Model-Free Emergency Frequency Control Based on Reinforcement Learning

Chunyu Chen, *Member IEEE*, Mingjian Cui, *Senior Member, IEEE*, Fangxing Li, *Fellow IEEE*, Shengfei Yin, *Student Member, IEEE*, Xinan Wang, *Student Member, IEEE*,

***Abstract*—Unexpected large power surges will cause instantaneous grid shock, and thus emergency control plans must be implemented to prevent the system from collapsing. In this paper, by the aid of reinforcement learning (RL), novel model-free control (MFC)-based emergency control schemes are presented. Firstly, multi-Q-learning-based emergency plans are designed for limited emergency scenarios by using offline-training-online-approximation methods. To solve the more general multi-scenario emergency control problem, a deep deterministic policy gradient (DDPG) algorithm is adopted to learn near-optimal solutions. With the aid of deep Q network, DDPG-based strategies have better generalization abilities for unknown and untrained emergency scenarios, and thus are suitable for multi-scenario learning. Through simulations using benchmark systems, the proposed schemes are proven to achieve satisfactory performances.**

***Index Terms*—emergency frequency control, reinforcement learning, model-free control, deep Q network, deep deterministic policy gradient**

I. Introduction

Electrical power systems may sometimes experience significant disturbances [1], which can cause unexpected power surges and imbalances. Rapid corrective or emergency control actions must be applied in case of safety damages resulting from continuous abnormal operating conditions. The goal for power system emergency frequency control (PSEFC) is to quickly maintain system frequency at an acceptable level after large power disturbances.

Practically, control actions such as generator tripping and load shedding are widely used in emergency control. Dependent on control goals, load shedding can be categorized by under-voltage load shedding [2] and under-frequency load shedding (UFLS). Based on the control mode, load shedding schemes can be categorized into multi-stage [3], adaptive [4], [5] and semi-adaptive ones [6]. Conventionally, UFLS is implemented in a sequential offline-predetermination-online-application manner. Nevertheless, these schemes fail to achieve meticulous compensation based on the severity of disturbances. Therefore, researchers use system parameters and measurements to estimate the imbalance, thus shedding load adaptively [4]–[6]. However, the main problems of this adaptive mode lie in handling noisy raw data and information inconsistency among different generating units. Consequently, a multi-agent system is used to obtain global operating conditions information (i.e., the magnitude of total active power imbalance) [7], followed by a coordinated load shedding process. In addition, some researchers consider emergency control of microgrids [8], [9] and emergency control using fast-response inverter-based distributed energy resources [9] or large-scale electric vehicles [10]. With the recent development of the wide-area monitoring system, short-term prediction is also combined with UFLS to predetermine the amount of load shed [11], [12].

Unlike uncertain small disturbances, significant disturbances are anticipated known information to system operators with certain statistical regularity [13]. Therefore, if schemes corresponding to known scenarios can be predetermined, these predetermined schemes for the most similar off-line disturbance scenarios can be instantly executed online. This off-line-predetermination-online-practice (OPOP) mode offers more adaptability and accuracy of control.

In this paper, the OPOP mode is incarnated into reinforcement learning (RL)-based model-free control (MFC). RL is a popular technique in machine learning [14], and off-policy RL (Q-learning) is widely studied for various industry applications [15]. In general the Q-learning approach is computationally challenging, deep Q network relieves the dimensional burdens by using deep neural networks to estimate Q values [16]. Based on the baseline deep Q network, various extensions ranging from a double deep Q network to advanced sampling are designed to improve data efficiency or performances [17], [18]. Furthermore, the actor-critic technique is combined with a deep Q network for RL with continuous actions. A deep deterministic policy gradient (DDPG) method is thus developed and applied in automatic assembly task [19] and energy management [20].

In power system applications, RL-based MFC also wields great power [21]–[23]. Due to the complexity and multiple inertial components of power systems, on-policy RL can barely satisfy the control speed requirement. Therefore, off-policy RL (Q-learning) is used for OPOP. Since multiple agents could be participating in emergency control, the multi-agent RL method is adopted. Furthermore, due to the requirements of coordination and asynchronous machine speed response caused by the difference actions, single-agent RL is modified to adapt the learning requirements of multi-agent scenarios.

Since various disturbances such as generator tripping or load change could happen with different size or timing, multiple emergency scenarios would be generated. Pure Q-learning requires space discretization, and the cost could grow exponentially when the number of scenarios becomes large. Therefore, the continuous action and state spaces in emergency control are also considered by using a DDPG algorithm.

This work is supported by the 'Fundamental Research Funds for the Central Universities' (2020QN62). Mingjian Cui is the corresponding author.

The main contributions of this work is summarized as follows:

- A systematic PSEFC framework using off-policy RL-based MFC is presented. Under the designed framework optimal control actions can be trained for various scenarios. Multi-Q learning is used for RL under multiple limited scenarios, and the predetermined actions which correspond to the most similar pre-trained scenarios when certain emergency scenarios occur online are immediately executed, providing flexibility and convenience of application lacking in other methods.
- The multiple agents which correspond to multiple regulation means are aggregated to a single agent to reduce the high dimensionality and allow the synchronousness of execution, thus guaranteeing the learning efficiency and angle stability.
- A DDPG is adopted for continuous emergency frequency control under various scenarios by using a deep Q network. This method fills in the gap in multi-Q-learning, which would generate considerable costs by using tabular methods.
- Benchmark power system models, which can reflect the real power system dynamics more than simplified models, are constructed to test the effectiveness of the proposed scheme.

The remainder of the paper is as follows: Section II defines the problem and models of the environment in RL; the design of the PSEFC scheme using an off-policy RL technique is given in Section III, and both single agent and multi-agent-based RL are studied in the controller design; simulations using standard test systems is given in Section IV; concluding remarks are given in Section V.

II. Preliminaries

First, the problem definition of PSEFC is discussed, followed by the introduction of the fundamental environment upon which the proposed RL agents learn the strategies.

A. Problem Definition

PSEFC aims to maintain the system following a major disturbance with the aid of load shedding techniques. In the context of RL-based PSEFC, load shedding schemes are achieved by RL and the schematic diagram is shown in Fig. 1. The eventual load shedding strategy in Fig. 1 is model-independent and purely data-driven once the RL agent is well-trained. It means that the model information is not required for the controller design.

B. Multi-Unit High Order Power System Model-Based Environment

Though the strategy is model-free, the model used to simulate the environment with which the agent interacts should still be built.

The system frequency response (SFR) or equivalent unit model is prevalently adopted to analyze the frequency response and design frequency controller [24]. Though simple, the system cannot completely reflect the dynamic characteristics of real-life systems [25]. Therefore, more detailed models that reflect the real systems more than the SFR model are built for the study.

