Secondary Frequency Control of Microgrids
An Online Reinforcement Learning Approach

Adibi, Mahya; Van Der Woude, Jacob

DOI
10.1109/TAC.2022.3162550

Publication date
2022

Document Version
Final published version

Published in
IEEE Transactions on Automatic Control

Citation (APA)
Adibi, M., & Van Der Woude, J. (2022). Secondary Frequency Control of Microgrids: An Online Reinforcement Learning Approach. *IEEE Transactions on Automatic Control*, *67*(9), 4824-4831. https://doi.org/10.1109/TAC.2022.3162550

Secondary Frequency Control of Microgrids: An Online Reinforcement Learning Approach

Mahya Adibi and Jacob van der Woude

***Abstract*—In this article, we present a reinforcement learning-based scheme for secondary frequency control of lossy inverter-based microgrids. Compared with the existing methods in the literature, we relax the common restrictions on the system, i.e., being lossless, and the transmission lines and loads to have known constant impedances. The proposed secondary frequency control scheme does not require *a priori* information about system parameters and can achieve frequency synchronization within an ultimate bound in the presence of dominantly resistive and/or inductive line and load impedances, model parameter uncertainties, and time varying loads and disturbances. First, using Lyapunov theory, a feedback control is formulated based on the unknown dynamics of the microgrid. Next, a performance function is defined based on cumulative costs toward achieving convergence to the nominal frequency. The performance function is approximated by a critic neural network in real-time. An actor network is then simultaneously learning a parameterized approximation of the nonlinear dynamics and optimizing the approximated performance function obtained from the critic network. Furthermore, using the Lyapunov approach, the uniformly ultimate boundedness of the closed-loop frequency error dynamics and the networks' weight estimation errors are shown.**

***Index Terms*—Microgrids, neural network, reinforcement learning, secondary frequency control.**

I. Introduction

The microgrid concept has been identified as a solution to facilitate the integration of large shares of renewable distributed generation (DG) units to the power networks [1], [2]. Most renewable units are connected to low- and medium-voltage distribution networks via inverters. The physical characteristics of these inverters significantly differ from the characteristics of synchronous generators. Hence, different control techniques are required to guarantee the stability of the system frequency in case of an imbalance between the generated power and the demand in the network [3]. To stabilize the system, primary droop controllers are widely employed. However, steady-state deviations from the nominal frequency are observed. Therefore, a secondary control layer must be implemented to achieve the ultimate frequency synchronization and power sharing; see [4].

A conventional approach to deal with the frequency synchronization problem consists of using a primary droop controller and a secondary proportional-integral control scheme. However, the performance and robustness of secondary microgrid controllers, when the system is driven by measurement noise and disturbances, is an open concern. In [5], a secure secondary controller for inverter-based distributed energy resources in ac microgrids is proposed. However, the existing work is limited to constant impedance loads. Secondary frequency controllers for steady-state network frequency restoration and power sharing in the presence of uncertainties (clock drifts) are derived in [6]. In [7], a modified adaptive droop controller is proposed in which the frequency restoration controller works tightly with the primary controller with short time constants. However, both works [6] and [7] are limited to lossless network scenarios. In [8], semidecentralized frequency synchronization schemes are presented without taking the transmission losses into account. Moreover, to achieve frequency and voltage regulation, microgrid controllers are designed in [9] based on network-reduced models of microgrids. However, such networks do not provide an explicit characterization of the loads. Hence, the controllers are not robust to load variations and model parametric uncertainties. A convex optimization scheme is proposed in [10] for smooth control of microgrids in the transitions between the grid-connected and islanded modes, with load curtailment as the key tuning knob. In [11], a sliding-mode controller is developed for the case of lossless microgrids and with the assumption of constant disturbances. In [12], an integral frequency control scheme, robust to disturbances, is proposed. However, similar to [11], the power network is assumed to be purely inductive (lossless). While this assumption is reasonable on the transmission level, it does not hold in general for a microgrid on the medium or low voltage levels.

In this article, we propose an actor-critic-based reinforcement learning approach for secondary frequency control of microgrids. Our adaptive secondary frequency controller acts on top of the local droop control level and handles lossy microgrids and does not rely on *a priori* known dynamics of the system. The adaptive actor-critic control scheme presented here compensates for the uncertain dynamics of DG units, parameter changes (for example, due to aging or thermal effects), disturbances and time-varying loads, as well as eliminating the steady-state error caused by the individual and isolated droop controllers.

The proposed reinforcement learning approach appropriately responds to changes in the system operating conditions and adjusts the control parameters in real-time. For the frequency regulation problem, a long-term performance function is defined based on instantaneous costs, but since the dynamics are unknown, we define a critic network to learn this performance function in real-time. Furthermore, an actor network aims at deriving an optimal control policy by approximating the unknown nonlinear dynamics and minimizing the learned performance function obtained from the critic network. We presented our preliminary results in [13]. Compared with [13], here, we further provide the proof of convergence of the learning algorithms and the sufficient conditions to guarantee the uniformly ultimately boundedness of the closed-loop frequency error system. Details of our proposed control design are provided in the following sections.

The rest of this article is organized as follows. Section II describes the model of the lossy microgrid and formulates the frequency control problem alongwith the closed-loop stability of the frequency error dynamics. Next, we present our proposed learning algorithm based on coupled critic and actor networks in Section III. We further provide the sufficient conditions on the critic and actor learning rates to guarantee

Manuscript received 21 June 2021; revised 6 January 2022; accepted 10 March 2022. Date of publication 25 March 2022; date of current version 30 August 2022. This work was supported by the NWO (Netherlands Organization for Scientific Research) program *Uncertainty Reduction in Smart Energy Systems* (URSES) under the auspices of the project ENBARK. Recommended by Associate Editor J. Lavaei. *(Corresponding author: Mahya Adibi.)*

The authors are with the Delft Institute of Applied Mathematics, Delft University of Technology, 2628, XE Delft, The Netherlands (e-mail: m.adibi@tudelft.nl; j.w.vanderwoude@tudelft.nl).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2022.3162550.

Digital Object Identifier 10.1109/TAC.2022.3162550

the convergence of the learning update rules (the proof is presented in the Appendix). Simulation results are discussed in Section IV. Finally Section V concludes this article.

II. Problem Statement and Theoretical Foundations

We consider an inverter-based microgrid modeled as a graph $\mathcal{G} = (N, E)$ with $N = \{1, 2, \ldots, n\}$ the set of nodes (generation buses with inverters as their grid interface) and $E \subseteq N \times N$ the set of undirected edges (network lines). We consider a model of a lossy microgrid in which two nodes $\{i, j\} \in E$ are connected by a complex nonzero admittance $Y_{ij} = G_{ij} - iB_{ij} \in \mathbb{C}$ with conductance $G_{ij} \in \mathbb{R}^+$ and susceptance $B_{ij} \in \mathbb{R}^+$ [14], [15]. Let $N_i = \{j \in N \mid j \neq i,\ \{i, j\} \in E\}$ denote the neighbor set of node i. We assign a time-dependent voltage phase angle $\delta_i \in \mathbb{T} := [0, 2\pi)$ and a voltage amplitude $V_i \in \mathbb{R}_{\geq 0}$ to each node i. The relative voltage phase angles are denoted by $\delta_{ij} := \delta_i - \delta_j$, $\{i, j\} \in E$.

