Journal of Operation and Automation in Power Engineering

Vol. 7, No. 1, May 2019, Pages: 107-118

http://joape.uma.ac.ir

Mini/Micro-Grid Adaptive Voltage and Frequency Stability Enhancement Using Q-learning Mechanism through the Offset of PID Controller

H. Shayeghi[*], *A. Younesi*

Department of Electrical Engineering, University of Mohaghegh Ardabili, Ardabil, Iran.

Abstract- *This paper develops an adaptive control method for controlling frequency and voltage of an islanded mini/micro grid (M/µG) using reinforcement learning method. Reinforcement learning (RL) is one of the branches of the machine learning, which is the main solution method of Markov decision process (MDPs). Among the several solution methods of RL, the Q-learning method is used for solving RL in this paper because it is a model-free strategy and has a simple structure. The proposed control mechanism is consisting of two main parts. The first part is a classical PID controller that is fixed tuned using Salp swarm algorithm (SSA). The second part is a Q-learning based control strategy which is consistent and updates it's characteristics according to the changes in the system continuously. Eventually, the dynamic performance of the proposed control method is evaluated in a real M/µG compared to fuzzy PID and classical PID controllers. The considered M/µG is a part of Denmark distribution system which is consist of three combined heat and power (CHP) and three WTGs. Simulation results indicate that the proposed control strategy has an excellent dynamic response compared to both intelligent and traditional controllers for damping the voltage and frequency oscillations.*

Keyword: Q-learning, Mini/Microgrid, Voltage control, Frequency oscillation damping, Fuzzy PID controller.

NOMENCLATURE

Abbreviations

AHC	Adaptive heuristic critic
AR	Average reward
CHP	Combined heat and power
DG	Distributed generation
GTG	Gas turbine generators
ISE	Integral of squared error
ITAE	Integral of time multiplied by absolute error
M/µG	Mini/Micro-grid
OS	Overshoot
PV	Photovoltaic
RL	Reinforcement Learning
SSA	Salp Swarm Algorithm
Ts	Settling time
US	Undershoot

1. INTRODUCTION

Developing the concept of M/µGs in the last decades, has been appeared as an effective response to different kind problems of classic power systems such as power system reliability, better power quality and reducing the environmental impacts [2]. M/µG is a small-scale power system with distinct electrical boundaries with the capability of supplying its loads autonomously when it is islanded from the main grid. [3]. One of the main characteristics of M/µG is that it is consists of different type power sources such DGs. Although this may be true, increasing the number of DGs will increase the M/µG availability when an error occurs on the main grid side in the islanded mode. But uncertain power output of renewable power resource like PV and WTG makes the control of voltage and frequency of M/µG a challenging work, which needs more effort and new adaptive control mechanisms [4, 5]. The literature on voltage and frequency control of M/µG shows a variety of approaches. As reported by Hirase et al. [6], the power system inertia is decreased due to increasing the number of DGs, therefore the frequency and voltage of power system are exposed to swing. In Ref. [7], a low voltage feedback controller is proposed, which in local level voltages is the second order, with the help of this method it is possible to use theoretical-circuit analysis techniques in closed loop systems. The authors in Ref. [8], have presented a control scheme based on the V-I drop and I-V drop characteristics for M/µG voltage and current conditions. The proposed method determines the output

Received: 08 Dec. 2018
Revised: 21 Mar. 2019
Accepted: 10 Apr. 2019
∗Corresponding author:
E-mail: hshayeghi@gmail.com (H. Shayeghi)
Digital object identifier: 10.22098/joape.2019.5542.1417
Research Paper

impedance of the resource subsystem along with the converter's dynamics, and analyzes the stability of the M/μG when is supplying constant loads. Firstly, the paper investigates the sustainability effects of key parameters such as loss coefficients, local loop control dynamics and the number of resources, and then compares the current and voltage status from a sustainability perspective. Asghar et al. [9], have developed a new control mechanism based on fuzzy logic and energy storage to control the frequency and voltage of the M/μG in the island mode. In this paper, battery storage and super-capacitor have been used to improve the M/μG frequency oscillations and voltage stability, respectively. Authors in Ref. [10], based on the output regulation theory, and fast-battery storage, have designed a controller to improve the frequency variations and the voltage stability of the M/μG. They attempt to improve the weaknesses of the drop-based controllers, including high settling time and poor transient performance. The effect of frequency and voltage oscillations on the operational performance of the M/μG is mathematically modelled in [6]. Finally, a proper control strategy based on the obtained mathematical model has been proposed to improve the frequency fluctuations and voltage deviations of the M/μG and has been tested experimentally. Various master-slave and drop based control methods for improving the frequency and voltage oscillations in an M/μG are presented and compared in Ref. [11]. In the master-slave based methods, the converter does not participate in the process of controlling the frequency and voltage, but in drop-based methods, it participates in the process of controlling the voltage and frequency. Although utilizing parallel converters in AC M/μGs and controlling them using drop based methods make the splitting of power between lines possible, but sometimes the difference of lines impedances in a sudden load change causes the instantaneous imbalance in the production and power absorbed by the parallel converter. Therefore, in Ref. [12], the authors have proposed a control method to improve the voltage stability of the M/μG by sampling the difference in lines impedances.

Reinforcement Learning is one of the important branches of machine learning in the field of artificial intelligence, and is a method for solving Markov decision process. The RL-based stability control of power system is explained in [13], in which the performance of this controller in power system oscillation enhancement was evaluated in a two-area four-machine power system. As reported in [13] , RL-based methods do not make any strong assumptions about the system dynamics. They can cope with partial information, nonlinear and stochastic behaviors. This characteristic is very useful for controlling power system, due to large size, huge information and a high degree of nonlinearities. When the real-world power system faces to a situation for the first time that was not experienced in simulations, this aspect is very important. The RL based controller continuously updates its knowledge about the system and therefore can adapt the changes in operating conditions or system dynamics. They can be used in combination with traditional control methods to improve performances. Yu and Zhen in [14] show two applications of RL method in power system stability control. In the first one, RL used for tuning the gain of a power system stabilizer and in the second application PSS was replaced with RL-based controller. They show that RL can be a complementary or a suitable alternative for conventional power system stabilizers. Reactive power control [15], power market applications [16], optimizing the dynamic performance of interline power flow controller [17] of RL-based methods was also reported. Literature shows the capability of RL methods to control major aspects of power systems.