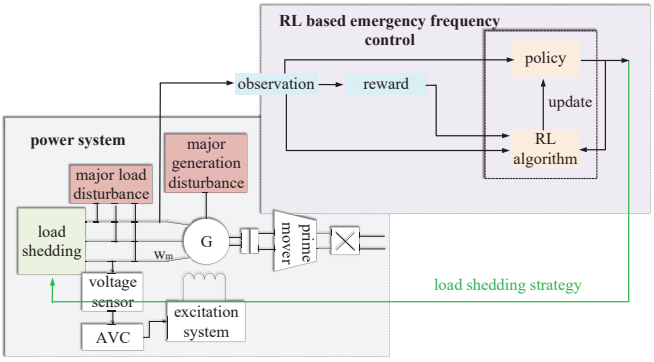


Fig. 1. Schematic diagram of RL-based emergency frequency control

The practical system contains multiple complex generating units including the auxiliary control systems and the network, meaning that a more practical power system should be characterized by the multi-input-multi-output high-order nonlinear system.

1) Electromechanical Model of Units: In this paper, the classic 7-order system containing both governor and excitation control systems is adopted:

$$\begin{cases} \dot{E}_{qi}^t = \frac{1}{T_{d0i}}\left[-E_{qi}^t + \left(x_{di} - x_{di}^t\right) I_{di} + E_{fi}\right] \\ \dot{E}_{di}^t = \frac{1}{T_{q0i}}\left[-E_{di}^t + \left(x_{qi} - x_{qi}^t\right) I_{qi}\right] \\ \dot{\delta}_i = \omega_i - \omega_0 \\ \dot{\omega}_i = \frac{\omega_0}{H_i}\left[P_{mi} - \left(E_{qi}^t I_{qi} + E_{di}^t I_{di} - \left(x_{di}^t - x_{qi}^t\right) I_{di} I_{qi}\right)\right] \\ \dot{P}_{mi} = \frac{1}{T_{ti}}\left(-P_{mi} + P_{gi}\right) \\ \dot{P}_{gi} = \frac{1}{T_{gi}}\left(-P_{gi} - \frac{(\omega_i - \omega_0)}{R_i} + u_{gi}\right) \\ \dot{E}_{fi} = -\frac{E_{fi}}{T_{ei}} + \frac{K_{ei}}{T_{ei}}\left(V_{ti}^r - V_{ti}\right) \end{cases} \tag{1}$$

where E_{qi}^t and E_{di}^t represent q-axis and d-axis components of the voltage behind transient reactance x_{qi}^t and x_{di}^t, respectively; I_{di} and I_{qi} represent q-axis and d-axis components of the current, respectively; T_{d0i} and T_{q0i} represent the open-circuit transient time constants, respectively; δ_i represents the machine angle; ω_i represents the machine speed; ω_0 represents the nominal machine speed; H_i represents the inertia constant; P_{mi} represents the mechanical power; P_{gi} represents the governor states; T_{ti} and T_{gi} represent the time constants in the turbine and governor, respectively; R_i represents the droop coefficient; u_{gi} represents the control input of the governor; E_{fi} represents the excitation voltage; T_{ei} and K_{ei} represent the control coefficients in the excitation system; V_{ti} represents the terminal voltage; and V_{ti}^r represents the reference terminal voltage.

2) Interface Model: The interface model contains stator voltage, dq/xy transformation and power equation. Together they depict the interface relation between the interior unit and the exterior transmission network. The stator voltage is

described by:

$$\begin{cases} U_{di} = x_{qi} I_{qi} - r_{ai} I_{di} \\ U_{qi} = E'_{qi} - x'_{di} I_{di} - r_{ai} I_{qi} \end{cases} \tag{2}$$

where U_{di} represents d-axis terminal voltage; U_{di} represents q-axis terminal voltage. The terminal voltage is $V_{ti} = \sqrt{U_{di}^2 + U_{qi}^2}$.

Based on $dq - xy$ transformation, the relation between dq and xy axis voltage (current) is given by:

$$\begin{bmatrix} U_{xi} \\ U_{yi} \end{bmatrix} = \begin{bmatrix} \cos\delta_i & \sin\delta_i \\ \sin\delta_i & -\cos\delta_i \end{bmatrix} \begin{bmatrix} U_{qi} \\ U_{di} \end{bmatrix} \tag{3}$$

where U_{xi} represents x-axis terminal voltage; U_{yi} represents y-axis terminal voltage.

$$\begin{bmatrix} I_{xi} \\ I_{yi} \end{bmatrix} = \begin{bmatrix} \cos\delta_i & \sin\delta_i \\ \sin\delta_i & -\cos\delta_i \end{bmatrix} \begin{bmatrix} I_{qi} \\ I_{di} \end{bmatrix} \tag{4}$$

where I_{xi} represents x-axis terminal current; I_{yi} represents y-axis terminal current. Based on (3) and (4), the output power is

$$\begin{cases} P_i = U_{xi} I_{xi} - U_{yi} I_{yi} \\ Q_i = U_{xi} I_{yi} + U_{yi} I_{xi} \end{cases} \tag{5}$$

where P_i represents the active power; Q_i represents the reactive power.

3) Network Model: The network is modelled by nodal active and reactive power injection equations:

$$P_i + jQ_i = \sum\nolimits_k V_i V_k Y_{ik} e^{j(\theta_i - \theta_k - \phi_{ik})} \tag{6}$$

where V_i represents the magnitude of the voltage at bus i; θ_i represents the angle of the voltage at bus i; Y_{ik} represents the admittance between bus i and k. The nonlinear power system model is derived by combining (1) to (6):

$$\begin{cases} \dot{x} = f(x, u, z) \\ 0 = g(x, u, z) \end{cases} \tag{7}$$

where x represents the states; u represents the control input; z represents the auxiliary variables. (7) is the environment upon which the agent executes its actions and observes the states while using the temporal difference (TD) learning methods.

III. PSEFC Using Off-Policy RL-Based MFC Technique

After establishing the environment in Section II, the off-policy RL-based MFC can be used to design PSEFC schemes. Before formally presenting the design procedures, the feasibility of RL's application into PSEFC is discussed. Then, the design of the RL-based PSEFC scheme is systematically studied in this section.

A. Feasibility Study

It is well known that compensating the exact amount of power imbalance (even if it is predicted with sufficient accuracy) cannot make the frequency converge to equilibrium. To illustrate this phenomenon, Kundur's 4-unit-13-bus system in Fig. 2 is simulated.

Assume the sudden loss of $5p.u.$ load at bus 4 at $t = 10s$; the relay time delay is $0.1s$; then the generator correspondingly

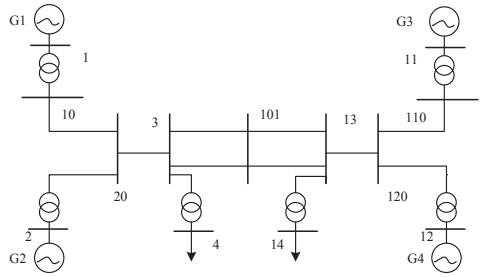


Fig. 2. Diagram of Kundur's 4-unit-13-bus system

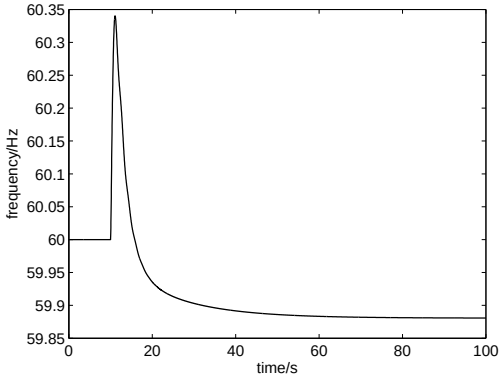


Fig. 3. Frequency response under equal compensation for Kundur's 4-unit-13-bus system

reduces the generation by $5p.u.$. As the SFR in Fig. 3 shows, there is an explicit frequency deviation, which is due to the fact that the real power balance can change dynamically during the compensation process. For example, the frequency-dependent load could dynamically change during the control process. Also, the input channels of the disturbance and control are not the same. Therefore, this 'equal' compensation strategy might not be effective enough to maintain the frequency, though it can improve the system frequency as shown in Fig. 3. The most desired outcome is that the frequency, via intelligent emergency controllers, is as close to the small neighborhoods of equilibrium as possible.