A. Microgrid Nonlinear Dynamical Model

We consider a microgrid model with discrete-time dynamics consisting of inverter-interfaced sources. The inverters are enhanced by standard primary voltage and frequency droop controllers as in [16]. We can formulate the system dynamics in the following form:

$$x_1(k+1) = f_1(x(k)) \tag{1}$$
$$x_2(k+1) = f_2(x(k)) + u(k) \tag{2}$$
$$x_3(k+1) = f_3(x(k)) \tag{3}$$

where

$$x_1(k) := [\delta_1(k), \delta_2(k), \ldots, \delta_n(k)]^{\mathrm{T}} \in \mathbb{T}^n \tag{4}$$
$$x_2(k) := [\omega_1(k), \omega_2(k), \ldots, \omega_n(k)]^{\mathrm{T}} \in \mathbb{R}^n \tag{5}$$
$$x_3(k) := [V_1(k), V_2(k), \ldots, V_n(k)]^{\mathrm{T}} \in \mathbb{R}^n \tag{6}$$
$$u(k) := [u_1(k), u_2(k), \ldots, u_n(k)]^{\mathrm{T}} \in \mathbb{R}^n \tag{7}$$
$$x(k) := [x_1^{\mathrm{T}}(k), x_2^{\mathrm{T}}(k), x_3^{\mathrm{T}}(k)]^{\mathrm{T}} \in \mathbb{R}^{3n}. \tag{8}$$

Here, $\omega_i \in \mathbb{R}$ is the inverter frequency corresponding to node i and u is the secondary frequency control input to be designed later in Section II-B. Note that the functions f_2 and f_3 include the local primary frequency and voltage droop control dynamics. Furthermore, $f_1 = \mathrm{mod}_{2\pi}\{\delta(k) + k\omega(k)\}$. We assume that the nonlinear dynamics of the DG units with their local primary droop controllers, i.e., functions $f_2(x(k))$ and $f_3(x(k))$ are *unknown*. The goal is to design a secondary frequency control scheme to compensate for frequency deviations while being robust against parametric uncertainties caused by the unknown dynamics and disturbances affecting the network.

Remark 1: Note that the aim of this article is to design a secondary frequency control scheme to guarantee frequency regulation and this is later on proved. However, although the overall system is equipped with primary *voltage* droop controllers and we will demonstrate that the voltage is stabilized under our secondary control scheme, this needs to be explicitly proved (for a lossy network) and/or a similar secondary voltage control algorithm needs to be designed. This is included in our future work.

Before starting with the control design procedure, we present the following definition that is required for stability analysis of the frequency error system which will be defined in Section II-B.

Definition 1: Consider the general nonlinear system $x(k+1) = f(x(k), k) + d(k)$ with $d(k)$ an unknown but bounded disturbance. If there exists a function $\mathcal{V}(x, k)$ with continuous partial differences, such that for x in a compact set $S \subset \mathbb{R}^n$

1) $\mathcal{V}(x(k), k)$ is positive definite, $\mathcal{V}(x(k), k) > 0$.
2) $\dot{\mathcal{V}}(x(k), k) < 0$ for $\|x\| > \beta$

for some $\beta > 0$ such that the ball of radius β is contained in S, and then the system is uniformly ultimately bounded and the norm of the state is bounded within a neighborhood of β ([17], [18] ch. 2.3.1).

Based on this definition, the stability of the dynamical frequency error system can be investigated by choosing an appropriate function $\mathcal{V}$. We will use Definition 1 to prove the stability of the closed-loop frequency regulation dynamics and the boundedness of the parameter estimation errors. In the next section, the regulation error dynamics and the structure of the control input are defined which are the bases for our adaptive learning-based control design in Section III.

B. Regulation Error Dynamics and Control Policy

Consider system dynamics (1)–(3) with unknown nonlinear functions $f_2(x(k))$ and $f_3(x(k))$, and the control input $u(k)$ to be designed. Let us define the nominal frequency of the system as $\omega^\star \in \mathbb{R}^+$ and the vector of the desired frequency signals as $x_2^\star := \omega^\star \mathbb{1}_n \in \mathbb{R}^n$. It is assumed that for the (lossy) system (1)–(3) with integrated local droop controllers, there exists an isolated frequency-synchronized solution (see [19, Assumption 2] and [20, Assumption 4.3]), which can be different from the nominal frequency. It has been shown that there is a steady-state error and deviation from the nominal frequency even for a lossless network. The secondary control objective is to compensate the deviation of frequency signals (5) from their nominal value $\omega^\star$ and make frequencies converge to the desired signal $x_2^\star$.

To accomplish this, we define the regulation error $e(k) \in \mathbb{R}^n$ as

$$e(k) = x_2^\star - x_2(k) \tag{9}$$

which results in the error dynamics

$$\begin{aligned} e(k+1) &= x_2^\star - x_2(k+1) \\ &= x_2^\star - f_2(x(k)) - u(k). \end{aligned} \tag{10}$$

To design $u(k)$ such that (10) is stabilized, we define the candidate Lyapunov function as

$$\mathcal{V}_e(k) = e^{\mathrm{T}}(k)e(k). \tag{11}$$

Taking the difference $\Delta\mathcal{V}_e(k)$ results in

$$\Delta\mathcal{V}_e(k) = e^{\mathrm{T}}(k+1)e(k+1) - e^{\mathrm{T}}(k)e(k). \tag{12}$$

Substituting the error dynamics (10) in (12), we obtain

$$\begin{aligned} \Delta\mathcal{V}_e(k) = {} & (x_2^\star - f_2(x(k)) - u(k))^{\mathrm{T}} \\ & \times (x_2^\star - f_2(x(k)) - u(k)) - e^{\mathrm{T}}(k)e(k). \end{aligned} \tag{13}$$

In order to have $\Delta\mathcal{V}_e(k) < 0$, we select the control input as

$$u(k) = x_2^\star - f_2(x(k)) + Ce(k) \tag{14}$$

where $C \in \mathbb{R}^{n \times n}$ is a constant diagonal positive definite matrix. If we assume $f_2(x(k))$ is known, substituting (14) in (13) yields

$$\Delta\mathcal{V}_e(k) = \sum_{i=1}^{n} (c_i^2 - 1)e_i^2 \tag{15}$$

where e_i is the ith element of $e(k)$ and c_i is the ith eigenvalue of the diagonal matrix C for $i \in N$. Hence, $\Delta\mathcal{V}_e(k) < 0$ and the error system (10) is asymptotically stable if

$$0 \leq c^{\max} < 1 \tag{16}$$

where $c^{\max} \in \mathbb{R}$ is the maximum eigenvalue of C.

However, the dynamics $f_2(x(k))$ is not known. Instead, we use the estimation of the function $f_2(x(k))$, i.e., $\hat{f}_2(x(k))$ ($\hat{f}_2(x(k))$ is approximated using the actor network and will be discussed in Section III-B). We design the control input (14) as follows:

$$u(k) = x_2^\star - \hat{f}_2(x(k)) + Ce(k) \tag{17}$$

which results in

$$\begin{aligned} \Delta\mathcal{V}_e(k) = {} & \left(\tilde{f}_2(x(k)) - Ce(k)\right)^{\mathrm{T}} \left(\tilde{f}_2(x(k)) - Ce(k)\right) \\ & - e^{\mathrm{T}}(k)e(k) \end{aligned} \tag{18}$$

where $\tilde{f}_2(x(k)) = \hat{f}_2(x(k)) - f_2(x(k))$ is the function estimation error. Therefore, $\Delta\mathcal{V}_e(k) < 0$ if

$$\left\|\tilde{f}_2(x(k)) - Ce(k)\right\| < \|e(k)\|. \tag{19}$$

Let the known value $f_2^{\max} \in \mathbb{R}^+$ be the upper bound of the function estimation error $\tilde{f}_2(x(k))$, such that $\|\tilde{f}_2(x(k))\| \leq f_2^{\max}$. Then

$$\begin{aligned}\left\|\tilde{f}_2(x(k)) - Ce(k)\right\| &\leq \left\|\tilde{f}_2(x(k))\right\| + \|Ce(k)\| \\ &\leq f_2^{\max} + c^{\max}\|e(k)\|.\end{aligned} \tag{20}$$

Considering (19) and (20), the error dynamics is stable if

$$f_2^{\max} + c^{\max}\|e(k)\| < \|e(k)\|. \tag{21}$$

Defining $e^{\max} := \frac{f_2^{\max}}{1-c^{\max}}$, it follows that

$$\Delta\mathcal{V}_e(k) < 0 \quad \forall \|e(k)\| > e^{\max}. \tag{22}$$

In other words, $\Delta\mathcal{V}_e(k)$ is negative outside of the compact set $S_e := \{\|e(k)\| \leq e^{\max}\}$, or equivalently, all the solutions that start outside of S_e will enter this set within a finite time and will remain inside the set forever. This means that

$$\|e(k)\| < \frac{f_2^{\max}}{1-c^{\max}}. \tag{23}$$

Based on Definition 1, the estimation errors are bounded from above with the ultimate bound $e^{\max}$.

