energies

MDPI

Article

# Assessing the Use of Reinforcement Learning for Integrated Voltage/Frequency Control in AC Microgrids

**Abdollah Younesi [1], Hossein Shayeghi [1,*] and Pierluigi Siano [2,*]**

[1] Electrical Engineering Department, Faculty of Engineering, University of Mohaghegh Ardabili, Ardabil 56199-11367, Iran; younesi.abdollah@gmail.com

[2] Department of Innovation and Management Systems, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy

* Correspondence: hshayeghi@gmail.com (H.S.); psiano@unisa.it (P.S.); Tel.: +98-4533512910 (H.S.); +39-089-96-4294 (P.S.)

Received: 4 February 2020; Accepted: 4 March 2020; Published: 8 March 2020


**Abstract:** The main purpose of this paper is to present a novel algorithmic reinforcement learning (RL) method for damping the voltage and frequency oscillations in a micro-grid (MG) with penetration of wind turbine generators (WTG). First, the continuous-time environment of the system is discretized to a definite number of states to form the Markov decision process (MDP). To solve the modeled discrete RL-based problem, Q-learning method, which is a model-free and simple iterative solution mechanism is used. Therefore, the presented control strategy is adaptive and it is suitable for the realistic power systems with high nonlinearities. The proposed adaptive RL controller has a supervisory nature that can improve the performance of any kind of controllers by adding an offset signal to the output control signal of them. Here, a part of Denmark distribution system is considered and the dynamic performance of the suggested control mechanism is evaluated and compared with fuzzy-proportional integral derivative (PID) and classical PID controllers. Simulations are carried out in two realistic and challenging scenarios considering system parameters changing. Results indicate that the proposed control strategy has an excellent dynamic response compared to fuzzy-PID and traditional PID controllers for damping the voltage and frequency oscillations.

**Keywords:** machine learning; microgrid control; Markov decision process; reinforcement learning

## 1. Introduction

In the last few years, there has been a growing interest in the development of microgrids (MGs) for enhancing power system reliability, better power quality and reducing the environmental impacts [1,2]. An MG 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 MG is that it consists of different types of power sources such as distributed generations (DGs) and energy storage systems [4,5]. While high number of DGs can increase the MG availability when an error occurs on the main grid side in the islanded mode, the uncertain power output of renewable power resource like PV and wind turbine generator (WTG) makes the control of voltage and frequency of MG a challenging work, which needs more effort and new adaptive control mechanisms [6,7]. Without a strong and high-efficiency control method, in this condition, the frequency and voltage of the microgrid may reach undesirable values [8].

The literature on voltage and frequency control of MG shows a variety of approaches [9–11]. As reported by Hirase et al. [12], the power system inertia decreases due to the higher number of DGs, therefore the frequency and voltage of the power system are exposed to swing. Authors in [13]

presented a static model based on the power flow and optimal power flow in order to control the under-frequency oscillations in an islanded microgrid. the application of this method is very limited and it only covers frequency oscillation of the microgrid. In [14], a low voltage feedback controller was proposed, which is based on theoretical-circuit analysis techniques in closed-loop systems. Authors in [15], presented a control scheme based on the V-I and I-V drop characteristics for MG voltage and current conditions. The proposed method determines the output impedance of the resource subsystem along with the converter's dynamics and analyzes the stability of the MG when it supplies constant loads. The paper investigates the sustainability effects of key parameters such as loss coefficients, local loop control dynamics and the number of resources and compares the voltage and current status from a sustainability perspective. Asghar et al. [16] developed a new control mechanism based on fuzzy logic and energy storage to control the frequency and voltage of the MG in the islanded mode. In this paper, both battery storage and super-capacitor have been used to improve the MG frequency oscillations and voltage stability, respectively. In [17] a decentralized robust fuzzy control strategy for islanded operation of an AC microgrid with voltage source inverters has presented. The objective of [17] is to design a robust controller for regulating the load voltage and sharing power among DGs in the presence of uncertainties in the system and non-linear loads. Authors in [18], based on the output regulation theory and fast-battery storage, designed a controller to improve the frequency variations and the voltage stability of the MG. They attempted 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 MG was mathematically modelled in [12]. In addition, a proper control strategy based on the obtained mathematical model has been proposed to improve the frequency fluctuations and voltage deviations of the MG and tested experimentally. Various master-slave and drop based control methods for improving the frequency and voltage oscillations in an MG are presented and compared in [19]. Differently from drop-based methods, in the master-slave based methods, the converter does not participate in the process of controlling the frequency and voltage. Although utilizing parallel converters in AC MGs and controlling them using drop based methods make the splitting of power between lines possible, sometimes, due to different lines impedances, a sudden load change causes the instantaneous imbalance in the production and power absorbed by the parallel converter. Therefore, in [20], the authors have proposed a control method to improve the voltage stability of the MG by sampling the difference in line impedances. A modified decentralized droop controller for inverter-based PV microgrids has presented in [21] that address the problem of instability and slower power-sharing between PV inverters.

The use of machine learning approaches in the areas of scheduling, maintenance management, quality improvement and control of MGs for effective solutions increased due to the development of new data measurement and communication techniques [22]. The Markov-decision process (MDP) is a discrete-time stochastic process partly under control of an action selector [23]. Recently, artificial intelligence methods are more considered by researchers for solving MDPs, and among these, machine learning techniques, including reinforcement learning (RL), are one of the most important methods. In [24], a thyristor controlled series capacitor (TCSC) is optimally controlled in order to achieve the stability of a multi-machine power system using the RL mechanism. According to the literature, the RL-based methods are model-free and do not require robust and accurate assumptions about system dynamics. In fact, these methods consider the system as a black box and model its dynamic behavior using its inputs and outputs, meaning that they can properly cope with nonlinearities and uncertainties despite partial information [25]. This is a useful property for controlling the nonlinear power systems, especially when, the complex and widespread power systems are forced to experience new undesirable conditions due to different events involving their various components. In a multi-agent system (MAS) RL, intelligent agents are compatible with their environment (system under control), meaning that when they perform an action, immediately receive feedback from the environment, thereby they update the probability of re-election of the elected action based on the received reward/penalty in

the corresponding state. One more interesting characteristic of the proposed RL based controllers is that they can be used as supervisory controllers for the classic controllers in a way that improves their dynamic response [25]. In [26] authors have presented a supervisory controller based on the RL method for controlling the frequency of a hybrid MG integrated with various energy storage systems. It is assumed the RL controller supervises the PID controller through its offset signal. Reference [27] presents two utilization for RL, at the first one RL used as an optimization tool for tuning the PID based power system stabilizer (PSS) control coefficients, while in the other one it is used as an appropriate alternative for the PSS system. The results prove that the RL controller can be a complimentary as well as a worthy alternative for classical controllers. Various parables have been reported in the literature for RL applications in power system control such as voltage control [28], market management [29], and stability analysis [30,31].

As it is concluded in [32], supervisory controllers such as rule-based system (RBS) and machine learning system (MLS) schemes have found to be flexible control methods with fast dynamic responses. RBS may be used in three categories such as fuzzy systems [33], neural networks [34], and the other standard RBS [35,36]. Although the fuzzy-based RBS and neural network schemes are more adaptable, still, they can produce unreliable results [32]. On the other hand, to ensure the robustness, reliability, and flexibility of MGs, MAS based decentralized control approaches have been recommended for MG management [25,30,37,38]. The prominent MAS intelligent agent's characteristics in the different aspects are classified as [32]: (i) reactivity: the ability to react to changes in the environment in a timely manner; (ii) pro-activeness: goal-directed behavior; (iii) social-ability: interaction with other agents. One of the main reasons for proposing the supervisory controllers in this paper is their capability to integrate with existing traditional control methods (such as PID) and improve their performance by adjusting their output. In other words, RL makes the traditional controllers be adaptive. This is important for several reasons. First, traditional controllers have acceptable reliability and are now used in various aspects of the industry. Secondly, the integration of RL supervisory control to them improves their performance while not requiring much cost. Moreover, basic controls already exist. In summary, the RL supervisory controller integrates the superb performance and fast dynamic response of RL with the reliability of traditional controllers and provides a robust and compatible controller at no much cost.

The primary objective of this paper is to formulate the problem of adaptive simultaneous control of voltage and frequency of a microgrid using game theory concept and try to find its optimal solution by multi-agent RL. In order to pursue the main goal of the paper, first, the dynamic nonlinear model of the microgrid test system is mathematically formulated and modeled with SIMULINK. Then the continuous-time nature of the system is discretized into a finite number of states to form the MDP. At this moment, the game theory formulation is done by determining the state, action, and reward/penalty factor characters. In this case, the complex control problem transforms into a game environment whose answer is the stability of the system. By placing two intelligent and autonomous agents in this environment (one for voltage and one for frequency) and giving them time to interact with the environment in order to reach to the optimal answer of the game (system stability) through an optimal control policy, it can be expected that these agents can always ensure the stability of the system in online simulations. The intelligent and autonomous agent means that each of these agents operates independently, and by using the reward/penalty that they receive from the environment can serve the main purpose of the game, which is reaching the game's answer. The Q-learning strategy, which is considered, to solve the RL-based game theory problem in this paper, is a model-free and a simple solution mechanism [39]. According to Figure 1, due to the nature of RL that learns the optimal control policy (OCP) by interacting with the environment [40], this paper first performs the offline simulations in which the intelligent agents use trial and error method to extract the OCP. The OCP means to determine the state-action pairs, $(s, a)$ in which the action $a$ is the best choice in the state $s$.