Also, it is challenging to achieve the exact imbalance computation in practice. For example, the prevalent adaptive methods are used to estimate the imbalance with the aid of the initial frequency change rate [4]. However, it is difficult to guarantee the accuracy of imbalance estimation due to the measurement noise and the challenge of capturing the initial time of the emergency's occurrence.

In this paper, the convergence of frequency is quantified as the reward and the controller is trained (using RL) to learn the actions to maximize this reward, which is equivalent to the minimization of the frequency deviation. The benefits of this methodology include:

- The timing of initial frequency change rate is not required in adaptive schemes is not needed.
- The action (load shedding) is in the form of optimal sequence and optimizes the overall control performance

in the control horizon, which cannot be achieved by conventional load shedding schemes.

The aforementioned presentation lays the foundation of the feasibility of RL's application in PSEFC.

B. Multi-Q-Learning Based Controller For Limited Emergency Scenarios

In this section, it is supposed that the following limited emergency scenarios would occur. There are two primary features of these scenarios:

- The operating conditions are comparatively more stable and significant disturbances seldom occur, e.g., the system is isolated and fluctuating resources including renewable energy are trivial or nonexistent.
- The operating conditions are somewhat known. The emergency scenarios can be predicted using modern communication and computing techniques. They possess certain statistical regularities.

Therefore, the OPOP can be achieved by multi-Q-learning under limited emergency scenarios.

1) Basics of The Control Scheme Based on Multi-Q-Learning: The general idea of multi-Q-learning is to learn optimal strategies under multiple off-line emergency scenarios and choose the strategies corresponding to the most similar off-line scenarios for the online scenarios, this method applies for situations where the precision requirement is low. That is to say, the off-line trained optimal policy can be online performed for the upcoming scenario which is similar to the off-line trained scenario. The essence of Algorithm 1 is matching the online unknown scenario with the off-line pre-trained ones, and the procedural form of the multi-Q-learning algorithm is as follows:

Algorithm 1 Multi-Q-learning based PSEFC for limited emergency scenarios

Initialization: Generate M limited emergency scenarios $\{e_i\}$ through simple random sampling.

Step 1. Train the agent off-line to obtain the optimal policy π_i^* for each emergency scenario e_i by using single-agent (Algorithm 2) or multi-agent Q-learning (Algorithm 3).

Step 2. Detect online emergency scenario e_j, identify the most similar off-line trained scenario e_i.

Step 3. Perform the pre-trained π_i^* of the most similar e_i for the online e_j.

The core of Algorithm 1 is the Q-learning agent design in Step 1. Q-learning transforms the original control problem into a dynamic optimization by modeling it as a Markov decision process. The aim is to maximize the expected discounted sum of rewards (Q values):

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \tag{8}$$

where s and a are the elements of the finite discrete set of the state $S=\{s\}$ and action $A=\{a\}$, respectively. $Q^{\pi}(s,a)$ can be expressed by:

$$Q^{\pi}(s,a) = E\left\{ \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \middle| s_t = s, a_t = a \right\} \tag{9}$$

where γ represents the discount coefficient; r_i represents the immediate reward at time i; and E is the mean operator. Eq. (8) can be further extended as:

$$\begin{aligned} Q^*(s,a) &= E\left\{ r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a') \middle| s_t = s, a_t = a \right\} \\ &= \sum_{s'} P_a(s,s') \left[R_a(s,s') + \gamma \max_{a'} Q^*(s',a') \right] \end{aligned} \tag{10}$$

Instead of using conventional dynamic programming which requires the knowledge of transition probability $P_a(s,s')$, TD method is used to solve (10) in a recursive manner, given by:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right] \tag{11}$$

The right hand side of (11) is used to approach state action values at each iteration, and action a is chosen by observing predefined rules (random or ε-greedy). Note that no information of the system model is required in (11), as it is a MFC problem. Based on the learning manner, on-policy Sarsa and off-policy Q-learning are developed to solve (11). The main difference between Sarsa and Q-learning is that the Q-function update policy and behavior policy are the same in Sarsa, while they are separate in Q-learning. There are also many extensions (e.g., actor-critic agent for continuous control and deep Q network for multi-scenario control) that are applicable for specific conditions. Nevertheless, they all essentially possess the recursive TD-based learning feature.

2) Single-Agent-Based Q-Learning: There are multiple generating units that can all participate in generation adjustment in the system, which implies that this is a multi-agent learning problem. Nevertheless, these multiple agents can still be regarded as a single agent to simplify the learning process. In this case, all the separate agents can be regarded as an aggregated agent, which assumes the role of the master agent in the process of RL, while the other slave agents (each separate agent corresponding to each generator) follow the command of the master agent, e.g., each slave agent equally shares the total adjustment (control action) of the master agent. The single agent-based RL control schematic is shown in Fig. 4. The off-policy Q-learning algorithm is then used to train the master agent. The procedural form of this algorithm is shown in Algorithm 2.

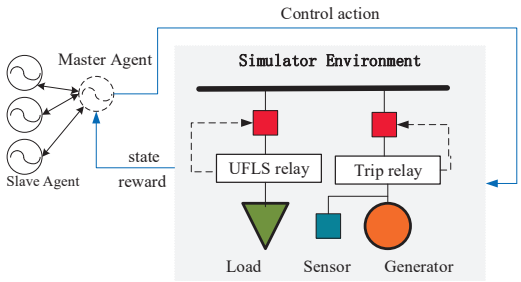


Fig. 4. Diagram of single agent-based learning structure

Algorithm 2 Off-Policy Single Agent-Based PSEFC Using Q-Learning

Initialization: Set $Q(s,a)$ ($\forall s \in \mathcal{S}$ $a \in \mathcal{A}$) and k to zero; choose values for λ, T_c, k_{max}.
Step 1. Observe the current state s_c; choose action a using ϵ-greedy methods.
Step 2. Execute a for T_c on the system in Section II, observe the next state s_n (the state values after time T_c), calculate the immediate reward $r(s_c, a, s_n)$.
Step 3. Update $Q(s,a)$

$$Q\left(s_c,a\right) \leftarrow \left(1-\alpha\right) Q\left(s_c,a\right) + \alpha\left[r + \lambda \max_{b \in \mathcal{A}(s_n)} Q\left(s_n,b\right)\right]$$

$k = k+1$
Step 4. If $k \leq k_{max}$, set $s_c \leftarrow s_n$, go to **Step 1.** Otherwise, go to **Step 5.**
Execution:
Step 5. For each $s \in \mathcal{S}$, select

$$a^*\left(s\right) \in \arg\max_{b \in \mathcal{A}(s)} Q\left(s,b\right)$$

Step 6. For each $s \in \mathcal{S}$, each generator executes $a^*\left(s\right)/N$.

