to cover multiple VSGs and reactive power loop control improvement.

References

[1] Q. Liu, T. Caldognetto, and S. Buso, "Review and comparison of grid-tied inverter controllers in microgrids," *IEEE Trans. Power Electron.*, vol. 35, no. 7, pp. 7624–7639, Jul. 2020.

[2] W. U. Rehman et al., "The penetration of renewable and sustainable energy in Asia: A state-of-the-art review on net-metering," *IEEE Access*, vol. 8, pp. 170364–170388, 2020.

[3] B. She et al., "Decentralized and coordinated VF control for islanded microgrids considering DER inadequacy and demand control," *IEEE Trans. Energy Convers.*, vol. 38, no. 3, pp. 1868–1880, Sep. 2023.

[4] W. U. Rehman, A. Moeini, O. Oboreh-Snapps, R. Bo, and J. Kimball, "Deadband voltage control and power buffering for extreme fast charging station," in *Proc. IEEE Madrid PowerTech*, 2021, pp. 1–6.

[5] A. Vasilakis, I. Zafeiratou, D. T. Lagos, and N. D. Hatziargyriou, "The evolution of research in microgrids control," *IEEE Open Access J. Power Energy*, vol. 7, pp. 331–343, 2020.

[6] K. M. Cheema, "A comprehensive review of virtual synchronous generator," *Int. J. Elect. Power Energy Syst.*, vol. 120, 2020, Art. no 106006.

[7] H. Luan et al., "Multiple mode operation and control for VSG interfacing DC distribution system," in *Proc. IEEE 4th Workshop Electron. Grid*, 2019, pp. 1–6.

[8] H. Wu et al., "Small-signal modeling and parameters design for virtual synchronous generators," *IEEE Trans. Ind. Electron.*, vol. 63, no. 7, pp. 4292–4303, Jul. 2016.

[9] G. Jianyi and F. Youping, "VSG-based parameter adaptive control strategy," in *Proc. E3S Web Conf.*, 2021, Art. no. 02041.

[10] B. She, F. Li, H. Cui, J. Wang, Q. Zhang, and R. Bo, "Virtual inertia scheduling for power systems with high penetration of inverter-based resources," 2022, *arXiv:2209.06677*.

[11] U. Markovic, Z. Chu, P. Aristidou, and G. Hug, "Fast frequency control scheme through adaptive virtual inertia emulation," in *Proc. IEEE Innov. Smart Grid Technol.-Asia*, 2018, pp. 787–792.

[12] U. Markovic, Z. Chu, P. Aristidou, and G. Hug, "LQR-based adaptive virtual synchronous machine for power systems with high inverter penetration," *IEEE Trans. Sustain. Energy*, vol. 10, no. 3, pp. 1501–1512, Jul. 2019.

[13] J. Alipoor, Y. Miura, and T. Ise, "Stability assessment and optimization methods for microgrid with multiple VSG units," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1462–1471, Mar. 2018.

[14] Z. Song et al., "Small signal modeling and parameter design of virtual synchronous generator to weak grid," in *Proc. IEEE 13th Conf. Ind. Electron. Appl.*, 2018, pp. 2618–2624.

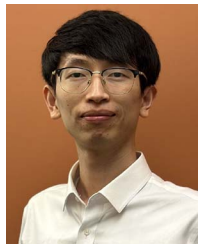
[15] M. Zhang, Z. Miao, L. Fan, and S. Shah, "Data-driven interarea oscillation analysis for a 100% IBR-penetrated power grid," *IEEE Open Access J. Power Energy*, vol. 10, pp. 93–103, 2022.

[16] K. Hatipoglu, M. Olama, and Y. Xue, "Model-free dynamic voltage control of distributed energy resource (DER)-based microgrids," *Energies*, vol. 13, no. 15, 2020, Art. no. 3838.

[17] Z. Hou and S. Xiong, "On model-free adaptive control and its stability analysis," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4555–4569, Nov. 2019.