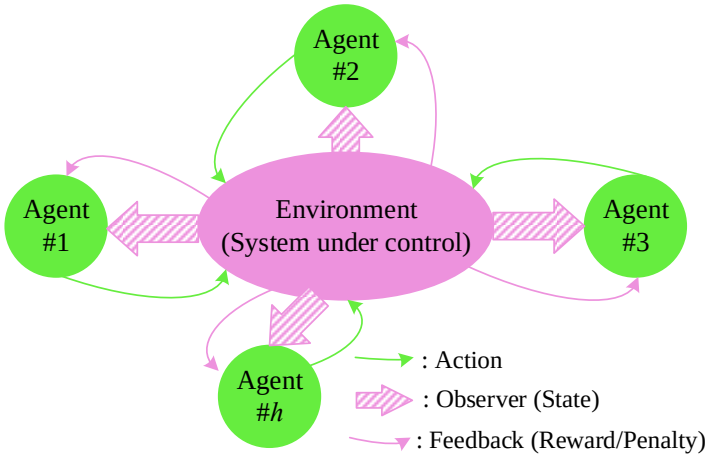



**Figure 1.** Multi-agent system (MAS) reinforcement learning procedure with $N$ intelligent agents.

Eventually, the simulations are continued online where the intelligent agents use the learned OCP in order to control the microgrid. A key point in online simulations is that each agent updates it's knowledge (OCP) about the environment (system under control) while performing the OCP to control the system. Accordingly, the proposed control mechanism is adaptive and can robustly cope with system uncertainties and operating condition changes.

The primary motivation of this work is to model the continuous-time environment of the MG control as an MDP and solve it using multi-agent reinforcement learning. The proposed control method is simple and adaptive and can properly cope with system nonlinearities and uncertainties. In order to provide the possibility of simultaneous voltage and frequency controlling in an MG, the reduced $Y_{bus}$ concept is used for the modeling of the system.

The innovative contributions of the present work are summarized below:

1. Modelling the continuous-time environment of MG control as an MDP and solve it using multi-agent reinforcement learning.
2. Considering independent intelligent agents to control the voltage and frequency in order to implement multi-agent reinforcement learning.
3. Using model-free Q-learning to cope with system nonlinearities.
4. Suggesting a simple strategy to assign the proper instant reward to the voltage and frequency agents according to system dynamics.
5. Employing the nonlinear model of a real microgrid at realistic scenarios for assessing the proposed MDP-based control strategy.

## 2. Materials and Methods

Reinforcement learning is originally developed for MDPs. In fact, the MDP provides a mathematical structure for modeling decision making in situations where outcomes are partly random and partly under the control of an agent [41]. Therefore, to employ RL in control theory, it is necessary to model the system under control as a finite MDP [24,42,43]. Given that the strategy of this paper is to control the frequency and voltage of the MG, in the following the control problem is modeled as a finite MDP and then solved using RL. Consequently, in this section, a short review of MDPs is provided and then the MG control modeling based on MDP along with solution methods using MAS-RL and Q-learning are proposed.

### *2.1. The Suggested Game Theory Approach*

#### 2.1.1. Markov Decision Process

As shown in Figure 2, the MDP is a stochastic process, which is comprised of discrete states along with a number of feasible actions for each state that are selected randomly by an action selector based on a distributed probability function [23,44]. In this paper, MDP is considered as a discrete-state and discrete-time systematic method to model the control of the MG. In the proposed MDP model in this paper, the set of states is referred to as the system states that are obtained using the rotational speed of the axis of system synchronous generators. The set of actions is indicated the available actions for system states. In addition, to show satisfaction with the action chosen by the intelligent agent, a reward/penalty function is considered. In the following, reinforcement learning is presented as an MDP solution method, and then the details of modeling the microgrid control problem using the MDP are expressed mathematically.

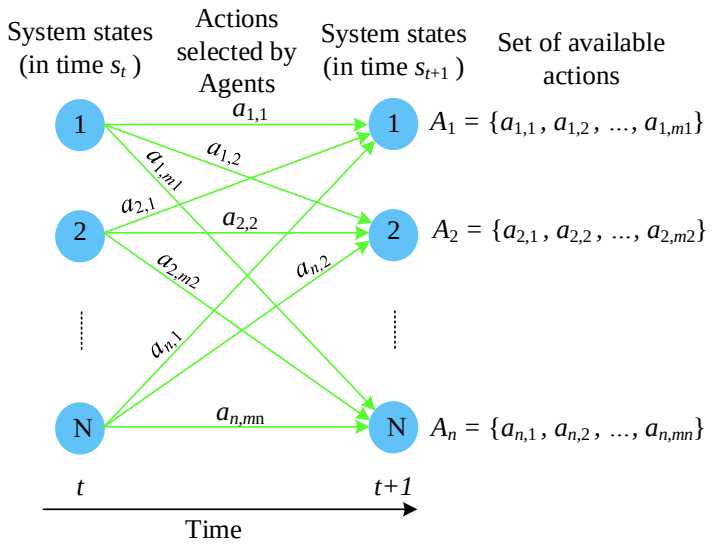



**Figure 2.** Stochastic Markov decision process (MDP) problem with states and the set of available actions.

#### 2.1.2. Reinforcement Learning

Reinforcement learning refers to the process of learning using the interaction [45]. An algorithmic approach, which one or more agents learn the OCP by interacting with the environment [46]. In this strategy, a fundamental assumption is that the environment is comprised of a finite number of states, each with a set of feasible actions to select. Continuous-time environments must be discretized into discrete spaces. The OCP means the intelligent agents determine the optimal action to be selected in each state [47]. Several model-based and non model-based methods such as adaptive heuristic critic (AHC), average reward (AR), and *Q*-learning are usually used for solving the RL problems. In this paper, the well-known *Q*-learning method, which is model-free with a simple mechanism is used [48].

#### 2.1.3. *Q*-Learning

*Q*-learning is one of the most interesting non-model based forms of RL. It can also be considered as a scheme of asynchronous dynamic programming. It is adaptive and can robustly cope with system uncertainties and changes in operating conditions without any strong assumptions about the system dynamics [49]. In this method, agents have the ability to learn to act optimally and autonomously in Markovian regions by receiving the feedback from the consequences of actions, without needing to

establish maps of the regions [48]. In order to mathematically model the proposed $Q$-learning strategy, suppose the system under control is comprised of a finite number of states ($n$) denoted by $S$.

$$S = \{s_1, s_2, s_3, \ldots, s_n\}, \quad (n \in \mathbb{N}) < \infty, \tag{1}$$

where, $S$ refers to system states and $\mathbb{N}$ is the set of natural numbers. For each intelligent agent, a competency matrix is formed, which is called the $Q$ matrix that accommodates the score of the selection of each action ($a$) for each state ($s$) of the system. In other words, the intelligent agent in each state of the system decides on the basis of the data in this matrix to select the appropriate action. The initial value of this matrix is set to zero for all pairs of ($s$, $a$), which are gradually updated as the learning process progresses. One of the acceptable criteria for ending the learning process is that the $Q$ matrix remains constant for two consecutive episodes. For example, assume an environment with $n$ states, which $m$ actions are available at each state. The initial $Q$ matrix is defined as given by (2).

$$Q_o = \begin{bmatrix} & a_1 & a_2 & a_3 & \ldots & a_m \\ s_1 & 0 & 0 & 0 & \ldots & 0 \\ s_2 & 0 & 0 & 0 & \ldots & 0 \\ s_3 & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s_n & 0 & 0 & 0 & \ldots & 0 \end{bmatrix}. \tag{2}$$

At each time step of the simulation, the autonomous agents obtain their current state ($s_t$) using a predetermined algorithm at the pre-processing stage. They then select the proper action using the $\varepsilon$-greedy method. In this way, in the current state $s_t$, with a probability of $\varepsilon$, the agent chooses an action with the highest value among the all available actions in $s_t$ accordance to $Q$ matrix, otherwise (with the probability of $1 - \varepsilon$), randomly chooses one of all actions available in-state $s_t$. The mechanism of $\varepsilon$-greedy is illustrated in Figure 3. Adjusting the amount of $\varepsilon$ can distinguish the learning and exploitation phases. The amount of $\varepsilon$ in the learning phase should be small enough to allow the agent to gain new experiences to achieve the OCP, but in the exploitation phase, the larger value should be selected so that the agent can perform the OCP while updating its knowledge ($Q$ matrix).

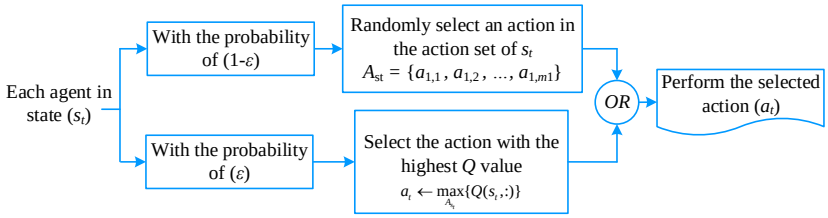



**Figure 3.** The $\varepsilon$-greedy mechanism for selecting the proper action.

It should be noted that although the agents operate independently of each other and they are completely autonomous, they all operate within the overall goal of the system. For this reason, each agent immediately receives feedback from the environment after selecting an action, so that if the action selected is in line with the overall goal, it will be rewarded but if it is in conflict with the overall goal it will be penalized. The value of this feedback is added to the knowledge of the intelligent agent by updating the $Q$ matrix and the agent will consider it in subsequent decisions. The flowchart of the proposed Q-learning strategy is depicted in Figure 4.