Based on the approach presented in our earlier works [17, 18] the purpose of this paper is to design a supervisory PID controller for damping the voltage and frequency oscillations in an islanded M/μG with high penetration of WTGs. The Q-learning method, which is used for solving RL in this paper, is a model-free and a simple solution mechanism of RL. The proposed control mechanism is consisting of two main parts. The first part is a classical PID controller that is fixed tuned using SSA. The second part is a Q-learning based control strategy which is consistent and updates it's characteristics according to the changes in the system continuously. Simulations are carried out in both offline and online modes. In offline mode 1000 episodes are considered (In this paper) for learning the optimal policy by the RL-based proposed controller. Then the proposed supervisory controller is applied in online mode to optimally damp the voltage and frequency oscillations of the M/μG. It should be mentioned that the proposed RL control strategy has an adaptive behavior, so it updates it’s knowledge about the system during online simulation continuously. Eventually, to evaluate the dynamic performance of the proposed control method compared to intelligent and traditional controllers, a real islanded M/μG is considered and simulated using MATLAB/SIMULINK. The considered M/μG is a part of Denmark distribution system which is consist of three CHPs and three WTGs. Simulation results indicate that the proposed control strategy has an excellent dynamic response compared to both intelligent fuzzy logic based and traditional controllers for damping the voltage and frequency oscillations.

The main investigations of the present work are:

- To simulation of a nonlinear model of an islanded M/μG in order to simultaneous control of voltage and frequency.
- To utilizing multi-agent reinforcement learning criteria to enhancing the voltage and frequency control in an islanded M/μG.
- To proposing a supervisory control strategy, which can be applied throughout the offset of the industrial controllers in order to improve their overall dynamic response.

2. ADAPTIVE RL-PID CONTROLLER

2.1. Reinforcement learning

Reinforcement learning is an algorithmic method based on trial and error, in which one or more agents learn an optimal control policy by interact with their environment (system under control) [19]. In other words, the environment is divided into several discrete states, in each state, there are a definite number of actions to be implemented. The intelligent agent learns to determine the optimal action that has to be applied to the system in each state [17]. In general, there are several methods for solving RL problems like AHC, Q-learning, AR, and etc. [20]. In this paper, Q-learning is used to solve the proposed RL based frequency controller.

2.2. Q-learning

The main advantages of Q-learning based controllers are a simple structure, independent of the model of the system under control, robustness against changes in the operating point and system uncertainties and adaptive behaviour [17, 21]. Q-learning based reinforcement learning assumes the environment (system under control) is divided into a finite number of states is shown with set **S**. Agent forms a matrix called **Q**, which has a value (initially ‘0’) for each set of action-state pairs and indicates the goodness of particular action in the corresponding state. In each time step, agent calculates its state st, and based on a defined strategy selects action a among available actions of stat st. Immediately after applying the action, the agent takes a reward *r* from the environment and calculates its next state s_{t+1}. Then it updates the corresponding element of the **Q** matrix. The goal of the agent in Q-learning method is to learn a strategy which maps the states to actions to maximize discounted long-term reward [23]. The discounted long-term reward of the system is given by Eq. (1).

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \qquad (1)$$

where *r* is the reward, *γ* is a number at the range 0 to 1 and is called discount factor. **Q** matrix is defined as:

$$Q^{\pi}(s,a) = E_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s,\ a_t = a \right\} \qquad (2)$$

where, *π*, *s*, *a*, and *r* are the control policy, current state, selected action, and the received reward, respectively. In each time step, Eq. (2) should be updated using optimal Bellman equation, which is given by Eq. (3).

$$\Delta Q \ = \ \alpha\left[r_{t+1} + \gamma \max_{a} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \qquad (3)$$

where *α* is ϵ (0,1) and is called attenuation factor. The steps of the proposed Q-learning method is summarized in Fig. 1. It is evident from Fig. 1 that after completing the learning phase (offline simulation), the system will be switched to online simulation.

2.3. RL-PID

Figure 2 shows the block diagram of the RL-PID controller. As can be seen, the RL-PID controller consists of two parts. The first part is a traditional PID controller that its coefficients are optimized using SSA [24] in this paper. It must be noticed that this section is fixed and is adjusted only once. The second part, which is a compatible controller, has two stages. In the preprocessing section, the system state after the previous action is determined by using the received signal discretization. In the other part, the RL control mechanism, in a supervisory manner, corrects the output of the voltage and frequency PID controllers utilizing information obtained in preprocessing stage. This part is variable and updated at any time step. As its name implies, reinforcement learning, this controller after applying an action to the system, receives the impact of it in a reward/penalty form and gives it a score in the corresponding state. Certainly, in each state of the system, an action with a higher score, is best suited to be implemented to the system.

3. SALP SWARM ALGORITHM (SSA)

Salp swarm algorithm is a particle-based optimization method which mathematically models the movement of Salp particles toward the food location that is considered as the best solution. In order to cope with multi parameter optimization problems an *n*-dimensional space is suggested, where *n* is the number of variables to be optimized [23]. In each iteration, the particle with the best position (nearest to the food location) is selected as the leader of the other particles and its position is updated by Eq. (5).

```
A. Finding optimal control policy (offline simulation)
   I. Start
  II. Define set of states [1], actions [1], and reward
 III. Set values of $\alpha$, $\gamma$, and $\varepsilon$.
  IV. Set all $Q(s,a) = 0$.
   V. While episode < max_episode
      (or until convergence of $Q$ values)
        a. Calculate current stat
        b. Until achieving the goal repeat
             i. Select an action among available actions of current state using ε-greedy* method, ($\varepsilon$ at range
                0.3 to 0.6)
            ii. Apply selected action and take the reward from the environment, calculate next state.
           iii. Update $Q$ matrix using the following equation.
                    $Q = Q + \Delta Q$        (4)
            iv. Set current action equal to next action.
        c. Go to A. b
  VI. Go to A. V.
 VII. End of the learning process.
B. Run the optimal policy (online simulation)
   I. Start
  II. Calculate current state.
 III. repeat
        a. Select an action among available actions of current state using the ε-greedy method, (ε at range 0.01
           to 0.03).
        b. Apply selected action and take the reward from the environment, calculate next state.
        c. Update $Q$ matrix using Eq. (4).
        d. Set current action equal to next action.
  IV. Go to B. III.
```

* In this method, the agent selects the action with the maximum value of Q with the probability of 1-ε, and selects an action randomly with the probability of ε.