III. Actor-Critic Learning Algorithm

We consider a neural network with one hidden layer for both actor and critic networks. In order to measure the long-term performance of the system, the cost function $J(k) \in \mathbb{R}^n$ is defined using the instantaneous cost [18] as

$$\begin{aligned}J(k) &= \sum_{m=k}^{\infty} \gamma^{m-k} r(m+1) \\ &= r(k+1) + \gamma r(k+2) + \gamma^2 r(k+3) + \cdots\end{aligned} \tag{24}$$

where $0 < \gamma < 1$ is the discount factor and $r(k) = [r_1(k), r_2(k), \ldots, r_n(k)]^{\mathrm{T}} \in \mathbb{R}^n$ is the vector of instantaneous costs (reinforcement learning signals) as follows (see [21]):

$$r_i(k) = \begin{cases} 0 & \text{if } \left|e_i(k)\right| \leq \mu \\ 1 & \text{if } \left|e_i(k)\right| > \mu \end{cases} \tag{25}$$

for $i \in N$ and $\mu \in \mathbb{R}^+$ is a fixed threshold. The instantaneous cost $r_i(k)$ is a measure of the current performance of the ith DG. To be more precise, it quantifies how the control input has performed; $r_i(k) = 0$ indicates a success in the frequency regulation and $r_i(k) = 1$ shows a performance degradation.

Since the dynamics is unknown, we define a critic network to learn the cost function $J(k)$ in real-time in Section III-A.

A. Adaptation of Critic Network

The critic neural network, with output $\hat{J}(k) \in \mathbb{R}^n$, learns to approximate the cost function $J(k) \in \mathbb{R}^n$. The output of the critic neural network can be described in the form

$$\hat{J}(k) = \hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}\left(v_1^{\mathrm{T}} x_2(k)\right) \tag{26}$$

such that $\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k) \in \mathbb{R}^{n\times n_1}$ represents the matrix of weights between the hidden and output layer and $v_1^{\mathrm{T}} \in \mathbb{R}^{n_1\times n}$ represents the matrix of weights between the input and hidden layer. We assume that the matrix of the weights v_1 is fixed and only the weights $\hat{\psi}_{\mathrm{c}}$ are being adapted. This assumption is common in practice and is a technique to accelerate the neural network training and reduce the training time [22]. In case of a poor approximation of the cost function, the parameters of the first layer can also be modified in the optimization process to have a more precise estimation. Moreover, $\phi_{\mathrm{c}}(v_1^{\mathrm{T}} x_2(k)) \in \mathbb{R}^{n_1}$ is the basis function vector in the hidden layer and n_1 is the number of the nodes in the hidden layer. In order to compress the notation, we introduce the shorthand notation $\phi_{\mathrm{c}}(k) = \phi_{\mathrm{c}}(v_1^{\mathrm{T}} x_2(k))$ for the value of the basis function at time instant k.

Let $e_{\mathrm{c}}(k) \in \mathbb{R}^n$ be the prediction error (temporal-difference error; see [23]) of the critic network as

$$\begin{aligned}e_{\mathrm{c}}(k) &= r(k) + \gamma\hat{J}(k) - \hat{J}(k-1) \\ &= r(k) + \gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k) - \hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\end{aligned} \tag{27}$$

and the objective function to be minimized as

$$J_{\mathrm{c}}(k) = \frac{1}{2}e_{\mathrm{c}}^{\mathrm{T}}(k)e_{\mathrm{c}}(k). \tag{28}$$

Applying the gradient descent algorithm for minimizing $J_{\mathrm{c}}(k)$, and hence $e_{\mathrm{c}}(k)$, results in

$$\begin{aligned}\hat{\psi}_{\mathrm{c}}(k+1) &= \hat{\psi}_{\mathrm{c}}(k) - \alpha_{\mathrm{c}}\frac{\partial J_{\mathrm{c}}(k)}{\partial e_{\mathrm{c}}(k)}\frac{\partial e_{\mathrm{c}}(k)}{\partial \hat{J}(k)}\frac{\partial \hat{J}(k)}{\partial \hat{\psi}_{\mathrm{c}}(k)} \\ &= \hat{\psi}_{\mathrm{c}}(k) - \alpha_{\mathrm{c}}\gamma\phi_{\mathrm{c}}(k)e_{\mathrm{c}}^{\mathrm{T}}(k)\end{aligned} \tag{29}$$

which leads to the following weight update rule for the critic network

$$\begin{aligned}&\hat{\psi}_{\mathrm{c}}(k+1) = \hat{\psi}_{\mathrm{c}}(k) - \alpha_{\mathrm{c}}\gamma\phi_{\mathrm{c}}(k)\times \\ &\quad\left(r(k) + \gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k) - \hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right)^{\mathrm{T}}\end{aligned} \tag{30}$$

where $\alpha_{\mathrm{c}} \in \mathbb{R}^+$ is the critic learning rate.

In Section III-B, the actor network is constructed to minimize both the function estimation error $\tilde{f}_2(x(k))$ and the cost function $\hat{J}(k)$.

B. Adaptation of Actor Network

The main goal of the actor network is to generate the approximation of the unknown nonlinear function $f_2(x(k))$ and then plug the estimated $\hat{f}_2(k)$ into the control policy (17). The estimated function is parameterized as

$$\hat{f}_2(k) = \hat{\psi}_{\mathrm{a}}^{\mathrm{T}}(k)\phi_{\mathrm{a}}\left(v_2^{\mathrm{T}} x_2(k)\right) \tag{31}$$

where $\hat{\psi}_{\mathrm{a}}^{\mathrm{T}}(k) \in \mathbb{R}^{n\times n_2}$ represents the matrix of weights between the hidden and output layer and $v_2^{\mathrm{T}} \in \mathbb{R}^{n_2\times n}$ represents the matrix of weights between the input and hidden layer. We assume that the matrix of the weight v_2 is fixed and only the weights $\hat{\psi}_{\mathrm{a}}$ are being adapted. Moreover, $\phi_{\mathrm{a}}(v_2^{\mathrm{T}} x_2(k)) \in \mathbb{R}^{n_2}$ is the basis function vector in the hidden layer and n_2 is the number of the nodes in the hidden layer. Similar to the critic network, we introduce the shorthand notation $\phi_{\mathrm{a}}(k) = \phi_{\mathrm{a}}(v_2^{\mathrm{T}} x_2(k))$ for the value of the basis function at time instant k to compress the notation.

We define the function estimation error $\tilde{f}_2(k) \in \mathbb{R}^n$ as

$$\tilde{f}_2(k) = \hat{f}_2(k) - f_2(k) \tag{32}$$

and the error between the desired cost function $J^\star(k) \in \mathbb{R}^n$ and the critic network output $\hat{J}(k)$ as

$$\tilde{J}(k) = \hat{J}(k) - J^\star(k). \tag{33}$$

The training of the actor network is done using $\tilde{f}_2(k)$ and $\tilde{J}(k)$ and defining the prediction error $e_{\mathrm{a}}(k) \in \mathbb{R}^n$ as

$$e_{\mathrm{a}}(k) = \tilde{f}_2(k) + \tilde{J}(k). \tag{34}$$

According to (24) and (25), the desired value for the function $J^\star(k)$ is 0. Thus, (34) becomes

$$e_{\mathrm{a}}(k) = \tilde{f}_2(k) + \hat{J}(k). \tag{35}$$

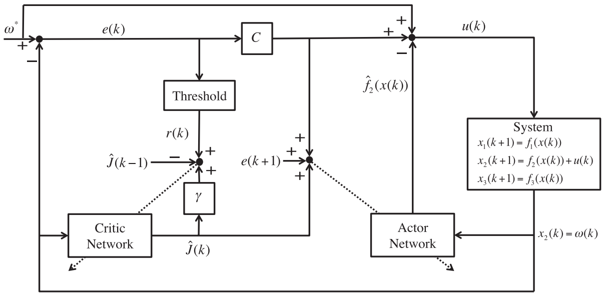


Fig. 1. Schematic overview of the closed-loop system. The dashed lines represent the updating mechanism for the critic and actor networks.