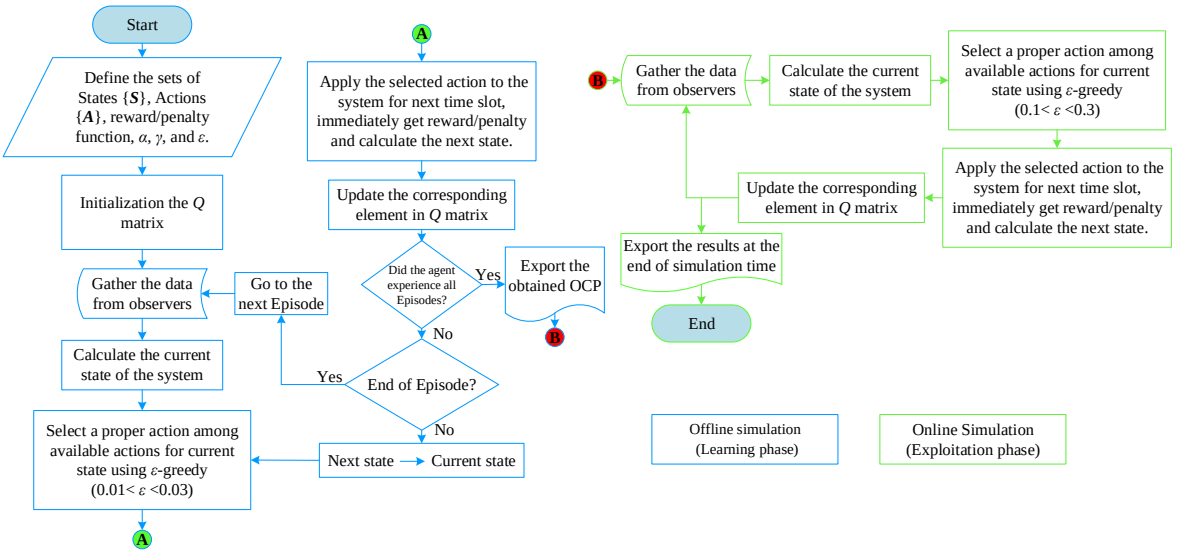



**Figure 4.** The flowchart of the proposed *Q*-learning control strategy.

In Figure 4, after defining the states, actions, and reward/penalty function, the intelligent agent interacts with the environment (system under control) to extract the OCP in the offline simulation. Once the learning process is complete (point B in Figure 4), the agent controls the system utilizing the OCP extracted in the previous step. After selecting and implementing the action, agents measure the satisfaction of the environment from the action selected through a concept known as the discounted long-term reward. The discounted long-term reward is calculated by (3) based on the instant feedback that is received from the environment [49]. In fact, the overall goal for all autonomous agents is to maximize the discounted long-term reward by finding the best action in each system state.

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \quad \gamma \in [01], \tag{3}$$

where, $r_t$ is the feedback that the agent receives from the environment after performing action $a_t$ at time step $t$ and $R_t$ indicates the discounted long-term reward. According to (3), the expected value of the $Q$ matrix is calculated by (4).

$$Q^{\pi}(s,a) = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|_{s_t=s,\ a_t=a}\right\}, \tag{4}$$

where, $\pi$ and $\pi^*$ are the current and optimal control policies, respectively. Considering the Bellman optimal equation, which is expressed by (5) [24,50,51], the $Q$ matrix elements are updated by (6) at each time step.

$$\Delta Q = \alpha\left\{r_{t+1}\ \gamma\ max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right\} \tag{5}$$

$$Q_{t+1} = Q_t + \Delta Q, \tag{6}$$

where, $\alpha$ is attenuation factor and $\alpha \in (0,1)$. It is worth remembering that the $Q$-learning-based controller's performance highly depends on how the states, actions, and feedback are defined, which are described in more detail below.

### 2.1.4. States

The MDP allows a single agent to learn a policy that maximizes a possibly delayed reward signal in a stochastic stationary environment. It guarantees convergence to the optimal policy, provided that the agent can sufficiently experiment and the environment in which it is operating is Markovian.

However, when multiple agents apply reinforcement learning in a shared environment, conditions may be a little different. The central idea of game theory is to model strategic interactions as a game between a set of players. A game is a mathematical object which describes the consequences of interactions between player strategies in terms of individual payoffs. In such systems, the optimal policy of an agent depends not only on the environment but on the policies of the other agents as well [52]. In the proposed framework two agents for frequency and voltage are considered that try to obtain the OCP. Note that, although the agents affect the other agent's operations, they are naturally autonomous. Figure 5 shows the suggested game-theory RL framework.

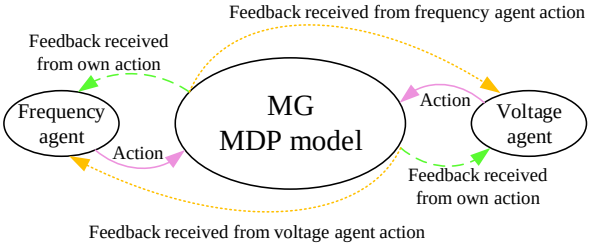



**Figure 5.** The proposed game-theory based reinforcement learning (RL) framework description considering the frequency and voltage agents.

As it was stated before, the control of frequency and voltage is the primary objective of this paper, therefore, the $\Delta\omega$ and $\Delta V$ signals are used as the feedback from the system under control to form the proposed MDP model. It should be noted that the angular velocity of synchronous generators in combined heat and power (CHP) units directly affects the frequency of the MG. According to literature, different signals can be considered to determine the state of the system [31], in this paper a combination of the angular velocity of the synchronous generator of CHP unit located at bus 01, $\Delta\omega_1$, along with the voltage deviation of bus 01, $\Delta V_1$ are considered for determination of the system oscillatory state, that is because the MGs are small and interconnected systems. Given the permissible range for frequency deviation (in p.u. system $\Delta F$ is equal to $\Delta\omega$) [53,54], the span of $-0.02$ to $+0.02$ is divided into 50 equal segments and (7) is utilized at each time step to determine the state of the system [31]. Figure 6 shows the discrete environment of the proposed MDP system. According to Figure 6, the zero-centred states that have been marked with the green box, are called the normal states, and the intelligent agent does not do anything in these states. In fact, this area is considered as the goal of the agents, meaning that their actions rewarded/penalized based on the distance from this area. According to (7), the mechanism of determining the state of the system consists of two components. The actual values of $\Delta\omega$ and $\Delta V$ plus their derivatives are intended to detect the severity and the elapsed time from the moment when the oscillation commences. The intelligent agents consider the sign of derivative components to make it clear whether the oscillations are going to the instability or moving towards the establishment [30].

$$s_t = \zeta(\Delta\omega_t^1, \frac{d\Delta\omega_t^1}{dt}, \Delta v_t^1, \frac{d\Delta v_t^1}{dt}), \tag{7}$$

where, $\Delta\omega_t^i$ and $\Delta V_t^j$ are the angular velocities of the generation unit $i$ and voltage deviation at bus $j$, respectively. The agent will analyze the feedback signals in pre-processing stage and detects the disturbance occurrence after an acceptable time. The process of determining the real state of the system in the pre-processing stage is illustrated in Figure 7.

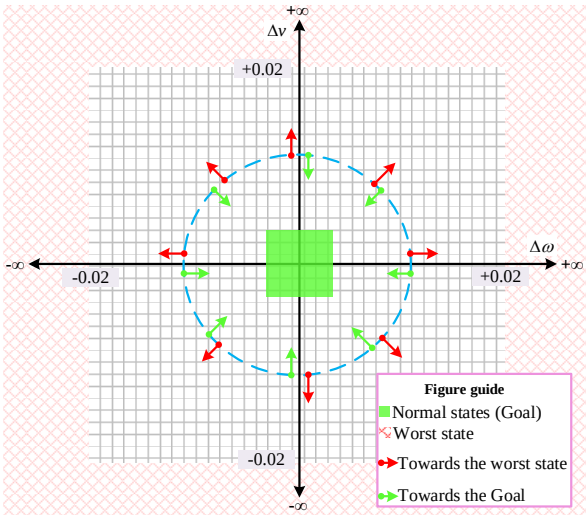



**Figure 6.** The two-dimensional perspective of the discrete environment of the proposed MDP model.

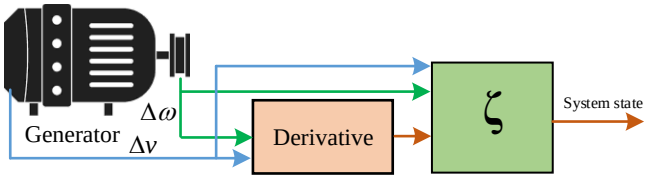


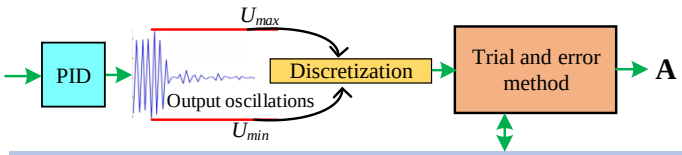

**Figure 7.** The process of determining the current state of the system.

### 2.1.5. Actions

Determining the accurate actions for different states of the system is a relatively complex issue. Since there is no specific law, it relies more on trial and error methods [55]. However, using the output of traditional control methods in the same application can be a good aid for this matter [31]. The set of feasible actions can be different for various states of the environment. Moreover, the number of actions can be greatly increased. Although this may improve the quality of the control system, from another side it increases the learning time extremely so it may make it impossible to obtain the OCP. For simplicity, this paper considers the same set of available actions for all states of the environment. According to (8) for each state, ten actions are suggested that five actions are considered for the frequency agent and five actions for the voltage agent.

$$A = \begin{Bmatrix} [-0.002, -0.001, 0, 0.001, 0.002] & For \quad Freq. \\ [-0.0002, -0.0001, 0, 0.0001, 0.0002] & For \quad Volt. \end{Bmatrix} \tag{8}$$

The proposed strategy for defining actions is shown in Figure 8. As can be seen from Figure 8, in the pre-processing stage, the information is gathered from the environment (system under control) and then the action set is defined based on a few simple assumptions about the range, size and number of actions.




**Considering some simple laws:**
1- Selecting a value out of the [$U_{min}$ $U_{max}$] makes the system unstable.
2- Increasing the number of actions increases the time of learning phase.
3- It is recommended that the actions be symmetrical.

**Figure 8.** A simple description for generating the action set.

#### 2.1.6. 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 $a_t$ among the actions available for state $s_t$. Immediately, the agent receives feedback (reward/penalty) from the (environment) system under control concerning the performed action. This scheme is depicted in Figure 9. Based on this feedback, the agent assigns a score (positive for reward and negative for the penalty) for the pair of $(s_t, a_t)$ and updates the corresponding element of $Q$ matrix. If the score is positive, the probability of performing the action $a$ at the state $s_t$ increases for the next time's experiences. Otherwise, if the score is negative (penalty), the agent selects the action with a lower probability in the state $s_t$, next time.