Fig. 1. Steps of the Q-learning solution method for RL

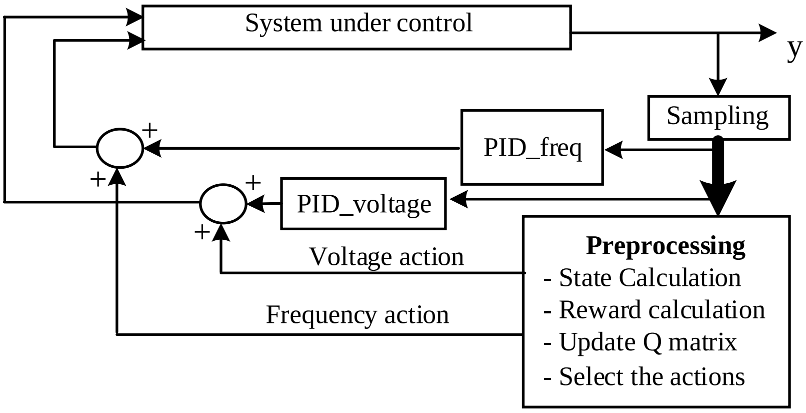

Fig. 2. Descriptive model of the proposed adaptive RL-PID controller

$$X_j^1 = \begin{cases} F_j + c_1((ub_j - lb_j)c_2 + lb_j) & c_3 \geq 0 \\ F_j - c_1((ub_j - lb_j)c_2 + lb_j) & c_3 < 0 \end{cases} \qquad (5)$$

Where X_j^1 is the first leader, F_j is food location, ub_j and lb_j are the upper and lower bonds of the parameters all in j^{st} dimension. c_1, c_2, and c_3 are random numbers. Parameter c_1 which is important because it balances the exploration and exploitation phases, is calculated by Eq. (6).

$$c_1 = 2e^{-(\frac{4l}{L})^2} \qquad (6)$$

Where l is the current iteration and L is the maximum number of iterations. c_2 and c_3 parameters are determined using normal distribution in the range [0 1].

According to Newton's displacement law, the position of the imitative Salp individuals is derived by Eq. (7).

$$X_j^i = \frac{1}{2}at^2 + v_0 t \tag{7}$$

Where i>2, X^i_j is the position of i^{th} Salp particle in j^{th} dimension, t refers to time, v_0 is the initial speed, $a = v_{final} / v_0$ which $v = (x - x_0) \times t^{-1}$. In the concept of optimization, time is equivalent with iteration and the space between the iterations is 1. With this in mind and proposing 0 for the initial speed of all individuals, Eq. (7) can be rewritten as Eq. (8).

$$X_j^i = \frac{1}{2}(X_j^i + X_j^{i-1}) \tag{8}$$

More detail about the SSA can be found in [23]. Figure 3 shows the pseudo code of the SSA.

```
Start
Initial population creation.
Repeat until the stop criteria is met
    Calculate the fitness function for all Salp individuals.
    Set the food location as the location of the best particle.
    Determine $c_1$ to $c_3$.
    Do for all Salp individuals
        Update the position of Leader individuals Salp using Eq. (5).
        Update the position of non-leader Salp individuals using Eq. (8).
        Check the constraint violations.
    End Do
End Repeat
Export the food location as the solution of the problem.
```

Fig. 3 The pseudo code of SSA.

4. MODEL OF THE M/µG UNDER STUDY

A typical European distribution power system owned by Himmerlands Elforsyning located in Aalborg, Denmark with high penetration of DG units is considered in this chapter. It consists of 10 loads, 3 fixed-speed stall-regulated WTGs and a CHP plant with three GTG namely CHP1, CHP2, and CHP3. It is assumed the WTG units have the capacitor banks for necessary compensation and they operate close to unity power factor. A test case situation is used to assess the dynamic performance of the proposed control method compared to classical PID controller to damp the frequency and voltage fluctuations with productions of 2.5 MW, 2.8 MW, and 2.8 MW from CHP1, CHP2, and CHP3 respectively, and 0.08 MW from each WTGs. Figure 4 shows the single line diagram of the proposed M/µG and locations of load and power plants [25]. The Simulink model of Figs. 5 and 6 are used for the induction and synchronous generators, respectively. Load, WTG, and exciter system data are given in Tables 1 to 3 [24]. In order to simulate the proposed M/µG in Simulink, the reduced Y_{bus} method is used. In this way, firstly, the Y_{bus} of the power system will be calculated. Then by eliminating the load buses, the reduced Y_{bus} will be formed. The voltage of the generation units multiplied by the reduced Y_{bus} and the input current of generation units will be determined. Figure 7 shows this process.

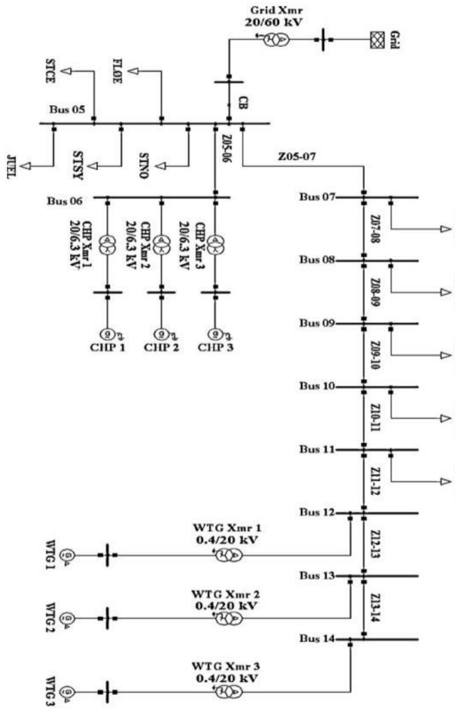

Fig. 4 Single line diagram of the proposed M/µG [25].

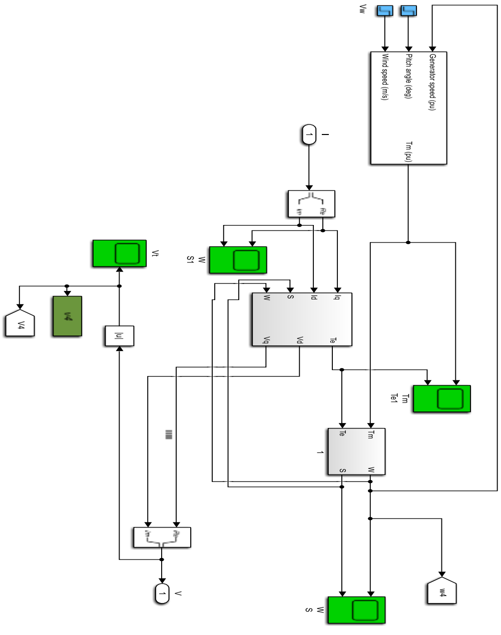

Fig. 5 Simulink model of the induction generator (WTG1 to WTG3 in Fig. 7)