The parameter of k_{max} is expected to be large to make the Q value converge, and can be obtained by trial and error (in this case the range from 5000 to 10000 is usually enough). N denotes the number of generators. The immediate reward r is defined by:

$$\begin{cases} r = 0 & \left|\frac{\sum_{T_c}|\Delta f_o|}{T_c}\right| \leq \eta \\ r = -1 & otherwise \end{cases} \tag{12}$$

where $\sum_{T_c}|\Delta f_o|/T_c$ denotes the average of the absolute center of inertia (COI) frequency deviation Δf_o over the control horizon T_c; η denotes the threshold value.

Other important elements in Algorithm 2 include the state and action set $\mathcal{S}$ and $\mathcal{A}$. $\mathcal{S}$ is defined by an evenly distributed discrete set of average COI frequency deviation $\{s_1, s_2, \cdots, s_m\}$, while $\mathcal{A}$ is defined by an evenly distributed discrete set of load shed $\{a_1, a_2, \cdots, a_n\}$.

3) Multi-Agent-Based Q-Learning: Multiple generators can also learn their respective emergency control actions in an interactive fashion, which essentially belongs to multi-agent reinforcement learning (MARL). The main challenge in MARL is to design the appropriate definitions of learning goal and interaction mechanism, which can either be cooperative, competitive or mixed. Though MARL has benefits such as experience sharing for a better performance, the curse of dimensionality, which is already present due to the growth of discrete state-action space, would further be enhanced by the increased number of agents. The multi-agent-based (taking 2- agent as an example) RL control schematic is shown in Fig. 5. In this section, it is supposed that all the agents form a cooperative relationship, which is understandable since they all aim at the same goal (namely frequency stabilization). All the agents, in this case, have the same reward function and the learning goal is to maximize the return [26]. Moreover, the optimal joint action at each state is assumed to be unique. There

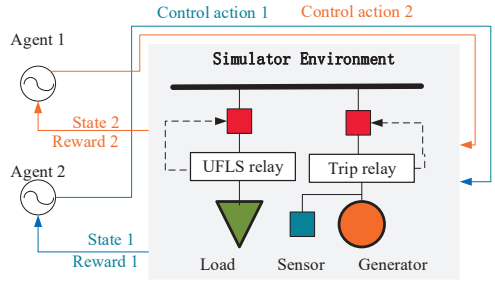


Fig. 5. Diagram of multiple agent-based learning structure

is a little need for coordination to obtain the fixed optimal joint action among multiple optimal joint actions. Team Q-learning is adopted to solve this problem. The procedure of this algorithm is shown in Algorithm 3.

Algorithm 3 Off-Policy Multi-Agent-Based PSEFC Using Q-Learning

Initialization: Set $Q(s,\mathbf{a})$ ($\forall s \in \mathcal{S}$ $\mathbf{a} \in \{\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_N\}$) and k to zero; choose values for λ, T_c, k_{max}.
Step 1. Observe the current state s_c; choose action $\mathbf{a}^*$ as $\mathbf{a}^* = \arg\max Q\left(s_c,\mathbf{a}\right)$.
Step 2. Execute $\mathbf{a}^*$ for T_c on the system in Section II, observe the next state s_n, calculate the immediate reward $r(s_c, \mathbf{a}^*, s_n)$.
Step 3. Update $Q(s,a)$

$$Q\left(s_c,\mathbf{a}^*\right) \leftarrow \left(1-\alpha\right) Q\left(s_c,\mathbf{a}^*\right) + \alpha\left[r + \lambda \max_{\mathbf{b} \in \mathcal{A}(s)} Q\left(s_n,\mathbf{b}\right)\right]$$

$k = k+1$
Step 4. If $k \leq k_{max}$, set $s_c \leftarrow s_n$, go to **Step 1.** Otherwise, go to **Step 5.**
Execution:
Step 5. For each $s \in \mathcal{S}$, the generators execute:

$$\mathbf{a}^*\left(s\right) \in \arg\max_{\mathbf{b} \in \mathcal{A}(s)} Q\left(s,\mathbf{b}\right)$$

The average operator in Step 6 of Algorithm 2 is neglected herein. All the actions $\mathbf{a}$ (load shedding amount) are determined during the TD learning process. The remaining parameters and settings are the same as (12) and omitted here for brevity.

C. DRL-Based Controller For Multiple Emergency Scenarios

In Section III-B, PSEFC for limited emergency scenarios is addressed. In this section, the limited emergency scenarios are extended to multiple various ones. Compared with the scenario in Section III-B, the main differences include:

- The emergency scenarios are various. Due to uncertain factors including irregular consumer behavior, unexpected generator fault and integration of renewable energy, there are a great number of emergency scenarios.

- The emergency scenarios are unknown. Not only are the emergency scenarios various, they are unpredictable as well.

The aforementioned features cause the high dimensionality of learning. Moreover, the discretization in Q-learning might not effectively meet the requirement of continuous control. Therefore, deep reinforcement learning (DRL) is used to improve the performance from the following two aspects:

- The generalization ability of deep neural networks in DRL can handle the uncertain scenario learning.
- The actor-critic structure in DRL can handle continuous action through training the actor using a policy gradient.

The principle of DRL-based PSEFC is shown in Fig. 6. The

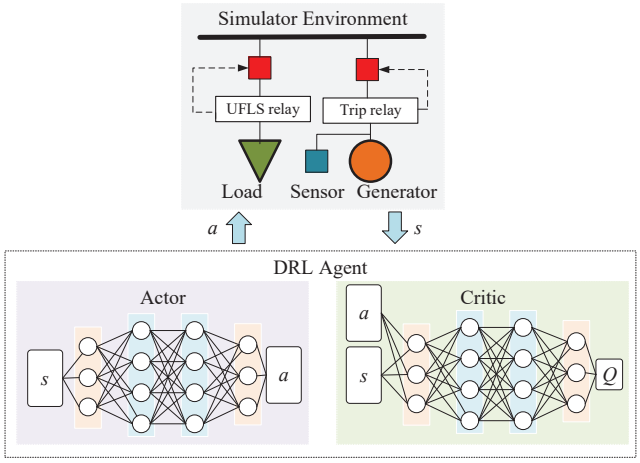


Fig. 6. Diagram of DRL-based PSEFC

actor and critic in Fig. 6 are both modeled by a deep neural network $\mu\left(s/\theta^{\mu}\right)$ and $Q\left(s,a/\theta^{Q}\right)$, respectively. During the training, the actor uses the network $\mu\left(s/\theta^{\mu}\right)$ to output continuous actions, which are judged by the critic through Bellman equations. The actor updates the policy parameter θ^{μ} by policy gradient:

$$\nabla_{\theta^{\mu}} J = E_s\left[\nabla_{\theta^{\mu}} Q\left(s,a/\theta^{Q}\right)\big/_{s=s_t,a=\mu(s_t/\theta^{\mu})}\right] \tag{13}$$

while the critic updates θ^{Q} by:

$$\nabla_{\theta^{Q}} L = E\left[\begin{array}{l}\left(y-Q\left(s,a/\theta^{Q}\right)\right)\\ \cdot\left[\nabla_{\theta^{Q}} Q\left(s,a/\theta^{Q}\right)\big/_{s=s_t,a=\mu(s_t/\theta^{\mu})}\right]\end{array}\right] \tag{14}$$

where

$$y = r+\gamma Q'\left(s',\mu\left(s'\Big/\theta^{\mu'}\right)\Big/\theta^{Q'}\right)$$

The details of DRL (DDPG) are given in [27] and not repeated here.