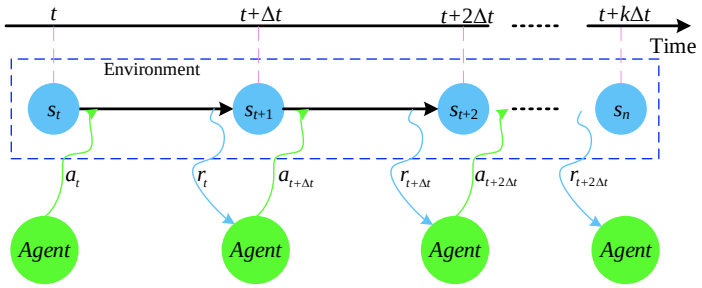



**Figure 9.** The operating process of the agent in a discrete environment in time steps.

With this intention that the primary objective of this paper is voltage and frequency control, therefore $\Delta\omega$ and $\Delta v$ signal of all generation units are selected for determination of reward/penalty for corresponding $(s_t, a_t)$ pairs. In essence, if an action causes the system to go out of the normal state (towards the red area in Figure 6), it will be marked as a wrong action in the current state and will be penalized. In return, if an action causes the system to go to the normal state (towards the green area in Figure 6), will receive the highest reward. In summary, the reward/penalty function is described by (9)–(12), in this paper. Based on these equations, it is assumed the reward/penalty factors of frequency (voltage) is 80% related to the action of frequency (voltage) controller and 20% related to voltage (frequency) controller.

$$r_t^f = \sum_{i=1}^{G} \left[ \frac{1}{\left(1 + \sum_{k=t-1}^{t} (\Delta\omega_i(k)\right)}, \right] \tag{9}$$

where, $G$ is number of generation buses. $r_t^f$ and $r_t^v$ indicate the frequency and voltage viewpoint of the reward, respectively.

$$r_t^v = \sum_{j=1}^{G} \left[ \frac{1}{\left(1 + \sum_{k=t-1}^{t} (\Delta v_j(k)\right)} \right] \tag{10}$$

$$\Re_t^f = \begin{Bmatrix} +1 & s_{t+1}\ normal\ state \\ -1 & s_{t+1}\ worst\ state \\ 0.8 \times r_t^f + 0.2 \times r_t^v & Otherwise \end{Bmatrix} \tag{11}$$

$$\Re_t^v = \begin{Bmatrix} +1 & s_{t+1}\ normal\ state \\ -1 & s_{t+1}\ worst\ state \\ 0.2 \times r_t^f + 0.8 \times r_t^v & Otherwise \end{Bmatrix}, \tag{12}$$

where, $\Re_t^f$ ($\Re_t^v$) is the score of the selected action $a_t$ at state $s_t$ and time step $t$ for frequency (voltage) agent. The implementation idea of the proposed control method based on RL is depicted in Figure 10.

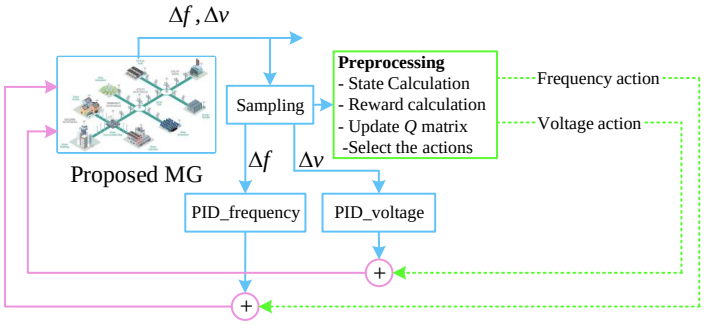



**Figure 10.** The implementation idea of the proposed control method based on RL.

## 2.2. Dynamic Modelling of the Microgrid Test Case

In order to assess the effectiveness of the proposed RL based control strategy, here, a typical distribution network located in Denmark with 10 loads and six generating units is considered. The microgrid understudy consists of 3 wind turbine generators namely $WTG_1$, $WTG_2$ and $WTG_3$ along with three gas turbine generators called $CHP_1$, $CHP_2$ and $CHP_3$. The $WTG$ units operation power factor is assumed close to unit (suppose they are equipped with proper compensators ). A normal operating point is considered for microgrid operation, which it is assumed the output power of each $WTG$ units is 0.08 MW and the output power of the $CHP$ units are 2.5, 2.8, and 2.8 MW for $CHP_1$, $CHP_2$, and $CHP_3$, respectively. The single-line representation of the proposed test microgrid with the location of loads and generating units is shown in Figure 11 [56].

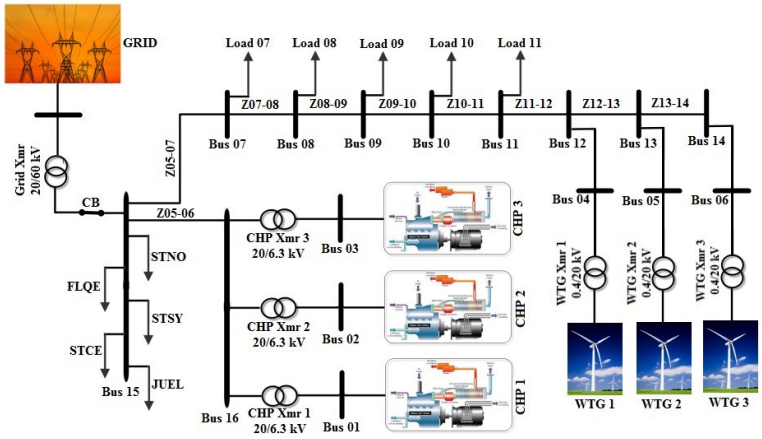



**Figure 11.** The single line diagram of the proposed real micro-grid (MG).

### 2.2.1. Fixed-Speed Wind Turbine Generator Model

The third-order model of the asynchronous generator is used in the present work. In the model shown in Figure 12, $C_p^i$, $\lambda_i$, $\beta_i$, and $T_{rot}^i$ are the aerodynamic power coefficient, speed ratio, pitch angle, and aerodynamic torque of the ith WTG, respectively. These components can be modelled by (13)–(15) [57].

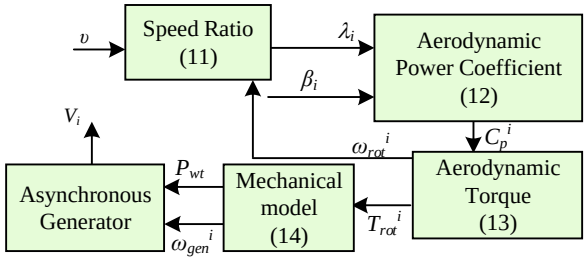



**Figure 12.** Conceptual model of the Fixed-speed wind generator [57].

$$\lambda_i = R_i \omega_{rot}^i V^{-1} \tag{13}$$

$$C_p^i = (0.44 - 0.0167\beta_i) \sin\left[\frac{3.1415(\lambda_i - 3)}{15 - 0.3\beta_i}\right] - 0.0184(\lambda_i - 3)\beta_i \tag{14}$$

$$T_{rot}^i = \frac{1}{\omega_{rot}^i} C_p^i \left[\frac{\rho_i \pi R_i^2 V^3}{2}\right], \tag{15}$$

where $\rho_i$, $R_i$, $v$, and $\omega_{rot}^i$ refer to air density, rotor radius, wind speed, and rotor speed, respectively. Finally, the dynamic model of the mechanical part of the asynchronous generator is described by (16)

$$\frac{d\omega_{rot}^i}{dt} = \frac{\omega_o}{2H_i}\left[T_{rot}^i - T_e^i - D_i \frac{\omega_{rot}^i}{\omega_o}\right] \tag{16}$$

### 2.2.2. Combined Heat and Power Plant Model

In general, a CHP unit consists of a gas turbine mechanically coupled with a synchronous generator and a droop controller is used to regulate the frequency. In this paper, a modified gas turbine model is used [56]. Since the operation of a CHP in the islanded mode of MG faces to complex challenges, the conventional droop controllers cannot properly regulate the frequency. Because their parameters are set once and at one operating point. Therefore, a supervisory RL based PID controller is added to the control loop of the gas turbine. Figure 13 shows the modified gas turbine model of a CHP.

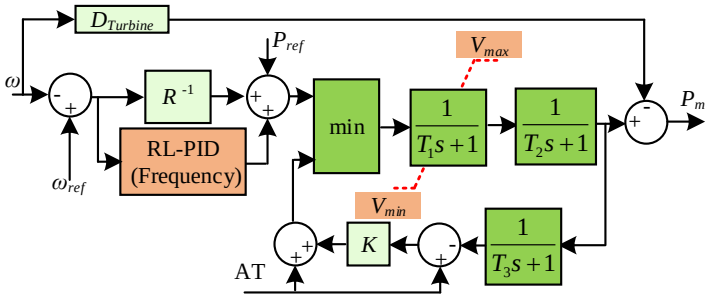



**Figure 13.** The modified model of the gas turbine of combined heat and power (CHP) with an extra adaptive controller [57].

In addition, the IEEE's type AC5A excitation system is used as the synchronous generator's excitation system [56]. As shown in Figure 14, an adaptive RL-PID controller is added to the typical AC5A as the secondary control tool.

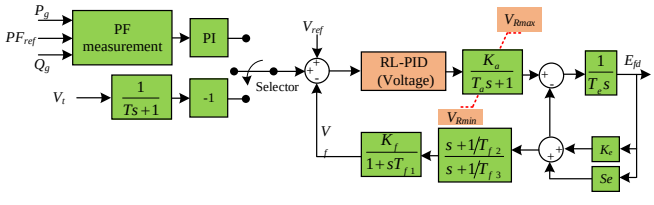



**Figure 14.** The modified IEEE's type AC5A with secondary voltage controller.

Load, *WTG*, and exciter system data are taken from [56]. The well-known reduced admittance matrix $Y_{bus}^{red}$ [58] is utilized here for modelling the proposed test system. For this aim, firstly the impact

of loads, capacitor and reactor compensators entered in the modelling as a constant admittance, which can be calculated using (17).

$$y_i = \frac{P_i - jQ_i}{V_{base}}, \tag{17}$$

where $P_i$, $Q_i$, and $V_{base}$ are the active power, reactive power at bus $i$, and the base voltage, respectively. The parameter $y_i$ is used as a shunt admittance in the corresponding bus and is entered into transmission admittance matrix $Y_{bus}$. The size of $Y_{bus}$ can be reduced by removing load buses from the original $Y_{bus}$. In this method, the impact of non-generator buses is transferred into generator buses using (18).

$$Y_{bus}^{red} = Y_{GG} - Y_{GN} \times Y_{NN}^{-1} \times Y_{NG}, \tag{18}$$

where

$$Y_{bus} = \begin{bmatrix} Y_{GG} & Y_{GN} \\ Y_{NG} & Y_{NN} \end{bmatrix}. \tag{19}$$

As an illustration, Figure 15 shows the final modeling stage, in which the dynamics of the micro energy grid can be evaluated.