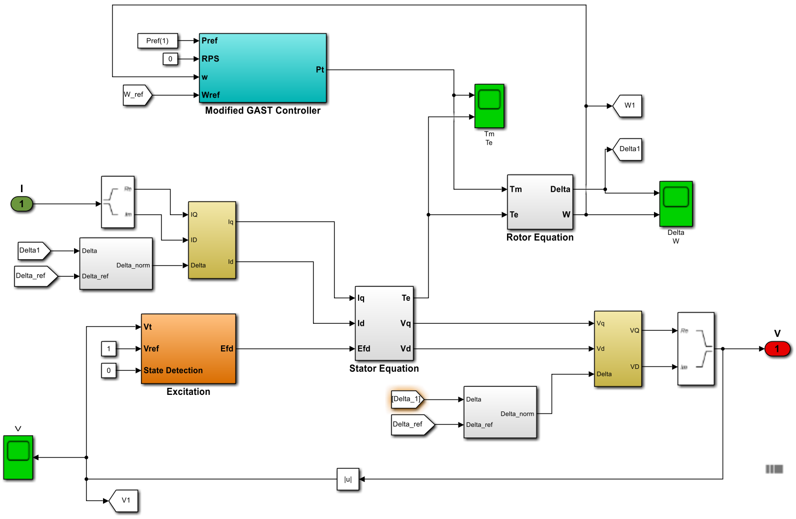

Fig. 6 Simulink model of the synchronous generator (CHP1 to CHP3 in Fig. 7)

Table 1. Load data

Load name	Active power (MW)	Reactive power (MVar)
FLØE	0.787	0.265
JUEL	2.442	0.789
STCE	1.212	0.16
STNO	2.109	0.286
STSY	0.4055	0.0735
Load 07	0.4523	0.2003
Load 08	0.7124	0.3155
Load 09	0.1131	0.0501
Load 10	0.1131	0.0501
Load 11	0.1131	0.0501

5. OPTIMIZATION RESULTS

5.1. Fuzzy PID and classical PID controllers

In this paper, some simplistic assumptions are considered due to complexity and economic aspects of designing multiple controllers in a complex power system. 1) Designing controller for WTG units is neglected because the capacity of WTGs is very small compared to CHP units. 2) The same voltage and frequency controllers are designed for all units, because they are completely identical. According to the above assumptions, two independent controllers are designed for the M/µG. An identical frequency controller and an identical voltage controller for all the CHP units. As Eq. (9) describes a time-domain objective function is considering for achieving the best dynamic response of the FPID [25] and classic PID controllers, which is based on the integral of time multiplied by absolute error criteria.

Table 2. Excitation system data

Parameter		Value
T_r	: Measurement delay (s)	0.0
K_a	: AVR dc gain	500
T_a	: AVR time constant (s)	0.02
K_e	: Exciter constant (p.u.)	1.0
T_e	: Exciter time constant (s)	0.9
K_f	: Stabilization path gain (p.u.)	0.03
Tf_1	: 1st stabilization path time constant (s)	0.6
Tf_2	: 2nd stabilization path time constant (s)	0.38
Tf_3	: 3rd stabilization path time constant (s)	0.058
E_1	: Saturation factor 1 (p.u.)	5.6
Se_1	: Saturation factor 2 (p.u.)	0.86
E_2	: Saturation factor 3 (p.u.)	4.2
Se_2	: Saturation factor 4 (p.u.)	0.5
V_{min}	: Controller minimum out (p.u.)	-7.3
V_{max}	: Controller maximum output (p.u.)	7.3

Table 3. WTG system data

Parameter	Value
Rotor inertia (kg.mm)	4×10^6
Drive train stiffness (N.m/rad)	8000
Drive train damping (N.m/rad)	0
Rotor (m)	34

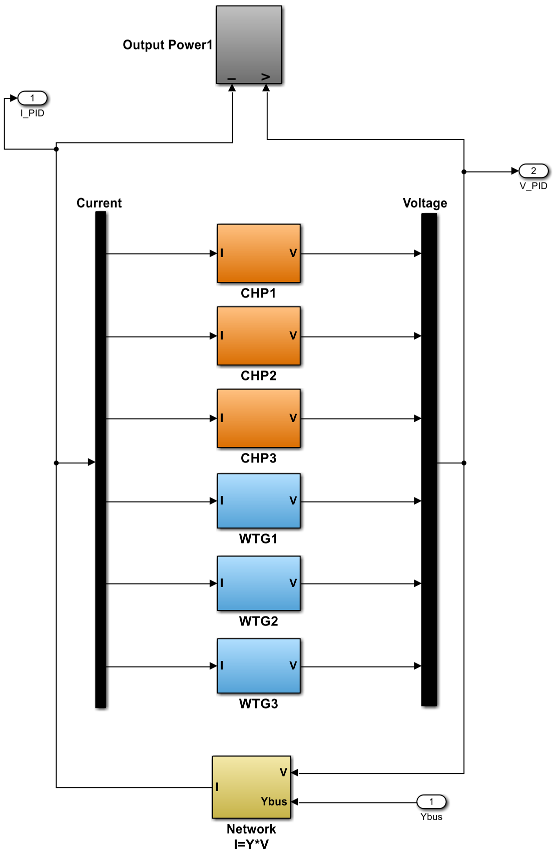

Fig. 7 Simulink model of the simulation process

Both frequency and voltage deviations in CHP buses are used for calculation of the objective function.

$$J = \int_{0}^{SimTime} t \times [(\Delta\omega_{CHP1} + \Delta\omega_{CHP2} + \Delta\omega_{CHP3})... \\ + \beta \times (\Delta V_{CHP1} + \Delta V_{CHP2} + \Delta V_{CHP3})]dt \qquad (9)$$

In Eq. (9), β is a weight factor that places the values of the voltage and frequency errors in the same range. Here, β is considered equal to 3 and *SimTime* considered as 40 s. The optimization problems are formed based on the control parameters of the FPID and PID controllers and are described by Eqs. (10.a) and (10.b) for FPID and PID controllers, respectively. In order to prevent an increase in paper volume, the structure, membership functions, and fuzzy rules of FPID controller are ignored here. Readers can refer to [25] for more details of FPID.

$$\begin{aligned} & Minimize \quad J \\ & subject\ to: \\ & K_1^{\min} < K_1^{f\&v} < K_1^{\max} \\ & K_2^{\min} < K_2^{f\&v} < K_2^{\max} \\ & K_3^{\min} < K_3^{f\&v} < K_3^{\max} \\ & K_4^{\min} < K_4^{f\&v} < K_4^{\max} \end{aligned} \qquad (10.a)$$

$$\begin{aligned} & Minimize \quad J \\ & subject\ to: \\ & K_p^{\min} < K_p^{f\&v} < K_p^{\max} \\ & K_i^{\min} < K_i^{f\&v} < K_i^{\max} \\ & K_d^{\min} < K_d^{f\&v} < K_d^{\max} \end{aligned} \qquad (10.b)$$

SSA is used for solving the optimization problems of Eqs. (10.a) and (10.b) with 50 initial populations and 50 iterations. The optimization result and the convergence process of the objective function for both FPID and PID controllers are shown in Table 4 and Fig. 8, respectively.