We consider the objective function to be minimized by the actor network in the form

$$J_{\mathrm{a}}(k) = \frac{1}{2} e_{\mathrm{a}}^{\mathrm{T}}(k) e_{\mathrm{a}}(k). \tag{36}$$

Using the gradient descent algorithm for minimizing $J_{\mathrm{a}}(k)$, and subsequently for $e_{\mathrm{a}}(k)$, we obtain

$$\begin{aligned}\hat{\psi}_{\mathrm{a}}(k+1) &= \hat{\psi}_{\mathrm{a}}(k) - \alpha_{\mathrm{a}} \frac{\partial J_{\mathrm{a}}(k)}{\partial e_{\mathrm{a}}(k)} \frac{\partial e_{\mathrm{a}}(k)}{\partial \tilde{f}_2(k)} \frac{\partial \tilde{f}_2(k)}{\partial \hat{\psi}_{\mathrm{a}}(k)} \\ &= \hat{\psi}_{\mathrm{a}}(k) - \alpha_{\mathrm{a}} \phi_{\mathrm{a}}(k) e_{\mathrm{a}}^{\mathrm{T}}(k) \end{aligned} \tag{37}$$

which results in

$$\hat{\psi}_{\mathrm{a}}(k+1) = \hat{\psi}_{\mathrm{a}}(k) - \alpha_{\mathrm{a}} \phi_{\mathrm{a}}(k) (\tilde{f}_2(k) + \hat{J}(k))^{\mathrm{T}} \tag{38}$$

where $\alpha_{\mathrm{a}} \in \mathbb{R}^+$ is the actor learning rate. However, we cannot use the weight update rule (38) in practice. This is due to the fact that the function estimation error $\tilde{f}_2(k)$ defined in (32) consists of the *unknown* nonlinear function $f_2(k)$. This problem can be addressed by substituting (17) in (10), which yields

$$\begin{aligned} e(k+1) &= -f_2(x(k)) + \hat{f}_2(x(k)) - Ce(k) \\ &= \tilde{f}_2(x(k)) - Ce(k). \end{aligned} \tag{39}$$

Hence, the function estimation error becomes

$$\tilde{f}_2(k) = e(k+1) + Ce(k). \tag{40}$$

Substituting (40) in (38) yields the following weight update rule for the actor network

$$\hat{\psi}_{\mathrm{a}}(k+1) = \hat{\psi}_{\mathrm{a}}(k) - \alpha_{\mathrm{a}} \phi_{\mathrm{a}}(k) \left(e(k+1) + Ce(k) + \hat{J}(k) \right)^{\mathrm{T}}. \tag{41}$$

The schematic structure of the reinforcement learning frequency control scheme is shown in Fig. 1. The actor is responsible for estimating the nonlinear dynamics of the system and generating the control input (17) such that it minimizes the cost function $\hat{J}(k)$ (estimated by the critic network). The critic adapts the estimation of the cost function, given $x(k)$, and the frequency regulation error signal. This process is repeated until we reach our control goal. In the following theorem, we present the conditions on the learning rates to guarantee the convergence of the learning algorithms. Under these conditions, the weights $\hat{\psi}_{\mathrm{c}}$ and $\hat{\psi}_{\mathrm{a}}$ of the critic and actor networks converge close to their optimal values $\psi_{\mathrm{c}}^\star$ and $\psi_{\mathrm{a}}^\star$, respectively, for the designed control policy (17). Before proceeding with the theorem, an assumption is presented that is required for the proof of theorem.

Assumption 1: It is assumed that the basis functions ϕ_{c} and ϕ_{a}, the elements of the weight matrix ψ_{c}, and the neural network approximation's error are bounded from above.

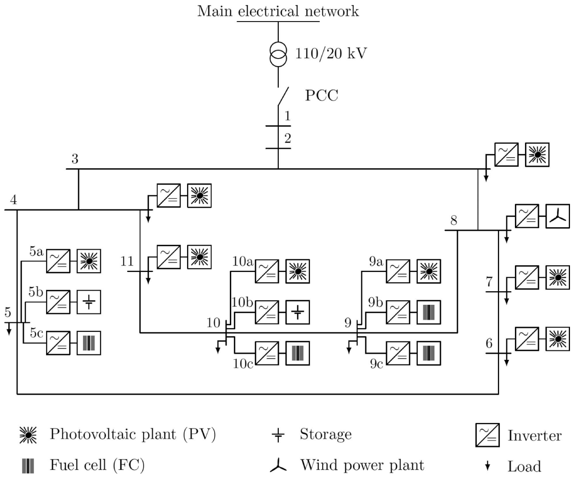


Fig. 2. Benchmark model adapted from [24] with 11 main buses and several inverter-interfaced DG and storage units.

TABLE I
NETWORK PARAMETERS

Base values	$S_{\text{base}} = 4.75$ MVA, $V_{\text{base}} = 20$ kV
$S_i^N, i = 1, \cdots, 6$	[0.505, 0.028, 0.261, 0.179, 0.168, 0.012] p.u.

Theorem 1: Consider the system (1)–(3) along with the control input (17), the critic update rule given by (30), and the actor update rule given by (41). The error between the estimated critic and actor parameter vectors $\hat{\psi}_{\mathrm{c}}$ and $\hat{\psi}_{\mathrm{a}}$ and their optimal values $\psi_{\mathrm{c}}^\star$ and $\psi_{\mathrm{a}}^\star$ converges to and stays within a compact set around zero (i.e., uniformly ultimately bounded stable) as long as the following conditions hold:

$$\alpha_{\mathrm{c}} < \frac{1}{\gamma^2 (\phi_{\mathrm{c}}^{\max})^2} \tag{42}$$

$$\alpha_{\mathrm{a}} < \frac{1}{(\phi_{\mathrm{a}}^{\max})^2} \tag{43}$$

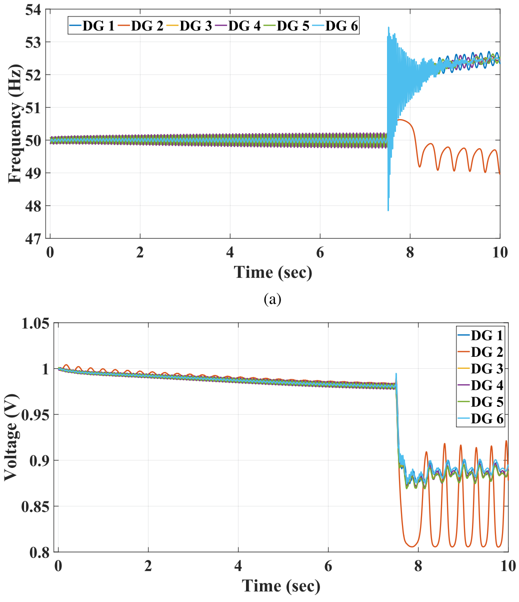
with $\phi_{\mathrm{c}}^{\max}$ and $\phi_{\mathrm{a}}^{\max}$ being the upper bounds of $\phi_{\mathrm{c}}(k)$ and $\phi_{\mathrm{a}}(k)$, respectively.

Proof: The proof is given in Appendix A. ■

In the following section, we validate the performance of the proposed control scheme via simulation on a benchmark microgrid in the presence of disturbances.