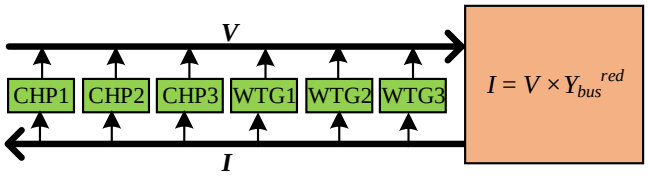



**Figure 15.** Dynamic $Y_{bus}^{red}-$ based model of the considered MG.

## 3. Simulation Results

In this paper, the design of the controller for WTG generation units is ignored in order to simplify the MG control strategy using the MDP system and solve it using RL. This is a reasonable assumption because the share of wind units is negligible compared to the total system power production. Since the CHP units are quite similar, the same controllers are considered for all of them. So, two intelligent agents are considered. One agent that is responsible for controlling the frequency oscillations and the other one is responsible for controlling the voltage fluctuations of all CHP buses. As mentioned earlier, these two intelligent agents (frequency agent and voltage agent) are completely independent of each other and operate autonomously. In this section, the effectiveness of the suggested RL-PID controller is evaluated compared to a classical PID and a fuzzy PID (F-PID) [59]. The optimal design procedure of the PID and F-PID controllers is not in the scope of this paper and therefore is ignored. The optimal setting of the controllers can be done in various methods including numerical methods such as Ziegler–Nichols and metaheuristic algorithms. In this paper, it is assumed the PID and F-PID controllers have optimized using the salp swarm algorithm (SSA) [60] and the results are available as tabulated in Table 1.

The details of the F-PID controller are given in [59]. Finally, two realistic scenarios are considered as challenging operating conditions of the MG in island mode.

**Table 1.** Control gains of fuzzy proportional integral derivative (F-PID) and PID controllers optimized by salp swarm algorithm (SSA).

| **Control Type** | **F-PID Control** | | | | | |
|---|---|---|---|---|---|---|
| Param. | | $K_1^f$ | $K_2^f$ | $K_3^f$ | $K_4^f$ | |
| Value | | 0.8812 | 1.8235 | 1.6520 | 0.9512 | |
| Param. | | $K_1^v$ | $K_2^v$ | $K_3^v$ | $K_4^v$ | |
| Value | | 0.0758 | 1.4510 | 1.3851 | 0.0851 | |
| **Control Type** | **PID Control** | | | | | |
| Param. | $K_p^f$ | $K_i^f$ | $K_d^f$ | $K_p^v$ | $K_i^v$ | $K_d^v$ |
| Value | 0.4978 | 0.1408 | 0.00117 | 0.0704 | 0.0383 | 0.0012 |

### *3.1. Scenario 1: Symmetric Three-Phase Fault*

In this scenario it is assumed that the MG is operated at its normal conditions and the parameter $T_2$ of gas turbine model of Figure 13 is changed +50%. At these conditions, a three-phase symmetrical fault occurs at time 5 sec near the bus 11. The fault causes the line between bus 10 and 11 to be tripped out and then re-connected after fault clearance in 40 ms. The $\Delta\omega$ and $\Delta v$ signals of all generators are plotted in Figures 16 and 17.

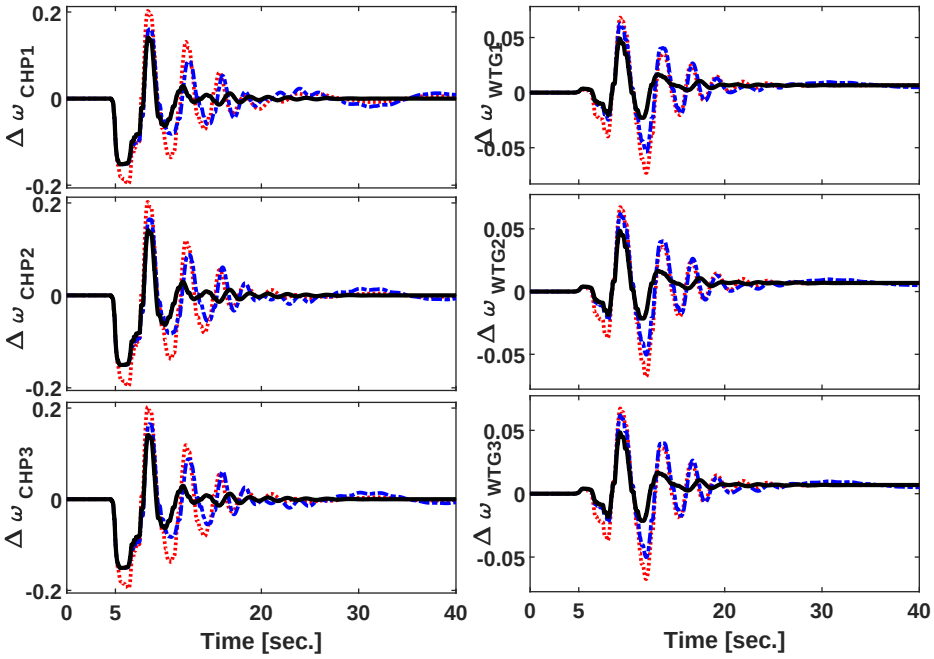



**Figure 16.** The oscillations of the angular velocity of generation units in *scenario* 1; black(solid): RL-PID, blue(dashed-dotted): F-PID, red(dotted): PID.

As can be seen from Figures 16 and 17, the proposed RL-PID controller has an extraordinary ability to stabilize the test system frequency and voltage variations compared to the traditional PID and F-PID controllers thanks to its flexible structure, which combines the machine learning compatibility feature along with PID controller precision and quick response property.

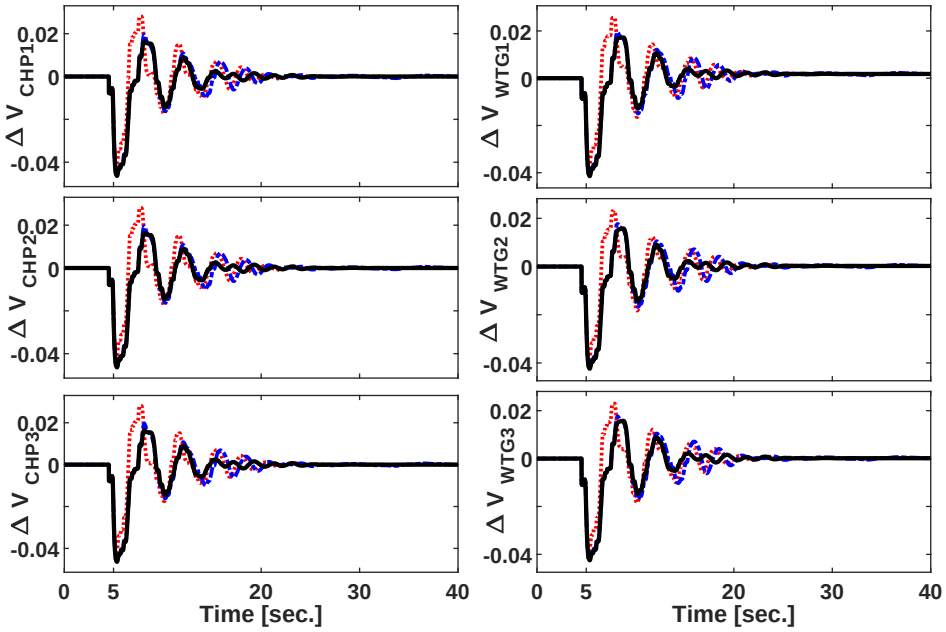



**Figure 17.** The oscillations of voltage at the generator buses in *scenario* 1; black(solid): RL-PID, blue (dashed–dotted): F-PID, red (dotted): PID.

Results show that the RL-PID makes the system dynamics better from the perspective of deviation overshoot(OS)/undershoot (US), decreasing settling time, and elimination of steady-state error.

### *3.2. Scenario 2: Sudden Load Connection/Disconnection*

In this scenario, it is assumed the parameter $T_2$ has decreased 50% and *STNO* load at bus 15 is off. At time 5 sec, the heavy load of *STNO* is connected suddenly. Immediately the generation units affected by increasing the load of the system and then their angular velocity is decreased. The control strategies try to damp the oscillations of $\Delta\omega$ and $\Delta v$ by modulating the error signal and applying it to the system through the turbine and excitation system of the generators. After a while, the oscillations will be damped and the system situation becomes stable again. Figures 18 and 19 show the $\Delta\omega$ and $\Delta v$ dynamics in scenario 2, respectively.

As can be seen from Figures 18 and 19, the proposed RL based controller has a superb dynamic characteristic compared to F-PID and PID controllers. In this study, the pre-processing analysis will be calculated in each 50 ms. This time is very important because it affects directly on the output of the RL controller. This means when a disturbance occurs, the agents can sense the effect of the disturbance utmost 50 ms after the occurrence regardless of the occurred disturbance's due. From Figures 18 and 19, that is why the RL controller has a poor effect on the first overshoot/undershoot after the fault occurrence. Furthermore, Figure 20 shows the actions were taken by the frequency and voltage agents in scenarios 1 and 2. It can be seen from Figure 20 that, the RL controller is inactive when the state of the system is normal but when the frequency/voltage of the system becomes unstable following a disturbance, RL starts generating the suitable control action. When oscillations are well-damped, the RL controller becomes inactive again.