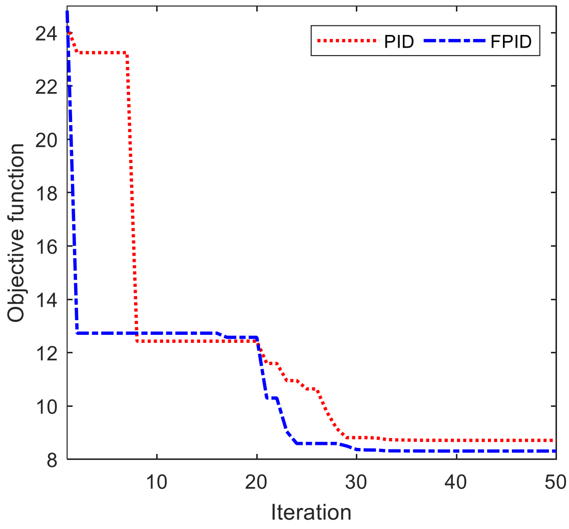

Fig. 8. The convergence curve of the objective function

5.2. The RL controller: state, action and reward/penalty function definition

Important to realize, the proposed RL controller is portable and can be added as a supervisory controller to any other kinds of controllers to improve their dynamic performance. In this paper, the PID controller is chosen, because along with its acceptable performance, has a simple structure and is widely used in the industry. Formerly, it was stated that the Q-learning method is used to solve the reinforcement learning in this paper. Another key point is that the Q-learning based controller's performance depends largely on how the states, actions, and reward/penalty functions are defined, which are described in more detail below.

5.2.1. States

Since the oscillation enhancement is the primary objective of this paper, the $\Delta\omega$ signal is sampled and used as the feedback signal from the system under control to determine the system state.

Table 4. SSA based control parameters of FPID & PID controllers

FPID

Parameter	K_1^f	K_2^f	K_3^f	K_4^f	K_1^v	K_2^v	K_3^v	K_4^v
Value	0.94512	2.00125	1.45218	0.751425	0.098452	1.625101	1.302158	0.0099541

PID

Parameter	K_p^f	K_i^f	K_d^f	K_p^v	K_i^v	K_d^v
Value	0.54821	0.18215	0.0001421	0.0824	0.0299	0.00154

For this aim, the time interval from -0.02 to + 0.02 divided into 50 equal segments and Eq. (11) is utilized at each time step to determine the state of the system. The zero-centred state is called normal state, and the intelligent agent does not do anything in the normal state. In fact, this equation, in addition to determining the value of system oscillations, it makes it clear whether the oscillations are going to the instability or moving towards the establishment [17].

$$s_t = (\Delta\omega, \frac{d\Delta\omega}{dt}) \tag{11}$$

where s_t is the state of the micro-grid at time t and is a function of $\Delta\omega$ and its derivative.

5.2.2. Actions

Although, there are no particular laws for defining of actions for RL based controllers, and this makes it a complex matter [26]. But it may be determined by inspiring from the output limits of the usual controllers that used for the same purpose [18]. Although increasing the number of actions for each system state can improve the dynamic performance of the controller by increasing the degree of freedom but on the other hand it increases the learning time extremely and makes it challenging (or even impossible) to find the optimal control policy. With this in mind, in this paper, the same actions are suggested for all states and expressed by Eq. (12). As a result of the simulation, the magnitude of voltage oscillations is smaller than the frequency oscillations. Also, according to Table 4, the control gains of the frequency PID controller are obtained approximately ten times greater than the gains of the voltage PID controller using SSA. With these in mind, the actions of the voltage controller are chosen smaller than the frequency actions.

$$A = \left\{ \begin{matrix} F \rightarrow & -0.05, \ \text{-0.025}, \ 0, \ 0.025, \ 0.05 \\ V \rightarrow & -0.005, \ \text{-0.0025}, \ 0, \ 0.0025, \ 0.005 \end{matrix} \right\} \tag{12}$$

where A is the action set for all states of the system.

5.2.3. Reward/Penalty function

The reward/penalty function is important because it assesses the degree of satisfaction from the action taken in the previous state in line with the overall goal. In the event that the system state is s_t, the agent utilizes its experience to perform the best action (a) among the actions defined for state s_t. Immediately; the agent receives a reward/penalty from the system under control concerning the performed action. Based on this reward/penalty, the agent assigns a score for a pair of (s_t, a) and updates the corresponding element of $\boldsymbol{Q}$ matrix. If the score is positive, the probability of performing the action a at the state s_t increases for the next times. Otherwise, if the score is negative (penalty), the agent selects the action a with a lower probability in the state s_t, in next times. With this intention that the primary objective of this paper is voltage and frequency control, therefore $\Delta\omega$ and Δv signal of all CHP generation units are selected for determination of reward/penalty for corresponding (s_t, a) pairs.

$$\Re_t = \begin{cases} +1 & \text{If } s_{t+1} \text{ is normal state} \\ -1 & \text{If } s_t \text{ is normal state and } s_{t+1} \text{ is not normal state} \\ \left(\dfrac{1}{(1 + \sum\limits_{k=t-1}^{t} (\Delta\omega_1(k) + \Delta\omega_2(k) + \Delta\omega_3(k) + \varsigma \times (\Delta V_1(k) + \Delta V_2(k) + \Delta V_3(k))))} \right) & \text{Otherwise} \end{cases} \tag{13}$$

In essence, if an action causes the system to go out of the normal state, it will be fined. In return, if an action causes the system to go to the normal state, will receive the highest reward. In summary, the reward/penalty function is described by Eq. (13), in this paper. where *t* is the time step.

6. SIMULATION RESULTS AND DISCUSSION

In order to evaluation of the dynamic performance of the proposed RL-PID controller compared to classical PID controller, M/μG of Fig. 4 is simulated with MATLAB/Simulink. A Three phase fault at bus 7 is considered as a challenging condition of the system. The angular velocity of the generators is plotted in Figs. 9 and 10. According to Figs. 9 and 10, both control strategies have the ability to eliminate the oscillations of angular velocity (frequency) of the generation units. However, it's clear that the RL-PID controller has a tremendous dynamic performance compared to classical PID controller. RL-PID controller has an excellent dynamic performance thanks to its flexible structure, which integrates the adaptive property of RL with the speed and precision of the PID controller.