IV. Case Studies

In Section III, both multi-Q-learning-based and DRL-based PSEFC schemes are addressed. To vividly demonstrate their performances, Kundur's 4-unit-13 bus system and the New England 68-bus system are used to verify the effectiveness of the proposed schemes. The specific simulation model information can be found in [28]. In the remainder of this section, the demonstration using Kundur's system is detailed in Section IV-A; while the New England 68-bus system is tested and analyzed in Section IV-B.

A. RL-Based PSEFC Scheme Simulation Using Kundur's System

Before the simulation, the emergency scenarios should be constructed. The scenarios considered are defined as follows:

- **Scenario 1:** The unit at Bus 1 loses $0.53p.u.$ generation.
- **Scenario 2:** The unit at Bus 2 loses $0.82p.u.$ generation.

1) Scenario 1 In Kundur's System Using Multi-Q-Learning: Based on Algorithm 1, M off-line scenarios should be trained. For brevity, these scenarios are defined by 11 different amounts of generation loss at Bus 1: from $0.1p.u.$ to $1p.u.$) with a step size of $0.1p.u.$. Hence, 11 agents should be trained for each scenario.

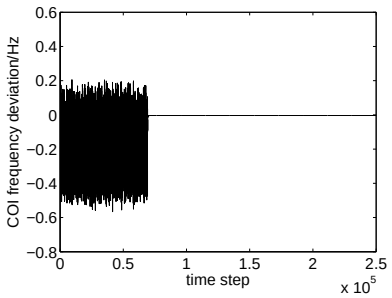
Before training the agent, $\mathcal{A}$ and $\mathcal{S}$ should be chosen. It is comparatively easy to choose the boundary (maximum) of $\mathcal{A}$ or $\mathcal{S}$. The former should be at the proximity of the significant disturbance, while the later can be identified by observing the maximum COI frequency deviation after simulating the joint-action of both the significant disturbance and control actions at the boundary. Thus, all the possible frequency deviations can be contained in $\mathcal{S}$.

The step size for discretizing $\mathcal{A}$ ($\mathcal{S}$) should also be set. The step size is expected to be small to consider all the states (actions) which could possibly occur during the dynamic process. Nevertheless, this would cause over-calculation due to the large dimensional $\mathcal{A}$ ($\mathcal{S}$). Since it is efficiency rather than accuracy that is emphasized in the multi-Q-learning based PSEFC scheme, $\mathcal{A}$ ($\mathcal{S}$) could be coarsely discretized for learning, so that COI frequency deviation under the trained controller would remain at the proximity of zero rather than reach zero exactly. After all, there are still other non-emergency frequency controllers (automatic generation control) which can further regulate system frequency.

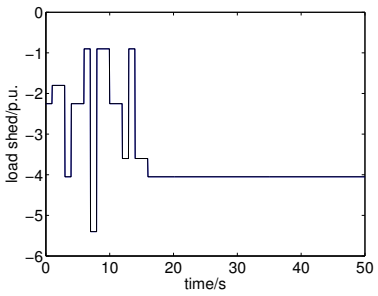
Based on the aforementioned rules, the state and action sets of the scenario where the generation loss is $0.5p.u.$ are set by: $\mathcal{A}:\{-0.6\times 9:0.05\times 9:0.1\times 9\}$; $\mathcal{S}:\{-0.4:0.05:0.4\}$. The usage of multiplier 9 is due to the difference in MVA of the generator and load side (the former is $900MVA$ while the latter is $100MVA$). Similarly, the sets for other scenarios can be achieved.

The training results, control action sequences, and machine speed deviations using payoffs in (12) are shown in Fig. 7.

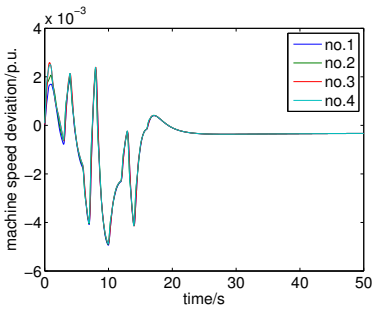
Fig. 7a shows the dynamic COI frequency response during the learning phase of one specific scenario, which converges to the equilibrium after sufficient time steps. As can be seen in Fig. 7b and 7c, both the control action sequences and machine speeds under the optimal policies for the most similar off-line emergency scenarios ($0.5p.u.$ generation loss at Bus 1) can be stabilized for online Scenario 1. This indicates that the single Q-learning agent has certain robustness for the approximate scenarios, which can be analogized to the robustness of the control system under uncertainties.



(a) Learning phase of single agent-based controller with payoff



(b) Control action of each agent with single agent-based controller



(c) Machine speed deviation of each generator with single agent-based controller

Fig. 7. Simulation results under Scenario 1 for Kundur's system using multi-Q-learning

2) Scenario 2 In Kundur's System: Remember that there exist two load buses (4 and 14) in Kundur's system. Instead of using the single agent-based Q-learning agent, in this section, a multi-agent-based Q-learning method is used for the two agents (load bus 4 and 14). The training results, control action sequences, and machine speed deviations using Algorithm 3 are shown in Fig. 8, respectively.

As can be seen in Fig 8, multi-agent Q-learning based optimal policies for the off-line emergency scenario ($0.8p.u.$ generation loss at bus 2) achieve satisfactory control performance for the online Scenario 2. It is clearly shown that each agent executes separate control action instead of the identical one when adopting Algorithm 3. Due to the difference of the control action sequences, the trembling of the machine speeds are unavoidable in Fig. 8c, which might be undesirable when considering power angle stability. The single agent-based scheme is thus preferred for emergency frequency control.

In the following, a deep reinforcement learning-based controller is simulated. The reward is defined by:

$$\begin{aligned} r = &-50\left|\Delta f_o\right| \\ &-8000\left(\left|\Delta f_o\right| \geq \lambda_1 or \left|\int \Delta f_o\right| \geq \lambda_2\right) - 25\left|\tfrac{d\Delta f_o}{dt}\right| \end{aligned} \tag{15}$$

where $\left(\left|\Delta f_o\right| \geq \lambda_1 or \left|\int \Delta f_o\right| \geq \lambda_2\right)$ represents the out-of-limit penalty term. λ_1 and λ_2 represent the threshold, the values of which are computed by computing either $\left|\Delta f_o\right|$ and $\left|\int \Delta f_o\right|$ with the boundary conditions, which is similar to establishing the boundary of $\mathcal{A}$ and $\mathcal{S}$ in a multi-Q-learning based controller. The coefficient -8000 means that the reward would be heavily penalized if the 'OR' condition holds. The last term $\left|\frac{d\Delta f_o}{dt}\right|$ is used to prevent low frequency oscillation. The structure of the critic network is shown in Figure 9.