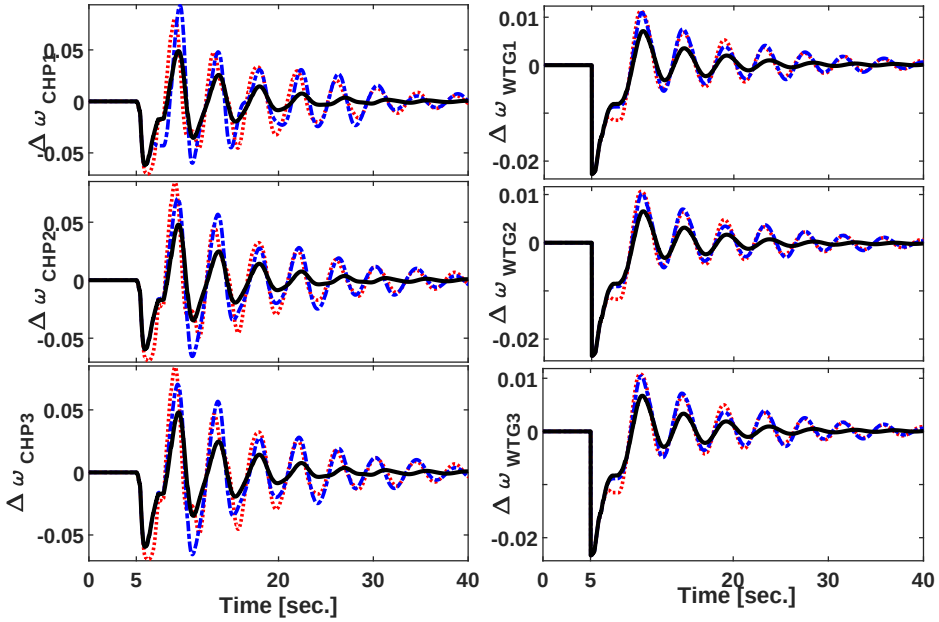



**Figure 18.** The oscillations of the angular velocity of generation units in *scenario* 2; black(solid): RL-PID, blue (dashed–dotted): F-PID, red (dotted): PID.

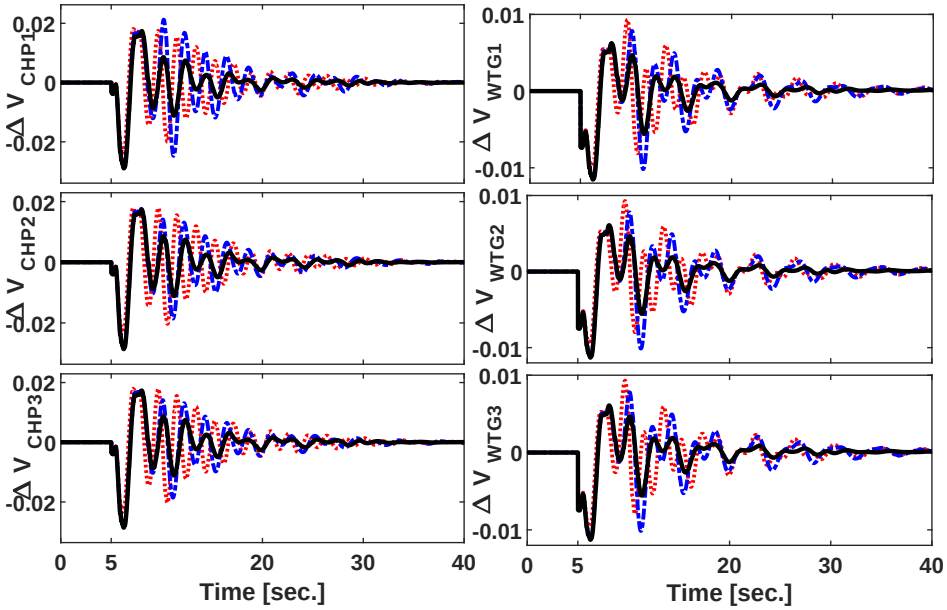



**Figure 19.** The oscillations of voltage at the generator buses in *scenario* 2; black(solid): RL-PID, blue (dashed–dotted): F-PID, red (dotted): PID.

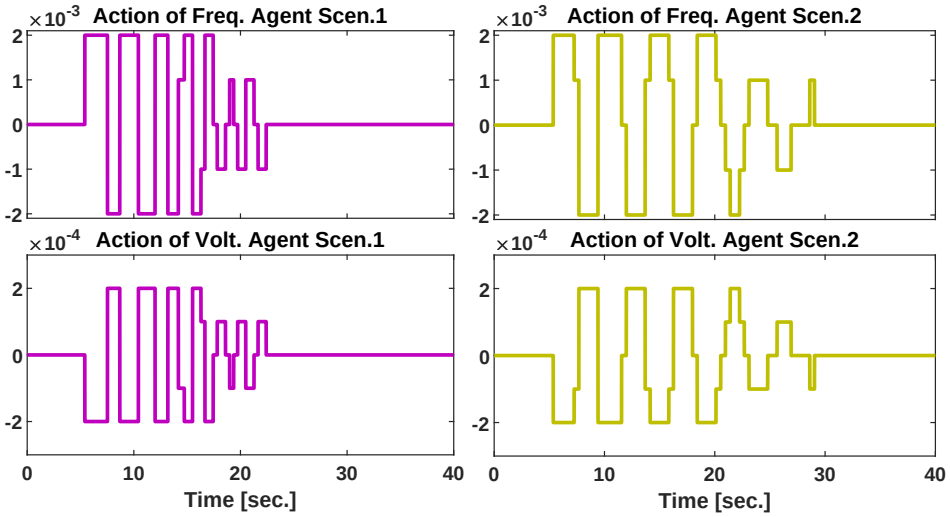



**Figure 20.** Actions selected by the RL frequency/voltage agents in different scenarios.

## 4. Discussion

According to simulation results provided in Section 3, the preponderance of the presented RL-PID controller compared to F-PID and PID control methods is evident. This paper has introduced a novel strategy for multi-agent learning used for integrated controlling the frequency and voltage of an MG. For more distinguishing the capabilities of the proposed game-theory based control strategy, time-domain analysis is provided in this section. For this purpose, two suitable time-domain indices based on the conventional integral of time multiplied by absolute error (ITAE), and integral of square error (ISE) are calculated and tabulated in Tables 2 and 3. The utilized ITAE and ISE indices are expressed by (20) and (21).

$$ITAE_y = \log_{10}\left(\int_0^{40} t \times |y| dt\right) \tag{20}$$

$$ISE_y = \log_{10}\left(\int_0^{40} 1000 \times y^2 dt\right). \tag{21}$$

As it can be seen from Tables 2 and 3, the suggested MDP based RL control scheme is successful in damping the oscillations of voltage and frequency of the test MG compared to F-PID and PID controller from the perspective of time-domain performance indices.

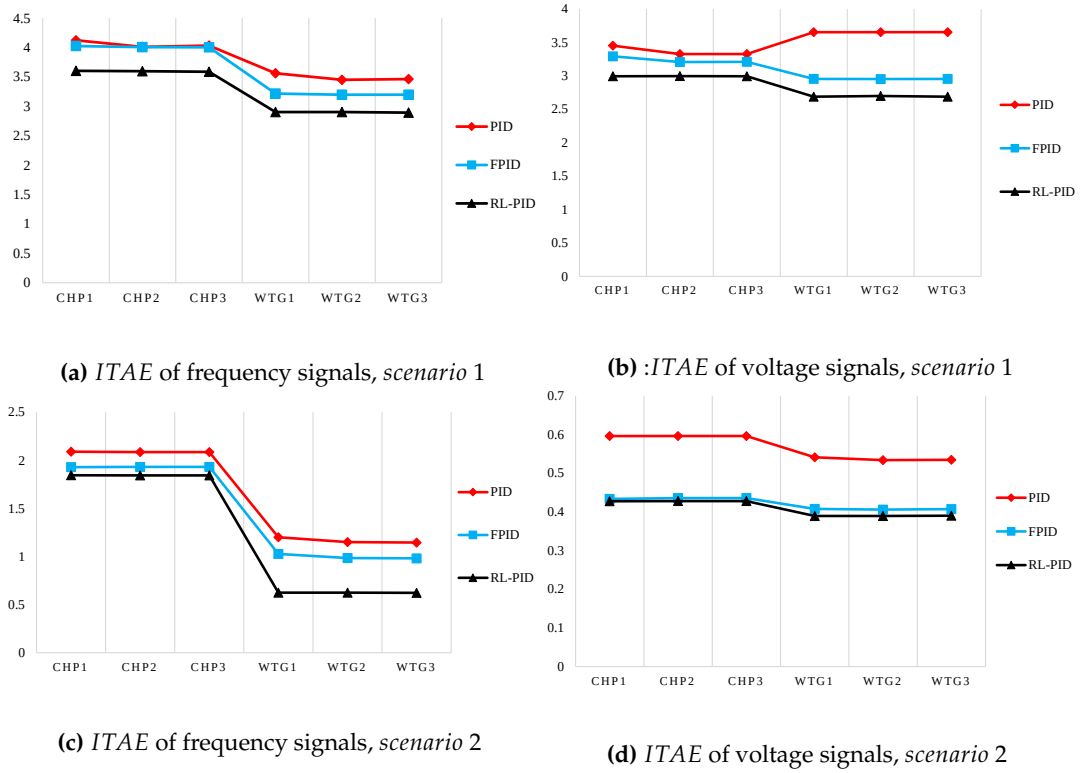
**Table 2.** Time domain performance indices in *scenario* 1.