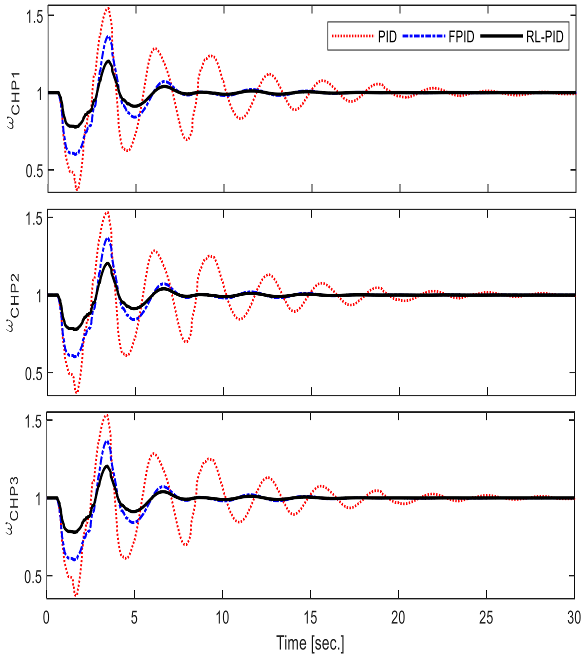

Fig. 9 The angular velocity of CHP units

In order to more highlight the capabilities of the RL-PID controller, the oscillations of the voltage of the generation buses are shown in Figs. 11 and 12. Under the conditions of the same fault, the amplitude of the voltage fluctuations is much smaller than the frequency fluctuations. In these circumstances, the superiority of the RL-PID controller is inferred from simulation results compared to the other control method. Results show that the RL-PID makes the system dynamics better in terms of overshoot /undershoot, increasing the speed of the oscillation damping, and the complete removal of the steady-state error.

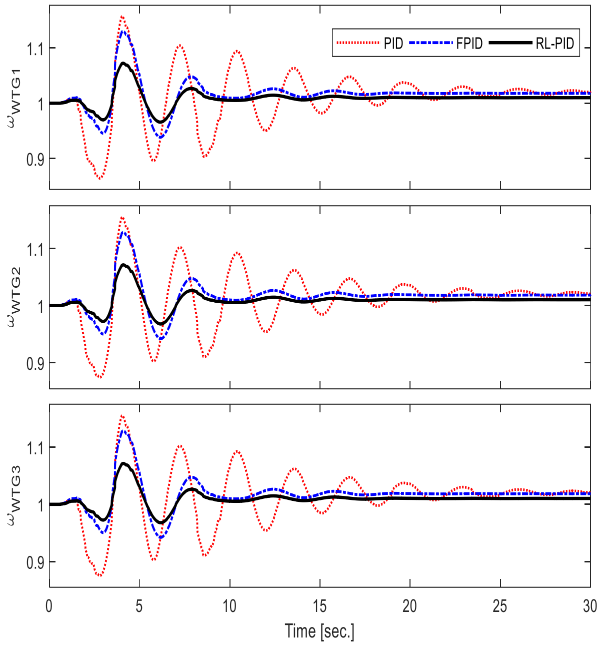

Fig. 10 The angular velocity of WTG units

Total production of active and reactive powers of M/μG during fault occurrence is shown in Fig. 13. The active power remains constant after some oscillations. Wherever it is evident that the RL-PID controller is superb in damping the oscillations compared to traditional PID controller.

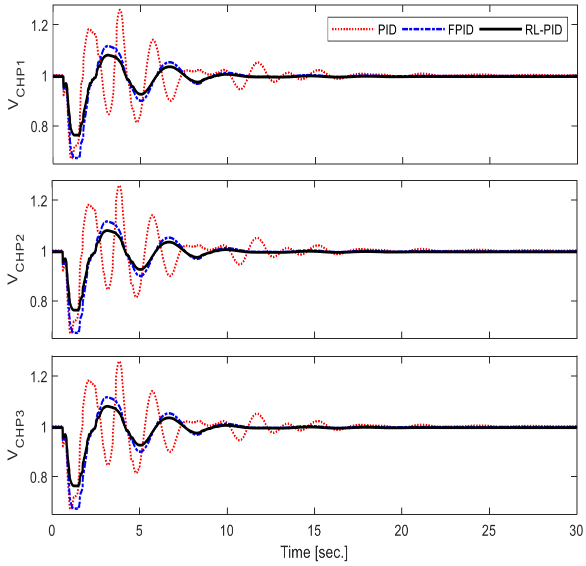

Fig. 11 The variations of voltage at the CHP unit terminals

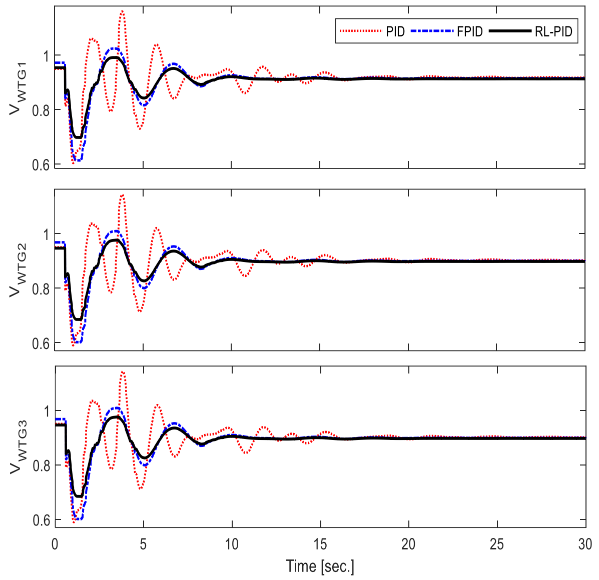

Fig. 12 The variations of voltage at the WTG unit terminals

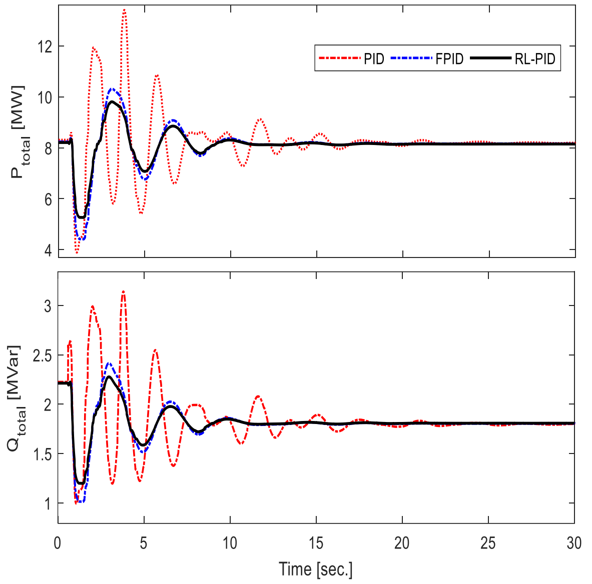

Fig. 13 Total production of the active and reactive power of M/μG

Figure 14 shows that how the M/μG power factor changes during the fault. Since the reactive power is decreased and active power remains constant, the power factor is increased. It can be revealed through Fig. 12 that the dynamic performance of the RL-PID controllers is much better than the classical PID controller. Figure 15 shows the actions were taken by the frequency and voltage controllers. It can be seen from Fig. 15 that, the RL controller is inactive in the normal state of the system but when a disturbance occurs in the system the frequency and voltage of the system become unstable, RL starts applying the suitable control action. When oscillations are well damped the RL controller become inactive again.