IV. Case Study

The effectiveness of our proposed reinforcement learning-based scheme is verified on the three-phase islanded Subnetwork 1 of the CIGRE benchmark medium voltage network as in [16] and [25]. The benchmark microgrid is shown in Fig. 2. The simulation is performed by considering $n = 6$ controllable generation sources at buses 5b ($i = 1$), 5c ($i = 2$), 9b ($i = 3$), 9c ($i = 4$), 10b ($i = 5$), and 10c ($i = 6$). All photovoltaic sources together with the wind turbine at bus 8 are assumed as noncontrollable units and are neglected. It is assumed that all controllable generation units are equipped with droop controllers. To each inverter $i \in N$, its power rating $S_i^N \in \mathbb{R}^+$ is assigned and is given in Table I. The gains and setpoints of the droop controllers are selected as $P_i^* = 0.6 S_i^N$ per unit, $k_{P_i} = 0.2/S_i^N$ Hz/per unit, as well as $Q_i^* = 0.25 S_i^N$ per unit, $k_{Q_i} = 0.1/S_i^N$ per unit/per unit. It is assumed that the batteries at nodes 5b and 10b are operated in charging mode, hence functioning as loads. Therefore, $P_i^* = -0.6 S_i^N$ for $i = 1, 5$. The loads at nodes 3–11 are specified in Table I of [25]. The load at node 1 is



(a)

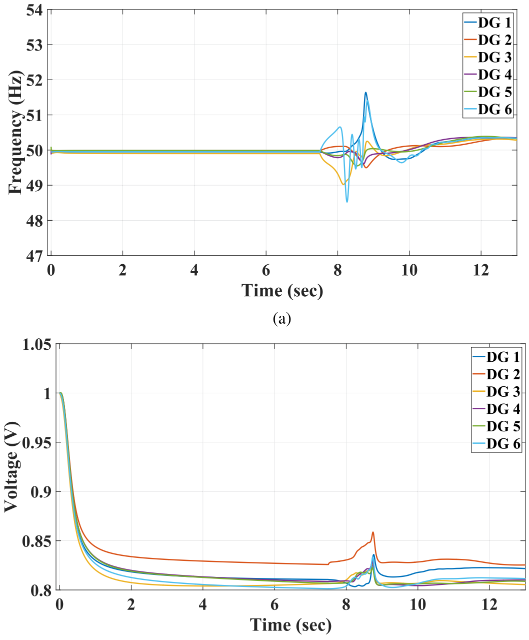
(b)

Fig. 3. Lossy microgrid case. Time evolution of (a) frequency and (b) voltage dynamics using only the primary droop controllers.

neglected. The line parameters are given in Table 3 of [25]. The nominal frequency and the sampling time are taken as $f^\star = 50$ Hz and $T = 1$ ms, respectively. The frequency signals $(x_2(k) := [\omega_1(k), \ldots, \omega_n(k)])$ of the units are measured and fed to the secondary control scheme. The elements of the diagonal gain matrix C is selected as $c_i = 0.1$ for $i \in N$. The threshold value μ is set to 0.02. We consider one hidden layer for both critic and actor neural networks, and we assume that each hidden layer contains 10 nodes, i.e., $n_1 = n_2 = 10$. For weight updating rules, the learning rates are selected as $\alpha_{\mathrm{c}} = 0.1$ and $\alpha_{\mathrm{a}} = 0.1$, and the discount factor is set as $\gamma = 0.5$. All the weight parameters of the matrices v_1 and v_2, between the input and hidden layer, are fixed at 1, as explained in Section III-A. For our case study, this choice yielded high level of performance and meanwhile reduced the order of complexity of optimizing the network. Next, the initial values for the adapting weights $\hat{\psi}_{\mathrm{c}}$ and $\hat{\psi}_{\mathrm{a}}$ are selected randomly (with uniform distribution) between 0 and 1. Furthermore, the activation functions are selected as hyperbolic tangent functions. Hence, the maximum of the activation functions $\phi_c(x)$ and $\phi_a(x)$ is 1. Consequently, conditions (42) and (43) imply $\alpha_{\mathrm{c}} < 4$ and $\alpha_{\mathrm{a}} < 1$, in order to have the frequency error and the neural networks' weight estimates uniformly ultimately bounded.

In this case study, we show the effectiveness of our adaptive control scheme under load variations. The initial voltage amplitude is selected as 1 per unit for all units. The initial frequency variables are selected as randomly distributed around 50 Hz with standard deviation of 0.1 Hz. The initial phase variables are selected at zero degree. The microgrid is assumed to be in the islanded mode.

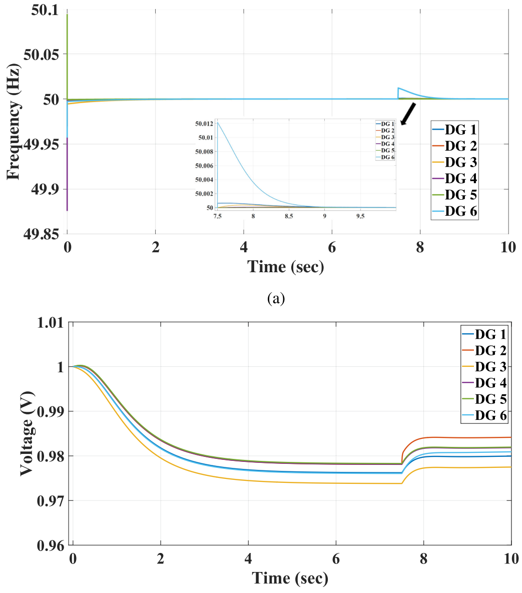
The trajectories of the frequencies $\frac{\omega_i}{2\pi}$ in Hz and the voltage amplitudes in per unit form for $i = 1, \ldots, 6$ of the controllable sources in the local droop control only case are shown in Fig. 3. As can be observed, using only the primary droop controller results in steady-state error in the frequency and voltage signals. Moreover, at $t = 7.5$ s, the values of the parameters $B_{24} = 178.3177$ and $G_{16} = 463.2297$ (nominal admittance and conductance values from [25]) are increased by 50%. The parameter changes that we impose are representative of cases where the transmission parameters vary and/or the connected resistive/inductive load alters. Note that in our simulation model of the microgrid, the load effect is absorbed in the transmission line model [16]. As a result of this, the frequency and voltage signals start oscillating and will have large errors with respect to the desired nominal values. To further provide evidence that control and stabilizing a lossy network is a big challenge and hard to be fulfilled without an adaptive and *online* learning-based control scheme, we have further simulated the distributed secondary control scheme from [26]. As can be seen in Fig. 4, the secondary distributed algorithm is able to stabilize frequency in the first period (although with a steady-state error that is larger than our proposed



(a)

(b)

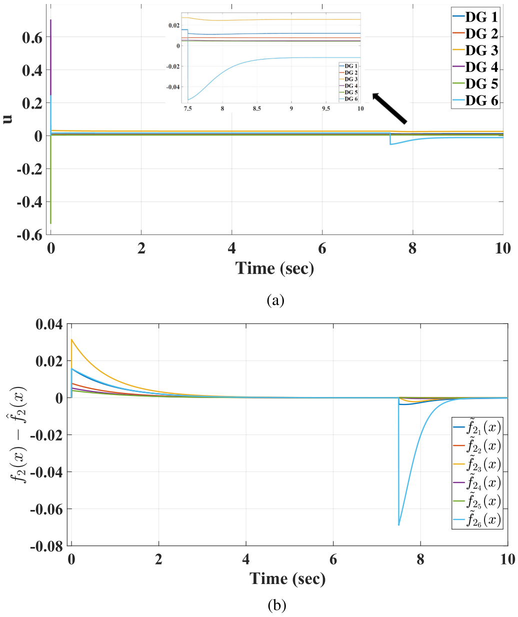
Fig. 4. Lossy microgrid case. Time evolution of (a) frequency and (b) voltage dynamics using the primary droop controllers plus the secondary control scheme from [26].



(a)

(b)

Fig. 5. Lossy microgrid case. Time evolution of (a) frequency and (b) voltage dynamics using primary droop controllers plus our proposed secondary *RL*-based control scheme.



(a)

(b)

Fig. 6. Lossy microgrid case. Time evolution of (a) *RL* secondary control input and (b) error in estimating f_2.