The FC in Fig. 9 represents fully connected networks and Relu represents that the activation is a rectified linear unit. Similarly, the single-layer actor-network can be built. The simulation time is $100s$ for each episode, and the sample time is $0.05s$. In each episode, a significant disturbance, which is defined by the generation loss of the generator, is simulated with the range of $0.5p.u.$ to $1p.u.$. For each episode, the simulation proceeds until either the out-of-limit condition is triggered or the simulation time reaches $100s$. The moving average of the reward is shown in Fig. 10, which shows that after approximately 40 episodes, the model converges to the optimal solution. After training the model, a random disturbance scenario is tested by setting the generation loss as $0.7p.u.$ at $1s$, and the results in Fig. 11 show that the generalization ability of the trained model is satisfactory with a good frequency response. Also, compared with Fig. 8, the action is much smoother due to the non-discretization learning.

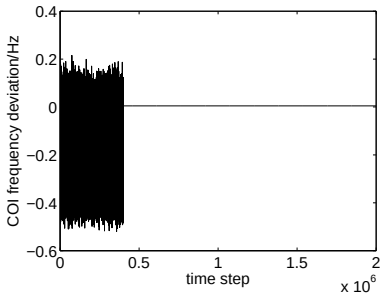
B. RL-Based PSEFC Scheme Simulation Using IEEE Standard System

To further verify the effectiveness of the proposed schemes for bulk power systems, the New England 68-bus system is tested. Due to page limit, detailed information of the New England system, which can be found in [29], is omitted in this paper. It is supposed that only the first 10 loads, i.e., the load bus 1, 3, 4, 7, 8, 9, 15, 16, 18 and 20 participate in load shedding. The online emergency scenario is defined as $5.2p.u.$ loss of generation at bus 63.

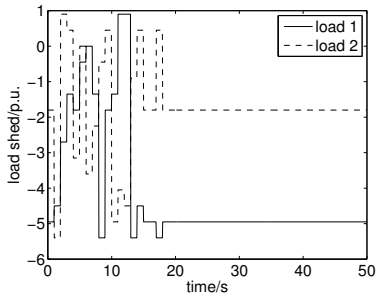
Firstly, the multi-Q-learning based controller is simulated. For brevity, the M off-line scenarios should be trained, and these scenarios are defined by 21 different amounts of generation loss at Bus 1: from $4p.u.$ to $6p.u.$) with a step size of $0.1p.u.$. Hence, 21 agents should be trained for each scenario.

The simulation results are shown in Fig. 12. Due to the coordination issue, only single-agent RL in Algorithm 2 is tested. Fig. 12a represents the dynamic COI frequency response during the learning phase of one specific scenario, which converges to the equilibrium after sufficient time steps. Fig. 12b shows the dynamic action sequences during the online operation. Fig. 12c shows the dynamic machine speed response of each generator. As can be seen, acceptable frequency regulation performance with rapid convergence velocity is achieved for the online scenario under the optimal policies of the off-line $5p.u.$ scenario. As is known, sluggish control speed is a problem for bulk power systems with multiple units due to coordination. Based on the results, the control speed in the 68-bus system is found to be on a par with that in the 4-unit (Kundur's) system, which proves that the single-agent RL-based PSEFC scheme can effectively adapt to the requirement of fast response in emergency scenarios.

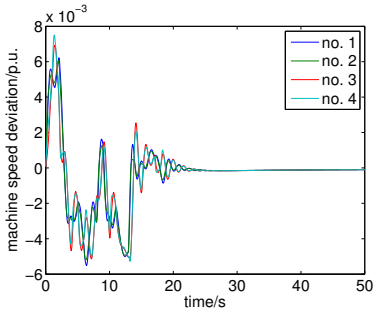
In the following, the DRL-based controller is simulated for the 68-bus system. The design procedure of the agent is similar to that in Section IV-A and omitted for brevity. After training



(a) Learning phase of multi-agent-based controller



(b) Control action of each agent with multi-agent based controller



(c) Machine speed deviation of each generator with multi-agent based controller

Fig. 8. Simulation results under Scenario 2 for Kundur's system using multi-Q-learning

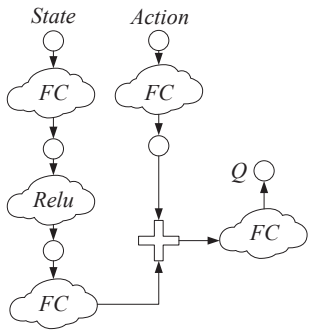


Fig. 9. Scheme diagram of the critic network

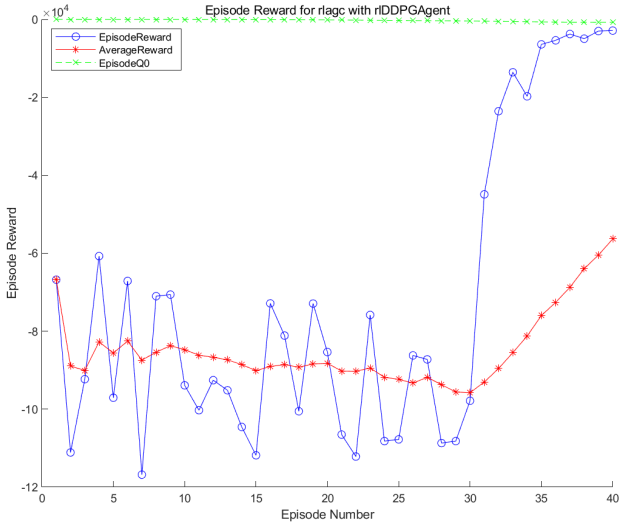


Fig. 10. The moving average of rewards during training

the model, a random disturbance is simulated by $3p.u.$ loss of generation of the first generator, and the results are shown in Fig. 13. As can be seen, machine speeds of all generators can be quickly regulated to the nominal values. Eventually, the proposed methods are compared with the staged scheme [4] and the scenario is set by the loss of $2.6p.u.$ of the generation at generator 13. The results are shown in Fig. 14a and 14b, respectively, where both proposed strategies perform better than the staged one: though the latter responds faster, the deviation is larger. To test the robustness of the proposed methods, the stability criterion C_s is given by:

$$\begin{cases} C_s = 1, & if \ \ |\Delta f_{COI}(t_d)| \leq \eta_0 \\ C_s = 0, & if \ \ |\Delta f_{COI}(t_d)| > \eta_0 \end{cases}$$

where $\Delta f_{COI}(t_d)$ represents the COI frequency deviation at the end time of each round of simulation. $C_s = 1$ means the deviation is acceptable and $C_s = 0$ is otherwise. η_0 is chosen as the maximal tolerable deviation which is $0.05Hz$ herein. 800 random scenarios are tested with major disturbance d conforming to the normal distribution $(\mu = 2, \sigma = 1)$ at generators 13 to 16 (each with 200). The robustness performance using the stability criterion C_s is verified by the acceptable $0-1$ ratio of the bar chart in Fig. 14c.