| Signal | ITAE | | | ISE | | |
|---|---|---|---|---|---|---|
| | PID | FPID | RLPID | PID | FPID | RLPID |
| $\Delta\omega_{CHP_1}$ | 81.687 | 80.602 | **50.0690** | 2.143 | 1.949 | **1.783** |
| $\Delta\omega_{CHP_2}$ | 78.603 | 76.702 | **50.028** | 2.137 | 1.945 | **1.781** |
| $\Delta\omega_{CHP_3}$ | 78.603 | 76.602 | **50.012** | 2.102 | 1.938 | **1.695** |
| $\Delta\omega_{WTG_1}$ | 62.781 | 60.813 | **54.851** | 1.155 | 1.023 | **0.635** |
| $\Delta\omega_{WTG_2}$ | 62.760 | 60.934 | **54.721** | 1.124 | 1.003 | **0.642** |
| $\Delta\omega_{WTG_3}$ | 62.766 | 60.950 | **54.896** | 1.121 | 1.003 | **0.642** |
| $\Delta v_{CHP_1}$ | 13.130 | 10.259 | **2.896** | 0.595 | 0.496 | **0.446** |
| $\Delta v_{CHP_2}$ | 3.050 | 10.132 | **2.901** | 0.596 | 0.496 | **0.435** |
| $\Delta v_{CHP_3}$ | 13.030 | 10.102 | **2.901** | 0.578 | 0.486 | **0.446** |
| $\Delta v_{WTG_1}$ | 27.243 | 26.425 | **5.135** | 0.552 | 0.428 | **0.387** |
| $\Delta v_{WTG_2}$ | 14.209 | 11.164 | **5.135** | 0.561 | 0.418 | **0.376** |
| $\Delta v_{WTG_3}$ | 13.787 | 10.102 | **3.790** | 0.564 | 0.431 | **0.377** |

**Table 3.** Time domain performance indices in *scenario* 2.

| Signal | ITAE | | | ISE | | |
|---|---|---|---|---|---|---|
| | PID | FPID | RLPID | PID | FPID | RLPID |
| $\Delta\omega_{CHP_1}$ | 4.124 | 4.026 | **3.604** | 2.562 | 2.397 | **1.959** |
| $\Delta\omega_{CHP_2}$ | 4.012 | 4.007 | **3.597** | 2.456 | 2.333 | **1.932** |
| $\Delta\omega_{CHP_3}$ | 4.035 | 4.006 | **3.589** | 2.465 | 2.334 | **1.954** |
| $\Delta\omega_{WTG_1}$ | 3.562 | 3.215 | **2.903** | 1.021 | 0.987 | **0.891** |
| $\Delta\omega_{WTG_2}$ | 3.452 | 3.198 | **2.903** | 1.102 | 0.996 | **0.889** |
| $\Delta\omega_{WTG_3}$ | 3.465 | 3.198 | **2.893** | 1.125 | 0.991 | **0.883** |
| $\Delta v_{CHP_1}$ | 3.452 | 3.292 | **2.994** | 1.452 | 1.298 | **1.060** |
| $\Delta v_{CHP_2}$ | 3.326 | 3.207 | **2.996** | 1.432 | 1.189 | **1.059** |
| $\Delta v_{CHP_3}$ | 3.326 | 3.208 | **2.994** | 1.441 | 1.196 | **1.059** |
| $\Delta v_{WTG_1}$ | 3.652 | 2.953 | **2.688** | 0.856 | 0.517 | **0.333** |
| $\Delta v_{WTG_2}$ | 3.652 | 2.952 | **2.698** | 0.857 | 0.510 | **0.323** |
| $\Delta v_{WTG_3}$ | 3.654 | 2.953 | **2.689** | 0.858 | 0.510 | **0.332** |

It can be noted that the RL based controller has enhanced the ITAE criteria 1.1% to 10.44% compared to F-PID and 2.5% to 23.45% compared to PID controller. The ISE index is enhanced 1.45% to 12.54% and 4.1% to 26.89% compared to F-PID and PID controllers, respectively. In order to clarify the conclusion of the simulation results and the data of Tables 2 and 3, the graphic diagram of Figure 21 is depicted in various scenarios.




**(a)** *ITAE* of frequency signals, *scenario* 1

**(b)** :*ITAE* of voltage signals, *scenario* 1

**(c)** *ITAE* of frequency signals, *scenario* 2

**(d)** *ITAE* of voltage signals, *scenario* 2

**Figure 21.** The presentation of the time domain performance indices in various scenarios.

## 5. Conclusions

The primary goal of the presented work is to design an adaptive integrated voltage/frequency controller for damping the oscillations in a microgrid with penetration of wind power using the machine learning theory concept and trying to find its optimal solution by utilizing the multi-agent reinforcement learning. For this aim, first, the dynamic nonlinear model of the microgrid test system is mathematically formulated and modeled with SIMULINK. Then the continuous-time nature of the

system is discretized into a finite number of states to form the Markov decision process (MDP). At this moment, the game theory formulation is done by determining the state, action, and reward/penalty factor characters. Each state of the system represents the condition of the MG from the viewpoint of frequency and voltage oscillations. In addition, there are some feasible control actions at each system state, that can be applied in a way that the stability of the MG is ensured. Then the problem of the stability of the voltage/frequency of the MG is formulated as a multi-agent reinforcement learning problem. Finally, the defined control strategy is solved using the Q-learning modeling of RL. In this way, the independent autonomous agents are assigned to control the voltage and frequency simultaneously. Each agent tries to control its corresponding parameter(voltage or frequency) regardless of the behavior of the other agents. Once independent agents in the offline simulation learned the optimal control policy by interaction with the environment (test system), they can simultaneously control the MG and also update their knowledge about the system under study. For assessing the dynamic response of the presented control scheme compared to fuzzy PID (F-PID) and traditional PID controllers, a real island MG is considered and simulated using MATLAB/SIMULINK. Simulations were carried out in two realistic and challenging scenarios such as symmetric three-phase fault and load shedding considering system parameter changes. Results indicate that the proposed control strategy has an excellent dynamic response compared to F-PID and PID controllers for damping the voltage/frequency oscillations. It has improved the performance of the F-PID controller approximately 1% to 34% and the PID controller 10% to 55%. From the research that has been conducted it is possible to show the awesome capabilities of reinforcement learning based controller which can cope with system nonlinearities. It is model-free and can control the system without any initial strong assumptions.

**Author Contributions:** Conceptualization, H.S., and A.Y.; methodology, A.Y.; software, A.Y.; validation, H.S., P.S.; formal analysis, A.Y.; investigation, A.Y.; resources, A.Y.; data curation, A.Y.; writing–original draft preparation, A.Y.; writing–review and editing, H.S. and P.S.; supervision, H.S. and P.S.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Morshed, M.J.; Fekih, A. A fault-tolerant control paradigm for microgrid-connected wind energy systems. *IEEE Syst. J.* **2016**, *12*, 360–372. [CrossRef]
2. Magdy, G.; Shabib, G.; Elbaset, A.A.; Mitani, Y. A Novel Coordination Scheme of Virtual Inertia Control and Digital Protection for Microgrid Dynamic Security Considering High Renewable Energy Penetration. *IET Renew. Power Gener.* **2019**, *13*, 462–474. [CrossRef]
3. Alam, M.N.; Chakrabarti, S.; Ghosh, A. Networked microgrids: State-of-the-art and future perspectives. *IEEE Trans. Ind. Inf.* **2018**, *15*, 1238–1250. [CrossRef]
4. Gungor, V.C.; Sahin, D.; Kocak, T.; Ergut, S.; Buccella, C.; Cecati, C.; Hancke, G.P. Smart grid technologies: Communication technologies and standards. *IEEE Trans. Ind. Inf.* **2011**, *7*, 529–539. [CrossRef]
5. Ahmed, M.; Meegahapola, L.; Vahidnia, A.; Datta, M. Analysis and mitigation of low-frequency oscillations in hybrid AC/DC microgrids with dynamic loads. *IET Gener. Transm. Distrib.* **2019**, *13*, 1477–1488. [CrossRef]
6. Amoateng, D.O.; Al Hosani, M.; Elmoursi, M.S.; Turitsyn, K.; Kirtley, J.L. Adaptive voltage and frequency control of islanded multi-microgrids. *IEEE Trans. Power Syst.* **2017**, *33*, 4454–4465. [CrossRef]
7. Wu, X.; Shen, C.; Iravani, R. A distributed, cooperative frequency and voltage control for microgrids. *IEEE Trans. Smart Grid* **2016**, *9*, 2764–2776. [CrossRef]
8. De Nadai Nascimento, B.; Zambroni de Souza, A.C.; de Carvalho Costa, J.G.; Castilla, M. Load shedding scheme with under-frequency and undervoltage corrective actions to supply high priority loads in islanded microgrids. *IET Renew. Power Gener.* **2019**, *13*, 1981–1989. [CrossRef]
9. Shrivastava, S.; Subudhi, B.; Das, S. Noise-resilient voltage and frequency synchronisation of an autonomous microgrid. *IET Gener. Transm. Distrib.* **2019**, *13*, 189–200. [CrossRef]