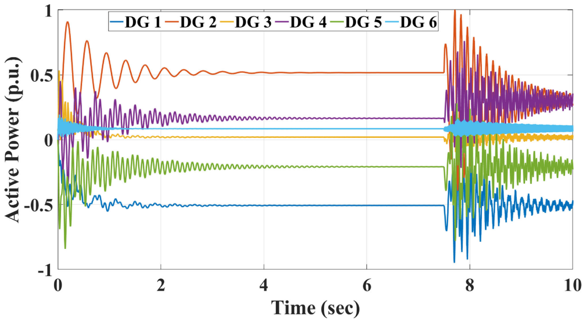


Fig. 7. Time evolution of active power outputs for our proposed scheme.

scheme and that is due to the lossy nature of network and the fact that the assumptions made in design of [26] are based on a lossless network). However, when the parameters of network change, similar to the primary droop control only case, the closed-loop system becomes unstable and we observe oscillations and large errors in tracking the nominal frequency and voltages. The severity is less than the droop control only case but it is not acceptable from the network operation and reliability aspects. On the other hand, if we deploy our proposed *RL*-based secondary control scheme (on top of the primary local droop controllers), as shown in Fig. 5, the microgrid will be entirely stable and even after the change in the dynamics, the secondary controller adapts the control input fast so that there is a very small jump in the state variables and after very short time, the frequency signals of the DG units converge to the nominal value of 50 Hz. The control input u and the error in estimating the unknown dynamics f_2 are illustrated in Fig. 6. In Fig. 7, active power signals are depicted. The power levels are below the nominal ratings and the proposed scheme does not require the generation units to inject unrealistic high instantaneous power to the grid.

V. Conclusion and Future Research

A reinforcement learning control scheme has been proposed for secondary frequency synchronization of lossy microgrids. Our method is able to efficiently handle general cases of resistive and inductive line and load impedances, parameter uncertainties, time varying loads, and disturbances. Using this adaptive control approach, no *priori* knowledge about the system dynamics is required. As next steps, we will extend our approach to address the secondary voltage control and reactive power sharing problem. Moreover, experimental validations of our proposed methods will be carried out as well.

Appendix A
Proof of Theorem 1

We begin the proof by defining the Lyapunov function candidate

$$\mathcal{V}(k) = \underbrace{\frac{1}{\alpha_{\mathrm{c}}}\mathrm{tr}[\tilde{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\tilde{\psi}_{\mathrm{c}}(k)]}_{\mathcal{V}_1(k)} + \underbrace{\frac{1}{\alpha\alpha_{\mathrm{a}}}\mathrm{tr}[\tilde{\psi}_{\mathrm{a}}^{\mathrm{T}}(k)\tilde{\psi}_{\mathrm{a}}(k)]}_{\mathcal{V}_2(k)} \tag{44}$$

where

$$\tilde{\psi}_{\mathrm{c}}(k) = \hat{\psi}_{\mathrm{c}}(k) - \psi_{\mathrm{c}}^{\star} \tag{45}$$

$$\tilde{\psi}_{\mathrm{a}}(k) = \hat{\psi}_{\mathrm{a}}(k) - \psi_{\mathrm{a}}^{\star} \tag{46}$$

and $\alpha > 0$ is constant. The first difference of $\mathcal{V}_1(k)$ is expressed by

$$\begin{aligned}\Delta\mathcal{V}_1(k) &= \mathcal{V}_1(k+1) - \mathcal{V}_1(k)\\ &= \frac{1}{\alpha_{\mathrm{c}}}\mathrm{tr}[\tilde{\psi}_{\mathrm{c}}^{\mathrm{T}}(k+1)\tilde{\psi}_{\mathrm{c}}(k+1) - \tilde{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\tilde{\psi}_{\mathrm{c}}(k)].\end{aligned} \tag{47}$$

Using (30) and noting that $\psi_{\mathrm{c}}^{\star}$ does not depend on k, we obtain

$$\begin{aligned}\tilde{\psi}_{\mathrm{c}}(k+1) &= \hat{\psi}_{\mathrm{c}}(k+1) - \psi_{\mathrm{c}}^{\star} = \hat{\psi}_{\mathrm{c}}(k) - \gamma\alpha_{\mathrm{c}}\phi_{\mathrm{c}}(k)e_{\mathrm{c}}^{\mathrm{T}}(k) - \psi_{\mathrm{c}}^{\star}\\ &= \tilde{\psi}_{\mathrm{c}}(k) - \alpha_{\mathrm{c}}\gamma\phi_{\mathrm{c}}(k)\times\\ &\qquad \Big(r(k) + \gamma\hat{\psi}_{\mathrm{c}}(k)\phi_{\mathrm{c}}(k) - \hat{\psi}_{\mathrm{c}}(k-1)\phi_{\mathrm{c}}(k-1)\Big)^{\mathrm{T}}.\end{aligned} \tag{48}$$

Based on the last expression, we can expand the multiplication term $\tilde{\psi}_{\mathrm{c}}^{\mathrm{T}}(k+1)\tilde{\psi}_{\mathrm{c}}(k+1)$ in the following way:

$$\begin{aligned}\tilde{\psi}_{\mathrm{c}}^{\mathrm{T}}(k+1)\tilde{\psi}_{\mathrm{c}}(k+1) &= \tilde{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\tilde{\psi}_{\mathrm{c}}(k) + \gamma^2\alpha_{\mathrm{c}}^2\left\|\phi_{\mathrm{c}}(k)\right\|^2\times\\ &(r(k) + \gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k) - \hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1))\times\\ &(r(k) + \gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k) - \hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1))^{\mathrm{T}}\\ &- 2\gamma\alpha_{\mathrm{c}}\Psi_{\mathrm{c}}(k)\times\\ &\Big(r(k) + \gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k) - \hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\Big)^{\mathrm{T}}\end{aligned} \tag{49}$$

where

$$\Psi_{\mathrm{c}}(k) = \tilde{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k) \tag{50}$$

is the approximation error of the critic network output. Utilizing the perfect square trinomial $(a-b)^2 = a^2 - 2ab + b^2$, we have

$$\begin{aligned}&\mathrm{tr}\left[-2\alpha_c\gamma\Psi_{\mathrm{c}}(k)\Big(r(k) + \gamma\hat{\psi}_{\mathrm{c}}(k)\phi_{\mathrm{c}}(k) - \hat{\psi}_{\mathrm{c}}(k-1)\phi_{\mathrm{c}}(k-1)\Big)^{\mathrm{T}}\right]\\ &= \alpha_c\left\|r(k) + \gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k) - \hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1) - \gamma\Psi_{\mathrm{c}}(k)\right\|^2\\ &\quad - \alpha_c\left\|r(k) + \gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k) - \hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2\\ &\quad - \alpha_c\gamma^2\left\|\Psi_{\mathrm{c}}(k)\right\|^2.\end{aligned} \tag{51}$$

Rewriting the first term in the abovementioned expression as

$$\begin{aligned}&\left\|r(k)+\gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)-\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)-\gamma\Psi_{\mathrm{c}}(k)\right\|^2\\&=\Big\|r(k)+\gamma(\hat{\psi}_{\mathrm{c}}(k)-\psi_{\mathrm{c}}^{\star})^{\mathrm{T}}\phi_{\mathrm{c}}(k)+\gamma\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\cdots\\&\quad\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)-\gamma\Psi_{\mathrm{c}}(k)^2\Big\|\\&=\left\|r(k)+\gamma\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2. \qquad (52)\end{aligned}$$

Considering (47) together with (49), (51), and (52), we obtain

$$\begin{aligned}\Delta\mathcal{V}_1(k)=&-(1-\alpha_{\mathrm{c}}\gamma^2\left\|\phi_{\mathrm{c}}(k)\right\|^2)\times\\&\left\|r(k)+\gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)-\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2\\&+\left\|r(k)+\gamma\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2\\&-\gamma^2\left\|\Psi_{\mathrm{c}}(k)\right\|^2. \qquad (53)\end{aligned}$$

Moreover, we can formulate $\Delta\mathcal{V}_2(k)$ as

$$\begin{aligned}\Delta\mathcal{V}_2(k)&=\mathcal{V}_2(k+1)-\mathcal{V}_2(k)\\&=\frac{1}{\alpha\alpha_{\mathrm{a}}}\mathrm{tr}[\tilde{\psi}_{\mathrm{a}}^{\mathrm{T}}(k+1)\tilde{\psi}_{\mathrm{a}}(k+1)-\tilde{\psi}_{\mathrm{a}}^{\mathrm{T}}(k)\tilde{\psi}_{\mathrm{a}}(k)]. \qquad (54)\end{aligned}$$