V. Conclusions

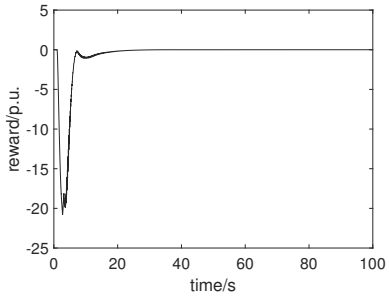
In this paper, novel MFC-based emergency frequency control schemes are designed by the aid of reinforcement learning techniques. Both multi-Q-learning and deep reinforcement learning (DRL) techniques are employed. The selection of either multi-Q-learning or DRL-based strategies depends on the capacity of the decision-maker. Due to the considerable computational costs resulting from the gradients update of multi-parametric deep neural networks, discrete tabular methods are more appealing to the decision-maker who has limited computational resources and memory. Nevertheless, DRL can be adopted with enhanced capacity. No method can gain an edge over the other absolutely in terms of the performance or operating cost; but it is clear that both can handle multi-scenario emergency frequency control with acceptable performances.

References

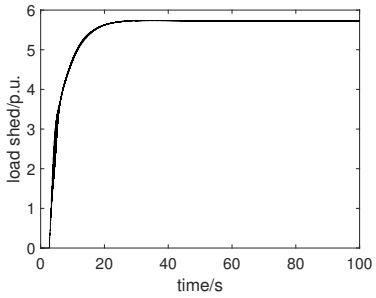
[1] L. Che, X. Liu, and Z. Shuai, "Optimal transmission overloads mitigation following disturbances in power systems," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 2592–2604, 2018.

[2] A. C. Adewole, R. Tzoneva, and A. Apostolov, "Adaptive under-voltage load shedding scheme for large interconnected smart grids based on wide area synchrophasor measurements," *IET Generation, Transmission & Distribution*, vol. 10, no. 8, pp. 1957–1968, 2016.

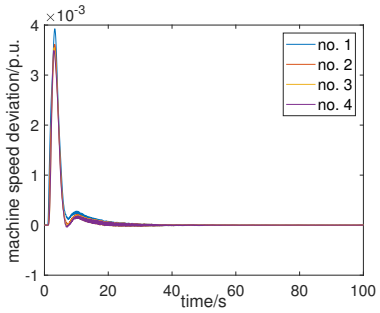
[3] M. G. Darebaghi and T. Amraee, "Dynamic multi-stage under frequency load shedding considering uncertainty of generation loss," *IET Generation, Transmission & Distribution*, vol. 11, no. 13, pp. 3202–3209, 2016.



(a) Dynamic payoff response

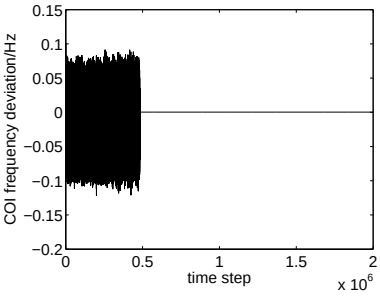


(b) Control action of each agent

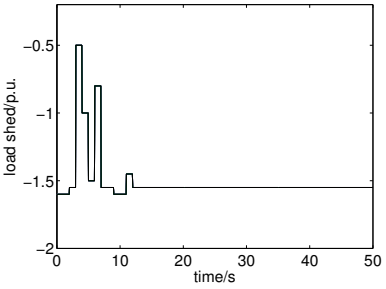


(c) Machine speed deviation of each generator

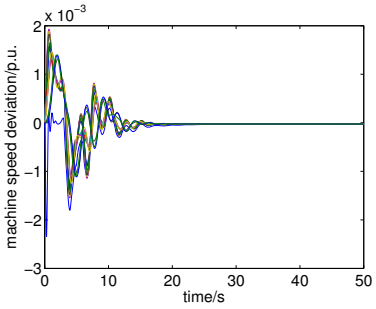
Fig. 11. Simulation results for Kundur's 4-unit-13-bus system using DDPG



(a) Learning phase of RL-based controller with payoff (12)

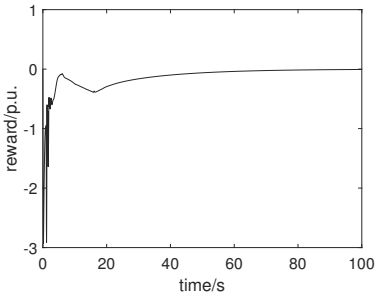


(b) Control action of each agent with payoff (12)

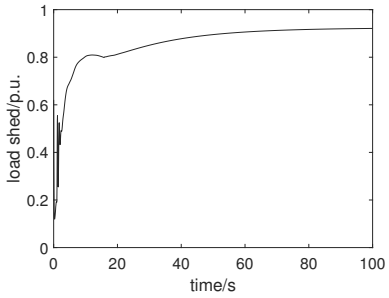


(c) Machine speed deviation of each agent with payoff (12)

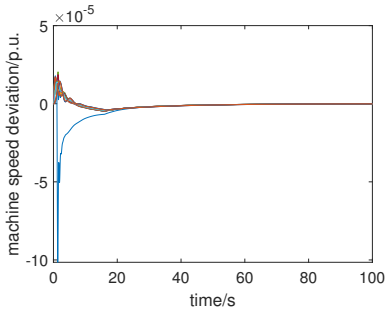
Fig. 12. Simulation results for IEEE 68-bus system using multi-Q-learning



(a) Dynamic payoff response

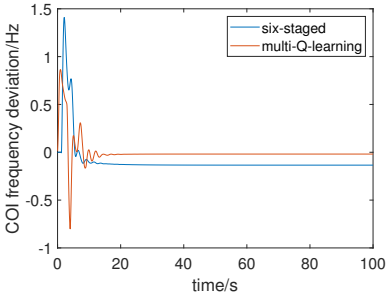


(b) Control action of each agent

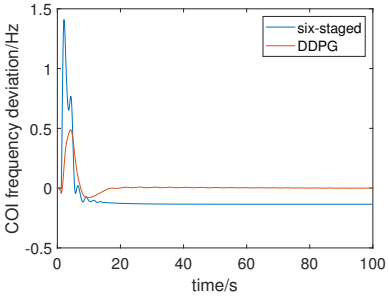


(c) Machine speed deviation of each generator

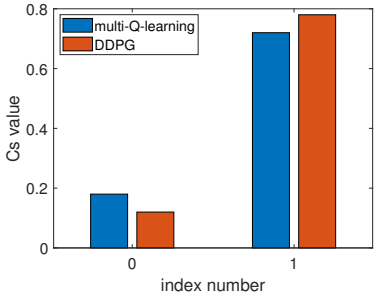
Fig. 13. Simulation results for IEEE 68-bus system using DDPG



(a) Comparison results for multi-Q-learning and staged schemes



(b) Comparison results for DDPG and staged schemes



(c) Bar chart of the $0-1$ of stability criteria C_s

Fig. 14. Comparison results and the robustness test using IEEE 68-bus system

[4] V. V. Terzija, "Adaptive underfrequency load shedding based on the magnitude of the disturbance estimation," *IEEE Transactions on Power Systems*, vol. 21, no. 3, pp. 1260–1266, 2006.