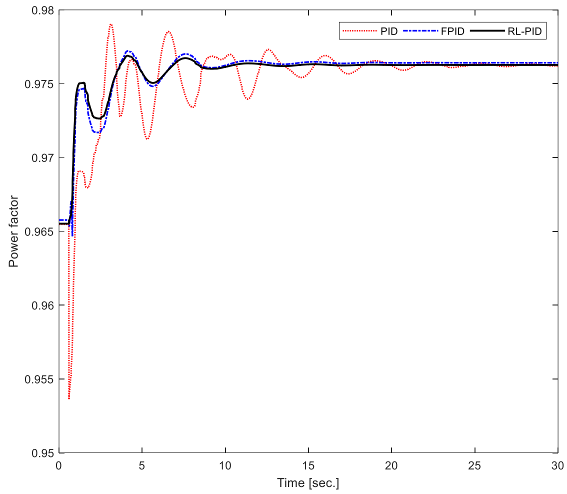

Fig. 14 Variations of M/μG power factor

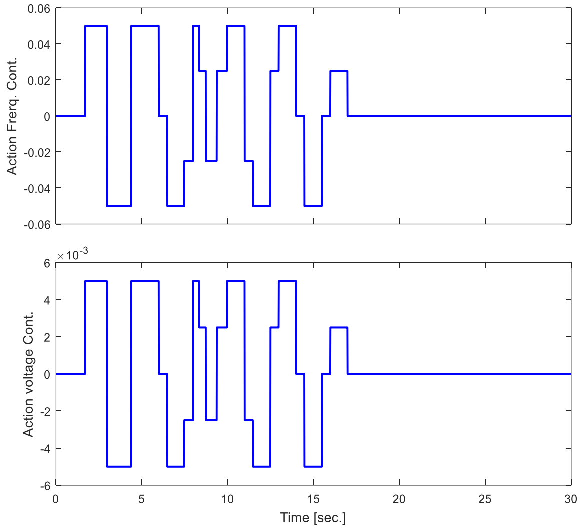

Fig. 15 Actions selected by the frequency and voltage controllers

In order to more identify the superiority of the proposed RL-based control strategy compared to traditional PID controller suitable time domain performance indices are selected and calculated for both RL and PID controllers. In this way, five suitable time domain indices based on *OS*, *US*, T_s, ITAE, and ISE are calculated and shown in Tables 5. Equations (14)-(17) calculate the proposed OS, US, ITAE, and ISE indices.

$$OS_i = \frac{\max(y_i) - y_i^{final}}{y_i^{final}} \times 100 \qquad (14)$$

$$US_i = \frac{y_i^{final} - \min(y_i)}{y_i^{final}} \times 100 \qquad (15)$$

Table 5. Time domain performance indices for RL-PID, FPID, and classical PID controllers

				Frequency			
Signal		$\Delta\omega_{CHP1}$	$\Delta\omega_{CHP2}$	$\Delta\omega_{CHP3}$	$\Delta\omega_{WTG1}$	$\Delta\omega_{WTG2}$	$\Delta\omega_{WTG3}$
OS%	PID	54.6116	53.1917	53.1916	15.7141	15.4995	15.4968
	FPID	36.6468	36.7063	36.7063	13.0022	12.8450	12.8492
	RLPID	**20.3593**	**20.3593**	**20.3593**	**7.2234**	**7.1361**	**7.1384**
US%	PID	62.7297	62.6482	62.6482	13.5939	12.5969	12.5010
	FPID	40.0135	39.8168	39.8168	6.1533	5.8015	5.7845
	RLPID	**22.2297**	**22.1204**	**22.1204**	**3.4185**	**3.2230**	**3.2136**
Ts	PID	24.642	24.832	24.832	19.620	19.630	19.630
	FPID	14.31	14.30	14.30	7.8965	7.910	7.910
	RLPID	**11.210**	**11.200**	**11.200**	**7.668**	**7.688**	**7.688**
ITAE	PID	22.6570	22.3870	22.3870	19.6876	19.8461	19.8593
	FPID	4.4438	4.4438	4.4438	15.2919	15.6749	15.7027
	RLPID	**2.4687**	**2.4632**	**2.4632**	**8.4955**	**8.7082**	**8.7237**
ISE	PID	87.9231	87.5949	87.5949	8.4800	7.9667	7.9317
	FPID	29.5241	29.3438	29.3438	3.3208	3.3093	3.3129
	RLPID	**9.1124**	**9.0567**	**9.0567**	**1.0249**	**1.0213**	**1.0225**
				Voltage			
Signal		ΔV_{CHP1}	ΔV_{CHP2}	ΔV_{CHP3}	ΔV_{WTG1}	ΔV_{WTG2}	ΔV_{WTG3}
OS%	PID	25.8371	25.7998	25.7998	16.1028	14.2935	14.2410
	FPID	11.5505	11.5488	11.5488	2.8549	5.6842	5.8452
	RLPID	**7.9342**	**7.9328**	**7.9328**	**0.8515**	**2.4063**	**2.4771**
US%	PID	33.0525	62.6482	62.6482	13.5939	12.5969	12.5010
	FPID	32.7502	39.8168	39.8168	6.1533	5.8015	5.7845
	RLPID	**23.7091**	**22.1204**	**22.1204**	**3.4185**	**3.2230**	**3.2136**
Ts	PID	33.590	33.620	33.620	39.497	39.490	39.490
	FPID	14.35	14.35	14.35	17.510	14.720	14.720
	RLPID	**9.950**	**9.950**	**9.950**	**10.038**	**10.028**	**10.018**
ITAE	PID	5.6965	5.6971	5.6971	7.5216	8.23404	8.31862
	FPID	5.4520	5.4520	5.4520	6.8954	6.8954	6.8954
	RLPID	**5.2710**	**5.2682**	**5.2682**	**6.68453**	**7.90976**	**7.99341**
ISE	PID	14.6909	14.6871	14.6871	41.8811	53.2423	54.1084
	FPID	10.9456	10.9539	10.9539	40.36195	52.3804	53.29198
	RLPID	**5.7889**	**5.7933**	**5.7933**	**37.8660**	**49.9451**	**50.8502**

$$ITAE_i = \int_0^{40} t \times |y_i|\, dt \tag{16}$$

$$ISE_i = \int_0^{40} (y_i)^2\, dt \tag{17}$$

Where y_i indicates the frequency and bus voltage of the CHP and WTG units.