Suppose that the unknown optimal weight of the output layer, for the actor network, is $\psi_{\mathrm{a}}^{\star}$. Then, we have

$$f_2(k)=\psi_{\mathrm{a}}^{\star\mathrm{T}}(k)\phi_{\mathrm{a}}(k)+\epsilon_2(x(k)) \qquad (55)$$

with $\epsilon_2(x(k))\in\mathbb{R}^n$ being the neural network approximation error. Using (31) and (55), we determine the estimation error $\tilde{f}_2(k)\in\mathbb{R}^n$ as

$$\tilde{f}_2(k)=\hat{f}_2(k)-f_2(k)=(\hat{\psi}_{\mathrm{a}}(k)-\psi_{\mathrm{a}}^{\star})^{\mathrm{T}}\phi_{\mathrm{a}}(k)-\epsilon_2(x(k)). \qquad (56)$$

Using (41) and noting that $\psi_{\mathrm{a}}^{\star}$ does not depend on k, we obtain

$$\begin{aligned}\tilde{\psi}_{\mathrm{a}}(k+1)&=\hat{\psi}_{\mathrm{a}}(k+1)-\psi_{\mathrm{a}}^{\star}\\&=\hat{\psi}_{\mathrm{a}}(k)-\alpha_{\mathrm{a}}\phi_{\mathrm{a}}(k)\left(\hat{J}(k)+Ce(k)-e(k+1)\right)^{\mathrm{T}}-\psi_{\mathrm{a}}^{\star}\\&=\hat{\psi}_{\mathrm{a}}(k)-\psi_{\mathrm{a}}^{\star}-\alpha_{\mathrm{a}}\phi_{\mathrm{a}}(k)\left(\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\tilde{f}_2(x(k))\right)^{\mathrm{T}}\\&=\tilde{\psi}_{\mathrm{a}}(k)-\alpha_{\mathrm{a}}\phi_{\mathrm{a}}(k)\times\\&\quad\left(\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k))\right)^{\mathrm{T}}. \qquad (57)\end{aligned}$$

Based on the last expression, we can formulate the multiplication term $\tilde{\psi}_{\mathrm{a}}^{\mathrm{T}}(k+1)\tilde{\psi}_{\mathrm{a}}(k+1)$ as

$$\begin{aligned}\tilde{\psi}_{\mathrm{a}}^{\mathrm{T}}(k+1)\tilde{\psi}_{\mathrm{a}}(k+1)=&\tilde{\psi}_{\mathrm{a}}^{\mathrm{T}}(k)\tilde{\psi}_{\mathrm{a}}(k)+\alpha_{\mathrm{a}}^2\left\|\phi_{\mathrm{a}}(k)\right\|^2\\&\times(\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k)))\\&\times(\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k)))^{\mathrm{T}}\\&-2\alpha_{\mathrm{a}}\tilde{\psi}_{\mathrm{a}}^{\mathrm{T}}(k)\phi_{\mathrm{a}}(k)\times\\&\left(\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k))\right)^{\mathrm{T}} \qquad (58)\end{aligned}$$

where

$$\Psi_{\mathrm{a}}(k)=\tilde{\psi}_{\mathrm{a}}^{\mathrm{T}}(k)\phi_{\mathrm{a}}(k) \qquad (59)$$

is the approximation error of the actor network output. Utilizing the perfect square trinomial $(a-b)^2=a^2-2ab+b^2$ yields

$$\begin{aligned}&\mathrm{tr}\left[-2\Psi_{\mathrm{a}}(k)\left(\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}^{\mathrm{T}}(k)-\epsilon_2(x(k))\right)\right]\\&=\left\|\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k))-\Psi_{\mathrm{a}}(k)\right\|^2\\&-\left\|\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k))\right\|^2-\left\|\Psi_{\mathrm{a}}(k)\right\|^2. \qquad (60)\end{aligned}$$

Considering the fact that

$$\begin{aligned}&\left\|\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)-\epsilon_2(x(k))\right\|^2\\&=\left\|\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)-\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)+\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\epsilon_2(x(k))\right\|^2\\&=\left\|\tilde{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\epsilon_2(x(k))\right\|^2\\&=\left\|\Psi_{\mathrm{c}}(k)+\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\epsilon_2(x(k))\right\|^2. \qquad (61)\end{aligned}$$

Subsequently, it follows that

$$\begin{aligned}\Delta\mathcal{V}_2(k)=&\frac{1}{\alpha}\Bigg(-(1-\alpha_{\mathrm{a}}\left\|\phi_{\mathrm{a}}(k)\right\|^2)\\&+\left\|\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k))\right\|^2\\&+\left\|\Psi_{\mathrm{c}}(k)+\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\epsilon_2(x(k))\right\|^2-\left\|\Psi_{\mathrm{a}}(k)\right\|^2\Bigg)\\\leq&\frac{1}{\alpha}\Bigg(-(1-\alpha_{\mathrm{a}}\left\|\phi_{\mathrm{a}}^{\mathrm{T}}(k)\right\|^2)\\&+\left\|\hat{\psi}_{\mathrm{c}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k))\right\|^2\\&+2\left\|\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\epsilon_2(x(k))\right\|^2+2\left\|\Psi_{\mathrm{c}}(k)\right\|^2-\left\|\Psi_{\mathrm{a}}(k)\right\|^2\Bigg). \qquad (62)\end{aligned}$$

Incorporating $\Delta\mathcal{V}_1(k)$ and $\Delta\mathcal{V}_2(k)$, $\Delta\mathcal{V}(k)$ is bounded by

$$\begin{aligned}\Delta\mathcal{V}(k)=&\Delta\mathcal{V}_1(k)+\Delta\mathcal{V}_2(k)\\\leq&-(1-\alpha_{\mathrm{c}}\gamma^2\left\|\phi_{\mathrm{c}}(k)\right\|^2)\\&\times\left\|r(k)+\gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)-\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2\\&-\frac{1}{\alpha}(1-\alpha_{\mathrm{a}}\left\|\phi_{\mathrm{a}}(k)\right\|^2)\left\|\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k))\right\|^2\\&+\left(\frac{2}{\alpha}-\gamma^2\right)\left\|\Psi_{\mathrm{c}}(k)\right\|^2-\frac{1}{\alpha}\left\|\Psi_{\mathrm{a}}(k)\right\|^2\\&+\frac{2}{\alpha}\left\|\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\epsilon_2(x(k))\right\|^2\\&+\left\|r(k)+\gamma\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\psi_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2. \qquad (63)\end{aligned}$$

Utilizing the inequalities

$$\begin{aligned}&\frac{2}{\alpha}\left\|\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\epsilon_2(x(k))\right\|^2\leq\frac{4}{\alpha}\left\|\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)\right\|^2+\frac{4}{\alpha}\left\|\epsilon_2(x(k))\right\|^2\\&\left\|r(k)+\gamma\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)-\psi_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2\leq\\&\quad 3\left\|r(k)\right\|^2+3\gamma^2\left\|\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)\right\|^2+3\left\|\psi_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2\end{aligned}$$

results in

$$\begin{aligned}\Delta\mathcal{V}(k) \leq &-(1-\alpha_{\mathrm{c}}\gamma^2\left\|\phi_{\mathrm{c}}(k)\right\|^2)\\ &\times\left\|r(k)+\gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)-\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2\\ &-\frac{1}{\alpha}(1-\alpha_{\mathrm{a}}\left\|\phi_{\mathrm{a}}(k)\right\|^2)\left\|\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k))\right\|^2\\ &+\left(\frac{2}{\alpha}-\gamma^2\right)\left\|\Psi_{\mathrm{c}}(k)\right\|^2+\left(\frac{4}{\alpha}+3\gamma^2\right)\left\|\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)\right\|^2\\ &+\left(\frac{-1}{\alpha}\right)\left\|\Psi_{\mathrm{a}}(k)\right\|^2+\left(\frac{4}{\alpha}\right)\left\|\epsilon_2(x(k))\right\|^2\\ &+3\left\|r(k)\right\|^2+3\left\|\psi_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2. \end{aligned} \tag{64}$$