[18] A. Hussain, A. O. Rousis, I. Konstantelos, G. Strbac, J. Jeon, and H.-M. Kim, "Impact of uncertainties on resilient operation of microgrids: A data-driven approach," *IEEE Access*, vol. 7, pp. 14924–14937, 2019.
[19] J. Lee, G. Jang, E. Muljadi, F. Blaabjerg, Z. Chen, and Y. C. Kang, "Stable short-term frequency support using adaptive gains for a DFIG-based wind power plant," *IEEE Trans. Energy Convers.*, vol. 31, no. 3, pp. 1068–1079, Sep. 2016.
[20] D. Li, Q. Zhu, S. Lin, and X. Y. Bian, "A self-adaptive inertia and damping combination control of VSG to support frequency stability," *IEEE Trans. Energy Convers.*, vol. 32, no. 1, pp. 397–398, Mar. 2017.
[21] J. Li, B. Wen, and H. Wang, "Adaptive virtual inertia control strategy of VSG for micro-grid based on improved bang-bang control strategy," *IEEE Access*, vol. 7, pp. 39509–39514, 2019.
[22] L. A. Zadeh, "Fuzzy sets," *Inform. Control*, vol. 8, pp. 338–353, 1965.
[23] Y. Hu, W. Wei, Y. Peng, and J. Lei, "Fuzzy virtual inertia control for virtual synchronous generator," in *Proc. IEEE 35th Chin. Control Conf.*, 2016, pp. 8523–8527.
[24] A. Karimi et al., "Inertia response improvement in AC microgrids: A fuzzy-based virtual synchronous generator control," *IEEE Trans. Power Electron.*, vol. 35, no. 4, pp. 4321–4331, Apr. 2020.
[25] O. Oboreh-Snapps, R. Bo, B. She, F. F. Li, and H. Cui, "Improving virtual synchronous generator control in microgrids using fuzzy logic control," in *Proc. IEEE/IAS Ind. Commercial Power Syst. Asia*, 2022, pp. 433–438.
[26] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA:A Bradford Book, 2018.
[27] J. Hao, D. W. Gao, and J. J. Zhang, "Reinforcement learning for building energy optimization through controlling of central HVAC system," *IEEE Open Access J. Power Energy*, vol. 7, pp. 320–328, 2020.
[28] K. Zhang, C. Zhang, Z. Xu, S. Ye, Q. Liu, and Z. Lu, "A virtual synchronous generator control strategy with Q-learning to damp low frequency oscillation," in *Proc. Asia Energy Elect. Eng. Symp.*, 2020, pp. 111–115.
[29] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[30] Y. Li et al., "Data-driven optimal control strategy for virtual synchronous generator via deep reinforcement learning approach," *J. Modern Power Syst. Clean Energy*, vol. 9, no. 4, pp. 919–929, 2021.
[31] K. Xiong, W. Hu, G. Zhang, Z. Zhang, and Z. Chen, "Deep reinforcement learning based parameter self-tuning control strategy for VSG," *Energy Rep.*, vol. 8, pp. 219–226, 2022.
[32] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A lyapunov-based approach to safe reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8103–8112.
[33] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 908–919.
[34] K. Sakimoto, Y. Miura, and T. Ise, "Stabilization of a power system with a distributed generator by a virtual synchronous generator function," in *Proc. IEEE 8th Int. Conf. Power Electron.*, 2011, pp. 1498–1505.
[35] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2018, pp. 1587–1596.
[36] B. She, F. Li, H. Cui, J. Zhang, and R. Bo, "Fusion of microgrid control with model-free reinforcement learning: Review and vision," *IEEE Trans. Smart Grid*, vol. 14, no. 4, pp. 3232–3245, Jul. 2023.
[37] B. She et al., "Inverter PQ control with trajectory tracking capability for microgrids based on physics-informed reinforcement learning," *IEEE Trans. Smart Grid*, early access, May 7, 2023, doi: 10.1109/TSG.2023.3277330.
[38] R.Salcedo et al., "Banshee distribution network benchmark and prototyping platform for hardware-in-the-loop integration of microgrid and device controllers," *J. Eng., vol.* 8, no. 8, pp. 5365–5373, 2019.

Oboreh-Snapps, Oroghene (Graduate Student Member, IEEE) received the B.Eng. degree in electrical engineering from Madonna University, Livonia, Nigeria in 2016, and the master's degree in electrical engineering with an emphasis on control and power systems in 2019 from the Missouri University of Science and Technology, Rolla, MO, USA, where he is currently working toward the Ph.D. degree in electrical engineering with an emphasis on control and power systems. His research interests include microgrid control, deep reinforcement learning, renewable energy and electric vehicle integration, power electronics control, and power system stability. In 2022, he was the recipient of the Prestigious College of Engineering Deans' Graduate Educator Award for his excellent performance as a Graduate Teaching Assistant.



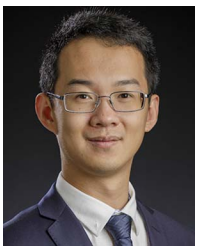
Buxin She (Graduate Student Member, IEEE) received the B.S.E.E. and M.S.E.E. degrees from Tianjin University, Tianjin, China, in 2017 and 2019, respectively. He is currently working toward the Ph.D. degree with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA. His research interests include microgrid operation and control, machine learning in power systems, distribution system plan, and power grid resilience. He was an Outstanding Reviewer of MPCE and IEEE Open Access Journal of Power and Energy. He is the Student Guest Editor of IET-RPG.

Shah Fahad received B.S. degree in electrical (power) engineering from COMSATS University, Abbottabad, Pakistan, in 2015, the M.S. degree in electrical power and control from the CECOS University of IT and Emerging Sciences, Pakistan, in 2018, and the Ph.D. degree in electrical engineering with the College of Electrical Engineering, Zhejiang University, Hangzhou, China, in 2022. He is currently a Postdoctoral Research Scholar with the Missouri University of Science & Technology, Rolla, MO, USA. His research interests include coordination control of converter-interfaced distributed generators, hierarchical control of microgrids, robust control, and applications of artificial intelligence/machine learning aided control techniques in RESs based Power systems. In 2018, he won a fully-funded scholarship from the CSC to support his Ph.D. studies. In March 2022, he was selected for a fully funded Visiting Researcher position with Qatar University, Qatar.