[5] M. Marzband, M. M. Moghaddam, M. F. Akorede, and G. Khomeyrani, "Adaptive load shedding scheme for frequency stability enhancement in microgrids," *Electric Power Systems Research*, vol. 140, pp. 78–86, 2016.

[6] S. S. Banijamali and T. Amraee, "Semi-adaptive setting of under frequency load shedding relays considering credible generation outage scenarios," *IEEE Transactions on Power Delivery*, vol. 34, no. 3, pp. 1098–1108, 2018.

[7] Y. Xu, W. Liu, and J. Gong, "Stable multi-agent-based load shedding algorithm for power systems," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 2006–2014, 2011.

[8] W. Gu, W. Liu, C. Shen, and Z. Wu, "Multi-stage underfrequency load shedding for islanded microgrid with equivalent inertia constant analysis," *International Journal of Electrical Power & Energy Systems*, vol. 46, pp. 36–39, 2013.

[9] Q. Zhou, Z. Li, Q. Wu, and M. Shahidehpour, "Two-stage load shedding for secondary control in hierarchical operation of islanded microgrids," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3103–3111, 2018.

[10] H. Liu, B. Wang, N. Wang, Q. Wu, Y. Yang, H. Wei, and C. Li, "Enabling strategies of electric vehicles for under frequency load shedding," *Applied Energy*, vol. 228, pp. 843–851, 2018.

[11] U. Rudez and R. Mihalic, "Wams-based underfrequency load shedding with short-term frequency prediction," *IEEE Transactions on Power Delivery*, vol. 31, no. 4, pp. 1912–1920, 2015.

[12] T. C. Njenda, M. Golshan, and H. H. Alhelou, "Wams based under frequency load shedding considering minimum frequency predicted and extrapolated disturbance magnitude," in *2018 Smart Grid Conference (SGC)*, pp. 1–5, IEEE, 2018.

[13] C. Rozyn, "National electricity amendment (emergency frequency control schemes) rule 2017." http://timmurphy.org/2009/07/22/line-spacing-in-latex-documents/. Accessed 30 March 2017.

[14] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 812–820, 2014.

[15] J. Yan, H. He, X. Zhong, and Y. Tang, "Q-learning-based vulnerability analysis of smart grid against sequential topology attacks," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 200–210, 2016.

[16] X. Han, H. Liu, F. Sun, and X. Zhang, "Active object detection with multi-step action prediction using deep q-network," *IEEE Transactions on Industrial Informatics*, 2019.

[17] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[18] K. Azizzadenesheli, E. Brunskill, and A. Anandkumar, "Efficient exploration through bayesian deep q-networks," in *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, IEEE, 2018.

[19] J. Xu, Z. Hou, W. Wang, B. Xu, K. Zhang, and K. Chen, "Feedback deep deterministic policy gradient with fuzzy reward for robotic multiple peg-in-hole assembly tasks," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1658–1667, 2018.

[20] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (ddpg) based energy harvesting wireless communications," *IEEE Internet of Things Journal*, 2019.

[21] J. Duan, H. Xu, and W. Liu, "Q-learning-based damping control of wide-area power systems under cyber uncertainties," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6408–6418, 2017.

[22] C. Wei, Z. Zhang, W. Qiao, and L. Qu, "Reinforcement-learning-based intelligent maximum power point tracking control for wind energy conversion systems," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 10, pp. 6360–6370, 2015.

[23] C. Wei, Z. Zhang, W. Qiao, and L. Qu, "An adaptive network-based reinforcement learning method for mppt control of pmsg wind energy conversion systems," *IEEE Transactions on Power Electronics*, vol. 31, no. 11, pp. 7837–7848, 2016.

[24] C. Chen, M. Cui, X. Wang, K. Zhang, and S. Yin, "An investigation of coordinated attack on load frequency control," *IEEE Access*, vol. 6, pp. 30414–30423, 2018.

[25] J. Zhang, M. Cui, and Y. He, "Robustness and adaptability analysis for equivalent model of doubly fed induction generator wind farm using measured data," *Applied Energy*, vol. 261, p. 114362, 2020.

[26] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," in *Innovations in multi-agent systems and applications-1*, pp. 183–221, Springer, 2010.

[27] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[28] C. Chen, K. Zhang, K. Yuan, L. Zhu, and M. Qian, "Novel detection scheme design considering cyber attacks on load frequency control," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 5, pp. 1932–1941, 2017.

[29] R. Yousefian and S. Kamalasadan, "A lyapunov function based optimal hybrid power system controller for improved transient stability," *Electric Power Systems Research*, vol. 137, pp. 6–15, 2016.

Chunyu Chen received the Ph.D. degree from Southeast University, Nanjing, China, in 2019. From 2017 to 2018, he was a Visiting Student with Southern Methodist University, Dallas, TX, USA. He is currently with School of Electrical and Power Engineering, China University of Mining and Technolgoy. His research interests include power system operation/control and power system cyber security.

Mingjian Cui received the B.S.and Ph.D. degrees in electrical engineering and automation from Wuhan University, Wuhan, China, in 2010 and 2015, respectively. From 2014 to 2015, he was a Visiting Scholar with the National Renewable Energy Laboratory, Transmission and Grid Integration Group, Golden, CO, USA. From 2016 to 2017, he was a Post-Doctoral Research Associate with The University of Texasat Dallas, Richardson, TX, USA. He is currently a Post-Doctoral Research Associate with Southern Methodist University, Dallas, TX, USA. He has published over 50 journal and conference papers. His research interests include power system operation, wind and solar forecasts, machine learning, data analytics, and statistics.

Fangxing Li received the B.S. and M.S. degrees in electrical engineering 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 of Electrical Engineering and the Campus Director of CURENT with the University of Tennessee, Knoxville, TN, USA. His research interests include renewable energy integration, distributed generation, energy markets, power system computing, reactive power and voltage stability, and measurement-based technology. Prof. Li is currently serving as the Vice Chair of IEEE PES PSOPE Committee, an Editor of the IEEE Transactions on Power Systems, an Editor of the IEEE Transactions on Sustainable Energy, an Editor of IEEE PES Letters.

Shengfei Yin received the B. Eng. degree from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2016, and the M.S. degree from the Illinois Institute of Technology, Chicago, IL, USA, in 2017. He is currently pursuing the Ph.D. degree in electrical engineering with Southern Methodist University, Dallas, TX, USA. His research interests include power market operation/optimization and data analysis in power systems.

Xinan Wang received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2013, and the M.S. degree from Arizona State University, Tempe, AZ, USA, in 2016, both in electrical engineering. He was a Research Assistant in the AI System Analytics Group at GEIRI North America, San Jose, CA, in 2016, 2017 and 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering at Southern Methodist University, Dallas, Texas, USA. His research interests include machine learning applications to power systems, wide-area measurement systems, data analysis and load modeling.