10. Liu, Z.; Miao, S.; Fan, Z.; Liu, J.; Tu, Q. Improved power flow control strategy of the hybrid AC/DC microgrid based on VSM. *IET Gener. Transm. Distrib.* **2019**, *13*, 81–91. [CrossRef]
11. El Tawil, T.; Yao, G.; Charpentier, J.F.; Benbouzid, M. Design and analysis of a virtual synchronous generator control strategy in microgrid application for stand-alone sites. *IET Gener. Transm. Distrib.* **2019**, *13*, 2154–2161. [CrossRef]
12. Hirase, Y.; Abe, K.; Sugimoto, K.; Sakimoto, K.; Bevrani, H.; Ise, T. A novel control approach for virtual synchronous generators to suppress frequency and voltage fluctuations in microgrids. *Appl. Energy* **2018**, *210*, 699–710. [CrossRef]
13. La Gatta, P.O.; Passos Filho, J.A.; Pereira, J.L.R. Tools for handling steady-state under-frequency regulation in isolated microgrids. *IET Renew. Power Gener.* **2019**, *13*, 609–617. [CrossRef]
14. Simpson-Porco, J.W.; Dörfler, F.; Bullo, F. Voltage stabilization in microgrids via quadratic droop control. *IEEE Trans. Autom. Control* **2016**, *62*, 1239–1253. [CrossRef]
15. Gao, F.; Bozhko, S.; Costabeber, A.; Patel, C.; Wheeler, P.; Hill, C.I.; Asher, G. Comparative stability analysis of droop control approaches in voltage-source-converter-based DC microgrids. *IEEE Trans. Power Electron.* **2016**, *32*, 2395–2415. [CrossRef]
16. Asghar, F.; Talha, M.; Kim, S. Robust frequency and voltage stability control strategy for standalone AC/DC hybrid microgrid. *Energies* **2017**, *10*, 760. [CrossRef]
17. Hosseinalizadeh, T.; Kebriaei, H.; Salmasi, F.R. Decentralised robust T-S fuzzy controller for a parallel islanded AC microgrid. *IET Gener. Transm. Distrib.* **2019**, *13*, 1589–1598. [CrossRef]
18. Zhao, H.; Hong, M.; Lin, W.; Loparo, K.A. Voltage and frequency regulation of microgrid with battery energy storage systems. *IEEE Trans. Smart Grid* **2017**, *10*, 414–424. [CrossRef]
19. Ahmarinejad, A.; Falahjoo, B.; Babaei, M. The stability control of micro-grid after islanding caused by error. *Energy Procedia* **2017**, *141*, 587–593. [CrossRef]
20. Issa, W.; Sharkh, S.M.; Albadwawi, R.; Abusara, M.; Mallick, T.K. DC link voltage control during sudden load changes in AC microgrids. In Proceedings of the 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), Edinburgh, UK, 19–21 June 2017; pp. 76–81.
21. Firdaus, A.; Mishra, S. Auxiliary signal-assisted droop-based secondary frequency control of inverter-based PV microgrids for improvement in power sharing and system stability. *IET Renew. Power Gener.* **2019**, *13*, 2328–2337. [CrossRef]
22. Susto, G.A.; Schirru, A.; Pampuri, S.; McLoone, S.; Beghi, A. Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Trans. Ind. Inf.* **2014**, *11*, 812–820. [CrossRef]
23. Elsherif, F.; Chong, E.K.P.; Kim, J. Energy-Efficient Base Station Control Framework for 5G Cellular Networks Based on Markov Decision Process. *IEEE Trans. Veh. Technol.* **2019**, *68*, 9267–9279. [CrossRef]
24. Ernst, D.; Glavic, M.; Wehenkel, L. Power systems stability control: reinforcement learning framework. *IEEE Trans. Power Syst.* **2004**, *19*, 427–435. [CrossRef]
25. Shayeghi, H.; Younesi, A., Adaptive and Online Control of Microgrids Using Multi-agent Reinforcement Learning. In *Microgrid Architectures, Control and Protection Methods*; Mahdavi Tabatabaei, N., Kabalci, E., Bizon, N., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 577–602.
26. Younesi, A.; Shayeghi, H.A. Q-Learning Based Supervisory PID Controller for Damping Frequency Oscillations in a Hybrid Mini/Micro-Grid. *Iran. J. Electr. Electron. Eng.* **2019**, *15*. [CrossRef]
27. Yu, T.; Zhen, W.G. A reinforcement learning approach to power system stabilizer. In Proceedings of the 2009 IEEE Power & Energy Society General Meeting, Calgary, AB Canada, 26–30 July 2009, pp. 1–5.
28. Vlachogiannis, J.G.; Hatziargyriou, N.D. Reinforcement learning for reactive power control. *IEEE Trans. Power Syst.* **2004**, *19*, 1317–1325. [CrossRef]
29. Nanduri, V.; Das, T.K. A reinforcement learning model to assess market power under auction-based energy pricing. *IEEE Trans. Power Syst.* **2007**, *22*, 85–95. [CrossRef]
30. Younesi, A.; Shayeghi, H.; Moradzadeh, M. Application of reinforcement learning for generating optimal control signal to the IPFC for damping of low-frequency oscillations. *Int. Trans. Electr. Energy Syst.* **2018**, *28*, e2488. [CrossRef]
31. Hadidi, R.; Jeyasurya, B. Reinforcement learning based real-time wide-area stabilizing control agents to enhance power system stability. *IEEE Trans. Smart Grid* **2013**, *4*, 489–497. [CrossRef]

32. Meng, L.; Sanseverino, E.R.; Luna, A.; Dragicevic, T.; Vasquez, J.C.; Guerrero, J.M. Microgrid supervisory controllers and energy management systems: A literature review. *Renew. Sustain. Energy Rev.* **2016**, *60*, 1263–1273. [CrossRef]
33. Bobyr, M.V.; Emelyanov, S.G. A nonlinear method of learning neuro-fuzzy models for dynamic control systems. *Appl. Soft Comput.* **2020**, *88*, 106030. [CrossRef]
34. Craven, M.P.; Curtis, K.M.; Hayes-Gill, B.H.; Thursfield, C. A hybrid neural network/rule-based technique for on-line gesture and hand-written character recognition. In Proceedings of the Fourth IEEE International Conference on Electronics, Circuits and Systems, Cairo, Egypt, 15–18 December 2020.
35. Gudyś, A.; Sikora, M.; Wróbel, Ł. RuleKit: A comprehensive suite for rule-based learning. *Knowl.-Based Syst.* **2020**, 105480. [CrossRef]
36. Jafari, M.; Malekjamshidi, Z. Optimal energy management of a residential-based hybrid renewable energy system using rule-based real-time control and 2D dynamic programming optimization method. *Renew. Energy* **2020**, *146*, 254–266. [CrossRef]
37. Das, P.; Choudhary, R.; Sanyal, A. *Review Report on Multi-Agent System Control Analysis for Smart Grid System*; SSRN 3517356; SSRN: Rochester, NY, USA, 2020.
38. Salgueiro, Y.; Rivera, M.; Nápoles, G. Multi-agent-Based Decision Support Systems in Smart Microgrids. In *Intelligent Decision Technologies 2019*; Springer: London, UK, 2020; pp. 123–132.
39. Shi, H.; Li, X.; Hwang, K.S.; Pan, W.; Xu, G. Decoupled visual servoing with fuzzyQ-learning. *IEEE Trans. Ind. Inf.* **2016**, *14*, 241–252. [CrossRef]
40. Lu, R.; Hong, S.H.; Yu, M. Demand Response for Home Energy Management Using Reinforcement Learning and Artificial Neural Network. *IEEE Trans. Smart Grid* **2019**, *10*, 6629–6639. [CrossRef]
41. Ruan, A.; Shi, A.; Qin, L.; Xu, S.; Zhao, Y. A Reinforcement Learning Based Markov-Decision Process (MDP) Implementation for SRAM FPGAs. *IEEE Trans. Circuits Syst. II: Express Briefs* **2019**. [CrossRef]
42. Wu, J.; Fang, B.; Fang, J.; Chen, X.; Chi, K.T. Sequential topology recovery of complex power systems based on reinforcement learning. *Phys. A: Stat. Mech. Its Appl.* **2019**, *535*, 122487. [CrossRef]
43. Busoniu, L.; Babuska, R.; De Schutter, B.; Ernst, D. *Reinforcement Learning and Dynamic Programming Using Function Approximators*; CRC Press: Boca Raton, FL, USA, 2017.
44. Song, H.; Liu, C..; Lawarree, J.; Dahlgren, R.W. Optimal electricity supply bidding by Markov decision process. *IEEE Trans. Power Syst.* **2000**, *15*, 618–624. [CrossRef]
45. Xiong, R.; Cao, J.; Yu, Q. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. *Appl. Energy* **2018**, *211*, 538–548. [CrossRef]
46. Weber, C.; Elshaw, M.; Mayer, N.M. *Reinforcement Learning*; BoD–Books on Demand: Norderstedt, Germany, 2008.
47. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [CrossRef]
48. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
49. Watkins, C.J.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
50. Bellman, R. Dynamic programming. *Science* **1966**, *153*, 34–37. [CrossRef]
51. Ernst, D. Near optimal closed-loop control. Application to electric power systems. Ph.D. Thesis. University of Liège, Liège, Belgium, 2003
52. Nowé, A.; Vrancx, P.; De Hauwere, Y.M. Game theory and multi-agent reinforcement learning. In *Reinforcement Learning*; Springer: London, UK, 2012; pp. 441–470.
53. Egido, I.; Fernandez-Bernal, F.; Centeno, P.; Rouco, L. Maximum frequency deviation calculation in small isolated power systems. *IEEE Trans. Power Syst.* **2009**, *24*, 1731–1738. [CrossRef]
54. Gupta, P.; Bhatia, R.; Jain, D. Average absolute frequency deviation value based active islanding detection technique. *IEEE Trans. Smart Grid* **2014**, *6*, 26–35. [CrossRef]
55. Shayeghi, H.; Younesi, A. 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.* **2017**, *13*, 385–398.
56. Mahat, P.; Chen, Z.; Bak-Jensen, B. Control and operation of distributed generation in distribution systems. *Electr. Power Syst. Res.* **2011**, *81*, 495–502. [CrossRef]
57. Saheb-Koussa, D.; Haddadi, M.; Belhamel, M. Modeling and simulation of windgenerator with fixed speed wind turbine under Matlab-Simulink. *Energy Procedia* **2012**, *18*, 701–708. [CrossRef]

58. Mondal, D.; Chakrabarti, A.; Sengupta, A. *Power System Small Signal Stability Analysis and Control*; Academic Press: Cambridge, MA, USA, 2014.
59. Shayeghi, H.; Younesi, A.; Hashemi, Y. Optimal design of a robust discrete parallel FP+ FI+ FD controller for the automatic voltage regulator system. *Int. J. Electr. Power Energy Syst.* **2015**, *67*, 66–75. [CrossRef]
60. Mirjalili, S.; Gandomi, A.H.; Mirjalili, S.Z.; Saremi, S.; Faris, H.; Mirjalili, S.M. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. *Adv. Eng. Softw.* **2017**, *114*, 163–191. [CrossRef]