Haotian Chen (Graduate Student Member, IEEE) received the B.S. degree in physical science from the University of Science and Technology of China, Hefei, China, in 2013, and the M.S. degree in electrical & computer engineering from the Florida Institute of Technology, Melbourne, FL, USA, in 2016. He is currently working toward the Ph.D. degree in electrical engineering with the Missouri University of Science & Technology, Rolla, MO, USA. His current interests include machine learning, developing strategies for agents in electricity markets, and optimization, modeling, and computer programming applications.

Jonathan W. Kimball (Senior Member, IEEE) received the B.S. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 1994, the M.S. degree in electrical engineering, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1996 and 2007, respectively. From 1996 to 1998, he worked for Motorola, Phoenix, AZ, designing IGBT modules for industrial applications. He then joined Baldor Electric, Fort Smith, AR, where he designed industrial adjustable speed drives ranging 1–150 hp. In 2003, he returned to Illinois as a Research Engineer (later a Senior Research Engineer). Later in 2003, he co-founded SmartSpark Energy Systems, Inc., in Champaign, IL, and was the Vice President of Engineering. In 2008, he joined the Faculty of Missouri S&T (formerly the University of Missouri-Rolla), as an Assistant Professor. He was promoted to Associate Professor in 2014 and to Professor of electrical and computer engineering in 2018. From 2016 to 2018, he was also the Dean's Scholar of the College of Engineering and Computing. From 2019 to 2022, he was the Director of Missouri S&T's Center for Research in energy and the environment. In 2022, he was named chair of the Electrical and Computer Engineering Department and the Fred W. Finley Distinguished Professor of electrical and computer engineering. His research interests include microgrids, switched-capacitor converters, and cyber-physical systems. Dr. Kimball is a member of Eta Kappa Nu, Tau Beta Pi, and Phi Kappa Phi. He is a licensed Professional Engineer with the State of Illinois (license 062-057980). He was the General Chair of the IEEE Applied Power Electronics Conference in 2017 and continues to serve on its steering committee.

Fangxing Li (Fellow, IEEE) is also known as Fran Li. He received the B.S.E.E. and M.S.E.E. degrees from Southeast University, Nanjing, China, in 1994 and 1997, respectively, and the Ph.D. degree from Virginia Tech, Blacksburg, VA, USA, in 2001. He is currently the James W. McConnell Professor in electrical engineering at the University of Tennessee, Knoxville (UTK), TN, USA. He is also a Founding Member of CURENT, an NSF/DOE Engineering Research Center headquartered at UTK, and is the UTK Campus Director of CURENT. His research interests include resilience, artificial intelligence in power, demand response, distributed generation and microgrid, and electricity markets. From 2020 to 2021, he ws the Chair of IEEE PES Power System Operation, Planning, and Economics (PSOPE) Committee. He has been the Chair of IEEE WG on Machine Learning for Power Systems since 2019 and the Editor-In-Chief of IEEE OPEN ACCESS JOURNAL OF POWER AND ENERGY (OAJPE) since 2020. Prof. Li was the recipient of numerous awards and honors including R&D 100 Award in 2020, IEEE PES Technical Committee Prize Paper award in 2019, five best or prize paper awards at international journals, and six best papers/posters at international conferences.



Hantao Cui (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Southeast University, Nanjing, China, in 2011 and 2013, respectively, and the Ph.D. degree in electrical engineering from the University of Tennessee, Knoxville, TN, USA, in 2018. He is currently an Assistant Professor with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK, USA. His research interests include power system modeling, simulation, and high-performance computing.

Rui Bo (Senior Member, IEEE) received the B.S.E.E. and M.S.E.E. degrees in electric power engineering from Southeast University, Nanjing, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Tennessee, Knoxville, TN, USA, in 2009. He is currently an Associate Professor of the Electrical and Computer Engineering Departmen, Missouri University of Science and Technology (formerly the University of Missouri-Rolla), Rolla, MO, USA. He was a Principal Engineer and Project Manager with Midcontinent Independent System Operator (MISO) from 2009 to 2017. His research interests include computation, optimization and economics in power system operation and planning, high performance computing, electricity market simulation, evaluation, and design. He is an Associate Editor for IEEE TRANSACTIONS ON POWER SYSTEMS, and IEEE TRANSACTIONS ON ENERGY MARKETS, POLICY AND REGULATION. He is currently the Chair of IEEE Power and Energy Society (PES) Bulk Power System Planning Subcommittee, and the secretary of IEEE PES Power System Economics Subcommittee. He was the recipient of number of awards and honors including 2020 University of Missouri System President's Award for Career Excellence - Early Career, 2020 Missouri S&T Faculty Excellence Award, 2018 DARPA Young Faculty Award, and 2015 MISO Outstanding Achievement Award.