As can be seen from Table 5 the proposed control strategy improved the OS and US more than 50% compared to classical PID controller. Settling time, the other time domain performance index which is important because it shows the capability of the proposed method in full damping of the oscillations has been enhanced between 54% to 74% with RL-PID controller compared to classical PID controller. The proposed controller has been reduced the ITAE and ISE perspective of different kind signals approximately 7% to 90% in some cases. As a result, it can be proved that the proposed RL-based control mechanism thanks to its flexible structure has a superb dynamic performance compared to the PID controller.

7. CONCLUSION

The purpose of this paper is to offer a two-stage control strategy for enhancing the stability of voltage and frequency an islanded M/µG with high penetration of WTGs. The first stage is a classical PID controller, which is fixed tuned using SSA. The second part is a Q-learning based control strategy which is consistent and updates it's characteristics according to the changes in the system continuously. The proposed RL control strategy has an adaptive behavior, so it updates it’s knowledge about the system during online simulation continuously. In order to evaluate the dynamic performance of the proposed control method compared to a traditional PID controller, a real islanded M/µG is considered and simulated using MATLAB/SIMULINK. Simulation results indicate that the proposed control strategy has an excellent dynamic response compared to the traditional PID controller for damping the voltage and frequency oscillations. Finally, suitable time domain performance indices such as overshoot, undershoot, settling time, ITAE and ISE are calculated and compared for both RL-PID and classic PID controller. The results indicate that the proposed RL based strategy can cope with system nonlinearities effectively and damp the oscillations of voltage and frequency in a M/µG. It is model-free and can control the system without any strong assumptions.

REFERENCES

[1] "IEEE recommended practice for excitation system models for power system stability studies," 1992.

[2] M. J. Morshed and A. Fekih, "A fault-tolerant control paradigm for microgrid-connected wind energy systems," *IEEE Syst. J.*, vol. 12, pp. 360-372, 2018.

[3] R. Ghanizadeh, M. Ebadian, and G. B. Gharehpetian, "Control of inverter-interfaced distributed generation units for voltage and current harmonics compensation in grid-connected microgrids," *J. Oper. Autom. Power Eng.*, vol. 4, pp. 66-82, 2016.

[4] D. O. Amoateng, M. A. Hosani, M. S. Elmoursi, K. Turitsyn, and J. L. Kirtley, "Adaptive voltage and frequency control of islanded multi-microgrids," *IEEE Trans. Power Syst.*, vol. 33, pp. 4454-4465, 2018.

[5] X. Wu, C. Shen, and R. Iravani, "A distributed, cooperative frequency and voltage control for microgrids," *IEEE Trans. Smart Grid*, vol. 9, pp. 2764-2776, 2018.

[6] Y. Hirase, K. Abe, K. Sugimoto, K. Sakimoto, H. Bevrani, and T. Ise, "A novel control approach for virtual synchronous generators to suppress frequency and voltage fluctuations in microgrids," *Appl. Energy*, vol. 210, pp. 699-710, 2018/01/15/ 2018.

[7] J. W. Simpson-Porco, F. Dörfler, and F. Bullo, "Voltage stabilization in microgrids via quadratic droop control," *IEEE Trans. Autom. Cont.*, vol. 62, pp. 1239-1253, 2017.

[8] F. Gao, S. Bozhko, A. Costabeber, C. Patel, P. Wheeler, C. I. Hill, *et al.*, "Comparative stability analysis of droop control approaches in voltage-source-converter-based DC microgrids," *IEEE Trans. Power Electron.*, vol. 32, pp. 2395-2415, 2017.

[9] F. Asghar, M. Talha, and S. Kim, "Robust frequency and voltage stability control strategy for standalone AC/DC hybrid microgrid," *Energies*, vol. 10, p. 760, 2017.

[10] H. Zhao, M. Hong, W. Lin, and K. A. Loparo, "Voltage and frequency regulation of microgrid with battery energy storage systems," *IEEE Trans. Smart Grid*, vol. PP, pp. 1-1, 2017.

[11] A. Ahmarinejad, B. Falahjoo, and M. Babaei, "The stability control of micro-grid after islanding caused by error," *Energy Procedia*, vol. 141, pp. 587-593, 12// 2017.

[12] W. Issa, S. M. Sharkh, R. Albadwawi, M. Abusara, and T. K. Mallick, "DC link voltage control during sudden load changes in AC microgrids," Proc. *IEEE 26th Int. Symp. Ind. Electron.*, 2017, pp. 76-81.

[13] D. Ernst, M. Glavic, and L. Wehenkel, "Power systems stability control: reinforcement learning framework," *IEEE Trans. Power Syst.*, vol. 19, pp. 427-435, 2004.

[14] T. Yu and W. G. Zhen, "A reinforcement learning approach to power system stabilizer," Proc. *IEEE Power & Energy Soc. Gen. Meet.*, Calgary, AB, 2009, pp. 1-5.

[15] J. G. Vlachogiannis and N. D. Hatziargyriou, "Reinforcement learning for reactive power control," *IEEE Trans. Power Syst.*, vol. 19, pp. 1317-1325, 2004.

[16] V. Nanduri and T. K. Das, "A reinforcement learning model to assess market power under auction-based energy pricing," *IEEE Trans. Power Syst.*, vol. 22, pp. 85-95, 2007.

[17] A. Younesi, H. Shayeghi, and M. Moradzadeh, "Application of reinforcement learning for generating optimal control signal to the IPFC for damping of low-frequency oscillations," *Int. Trans. Electr. Energy Syst.*, vol. 28, p. e2488, 2018.

[18] H. Shayeghi and A. Younesi, "An online q-learning based multi-agent LFC for a multi-area multi-source power system including distributed energy resources," *Iran. J. Electr. Electron. Eng.*, vol. 13, pp. 385-398, 2017.

[19] C. Weber, M. Elshaw, and N. M. Mayer, *Reinforcement learning, theory and applications*: I-TECH Education and Publishing, 2008.

[20] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: a survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237-285, 1996.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*: MIT Press, 2005.

[22] C. J. C. H. Watkins and P. Dayan, "Technical note: q-learning," *Mach. Learn.*, vol. 8, pp. 279-292.

[23] S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, "Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems," *Adv. Eng. Software*, vol. 114, pp. 163-191, 2017/12/01/ 2017.

[24] P. Mahat, Z. Chen, and B. Bak-Jensen, "Control and operation of distributed generation in distribution systems," *Electric Power Syst. Res.*, vol. 81, pp. 495-502, 2011/02/01/ 2011.

[25] H. Shayeghi, A. Younesi, and Y. Hashemi, "Optimal design of a robust discrete parallel FP + FI + FD controller for the Automatic Voltage Regulator system," *Int. J. Electr. Power Energy Syst.*, vol. 67, pp. 66-75, 2015.

[26] R. Hadidi and B. Jeyasurya, "Reinforcement learning based real-time wide-area stabilizing control agents to enhance power system stability," *IEEE Trans. Smart Grid*, vol. 4, pp. 489-497, 2013.