Assume that $r^{\max}$, $\psi_{\mathrm{c}}^{\max}$, $\phi_{\mathrm{c}}^{\max}$, $\phi_{\mathrm{a}}^{\max}$, and $\epsilon_2^{\max}$ are the upper bounds of $r(k)$, $\psi_{\mathrm{c}}^{\star}$, $\phi_{\mathrm{c}}(k)$, $\phi_{\mathrm{a}}(k)$, and $\epsilon_2(x(k))$, respectively, it yields

$$\begin{aligned}&\left(\frac{4}{\alpha}\right)\left\|\epsilon_2(x(k))\right\|^2+\left(\frac{4}{\alpha}+3\gamma^2\right)\left\|\psi_{\mathrm{c}}^{\star\mathrm{T}}\phi_{\mathrm{c}}(k)\right\|^2\\ &+3\left\|r(k)\right\|^2+3\left\|\psi_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2\\ &\leq\underbrace{\left(\frac{4}{\alpha}\right)(\epsilon_2^{\max})^2+\left(\frac{4}{\alpha}+3\gamma^2+3\right)(\psi_{\mathrm{c}}^{\max\mathrm{T}}\phi_{\mathrm{c}}^{\max})^2+3(r^{\max})^2}_{\Gamma^2}. \end{aligned} \tag{65}$$

Note that based on definition (25), $r^{\max}$ is 1. Using (65), we obtain

$$\begin{aligned}\Delta\mathcal{V}(k) \leq &-(1-\alpha_{\mathrm{c}}\gamma^2\left\|\phi_{\mathrm{c}}(k)\right\|^2)\\ &\times\left\|r(k)+\gamma\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)-\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k-1)\phi_{\mathrm{c}}(k-1)\right\|^2\\ &-\frac{1}{\alpha}(1-\alpha_{\mathrm{a}}\left\|\phi_{\mathrm{a}}(k)\right\|^2)\left\|\hat{\psi}_{\mathrm{c}}^{\mathrm{T}}(k)\phi_{\mathrm{c}}(k)+\Psi_{\mathrm{a}}(k)-\epsilon_2(x(k))\right\|^2\\ &+\left(\frac{2}{\alpha}-\gamma^2\right)\left\|\Psi_{\mathrm{c}}(k)\right\|^2-\frac{1}{\alpha}\left\|\Psi_{\mathrm{a}}(k)\right\|^2+\Gamma^2. \end{aligned} \tag{66}$$

Now by assuming $\alpha > \frac{2}{\gamma^2}$ and if the learning rates α_{a} and α_{c} satisfy

$$\alpha_{\mathrm{c}} < \frac{1}{\gamma^2(\phi_{\mathrm{c}}^{\max})^2} \quad \text{and} \quad \alpha_{\mathrm{a}} < \frac{1}{(\phi_{\mathrm{a}}^{\max})^2} \tag{67}$$

then, the difference $\Delta\mathcal{V}(k)$ is less than zero everywhere outside the compact set defined as

$$\mathcal{S}=\left\{(\Psi_{\mathrm{c}}(k),\Psi_{\mathrm{a}}(k))\,\middle|\,\left\|\Psi_{\mathrm{c}}(k)\right\|\leq\frac{\Gamma}{\sqrt{\gamma^2-\frac{2}{\alpha}}},\left\|\Psi_{\mathrm{a}}(k)\right\|\leq\Gamma\sqrt{\alpha}\right\}. \tag{68}$$

This implies that if the norm of any of the estimation errors is outside of the aforementioned set, it will be brought to inside the set and is guaranteed to stay inside the compact set. Therefore, based on Definition 1, the weight estimation errors of the critic and actor networks are uniformly ultimately bounded.

References

[1] R. Lasseter, "Conditions for stability of droop-controlled inverter-based microgrids," *Automatica*, vol. 1, pp. 305–308, 2002.

[2] J. M. Guerrero, M. Chandorkar, T. L. Lee, and P. C. Loh, "Advanced control architectures for intelligent microgrids—Part I and II," *Automatica*, vol. 60, no. 4, pp. 1254–1270, 2013.

[3] J. Schiffer, D. Zonetti, R. Ortega, A. Stankovic, T. Sezi, and J. Raisch, "A survey on modeling of microgrids—From fundamental physics to phasors and voltage sources," *Automatica*, vol. 74, pp. 135–150, 2016.

[4] C. De Persis, N. Monshizadeh, J. Schiffer, and F. Dorfler, "A Lyapunov approach to control of microgrids with a network-preserved differential-algebraic model," in *Proc. IEEE Conf. Decis. Control*, 2016, pp. 2595–2600.

[5] A. Bidram, B. Poudel, L. Damodaran, R. Fierro, and J. M. Guerrero, "Resilient and cybersecure distributed control of inverter-based islanded microgrids," *IEEE Trans. Ind. Inform.*, vol. 16, no. 6, pp. 3881–3894, Jun. 2020.

[6] A. Krishna, J. Schiffer, and J. Raisch, "Distributed secondary frequency control in microgrids: Robustness and steady-state performance in the presence of clock drifts," *Eur. J. Control*, vol. 51, pp. 135–145, 2019.

[7] M. Eskandari, L. Li, M. Moradi, P. Siano, and F. Blaabjerg, "Active power sharing and frequency restoration in an autonomous networked microgrid," *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 4706–4717, Nov. 2019.

[8] F. Dorfler and S. Grammatico, "Gather-and-broadcast frequency control in power systems," *Automatica*, vol. 79, pp. 296–305, 2017.

[9] C. De Persis and N. Monshizadeh, "Bregman storage functions for microgrid control," *IEEE Trans. Autom. Control*, vol. 63, no. 1, pp. 53–68, Jan. 2018.

[10] A. Gholami and X. Sun, "Towards resilient operation of multimicrogrids: An MISOCP-based frequency-constrained approach," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 3, pp. 925–936, Sep. 2019.

[11] S. Trip, M. Cucuzzella, C. D. Persis, A. van der Schaft, and A. Ferrara, "Passivity-based design of sliding modes for optimal load frequency control," *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 5, pp. 1893–1906, Sep. 2019.

[12] E. Weitenberg, Y. Jiang, C. Zhao, E. Mallada, C. De Persis, and F. Dörfler, "Robust decentralized secondary frequency control in power systems: Merits and trade-offs," *IEEE Trans. Autom. Control*, vol. 64, no. 10, pp. 3967–3982, 2019.

[13] M. Adibi and J. van der Woude, "A reinforcement learning approach for frequency control of inverted-based microgrids," in *Proc. IFAC Workshop Control Smart Grid Renewable Energy Syst.*, 2019, pp. 111–116.

[14] P. Kundur, *Power System Stability and Control*. New York, NY, USA: McGraw-Hill, 1994.

[15] F. Dorfler and F. Bullo, "Kron reduction of graphs with applications to electrical networks," *IEEE Trans. Circuits Syst. I, Regular Papers*, vol. 60, no. 1, pp. 150–163, Jan. 2013.

[16] J. Schiffer, R. Ortega, A. Astolfi, J. Raisch, and T. Sezi, "Conditions for stability of droop-controlled inverter-based microgrids," *Automatica*, vol. 50, no. 10, pp. 2457–2469, 2014.

[17] J. Sarangapani, *Neural Network Control of Nonlinear Discrete-Time Systems*. New York, NY, USA: Taylor & Francis, 2006.

[18] F. L. Lewis, A. Yesildirak, and S. Jagannathan, *Neural Network Control of Robot Manipulators and Nonlinear Systems*. New York, NY, USA: Taylor & Francis, 1998.

[19] H. Bouattour, J. W. Simpson-Porco, F. Dorfler, and F. Bullo, "Further results on distributed secondary control in microgrids," in *Proc. 52nd IEEE Conf. Decis. Control*, 2013, pp. 1514–1519.

[20] J. Schiffer, D. Goldin, J. Raisch, and T. Sezi, "Synchronization of droop-controlled microgrids with distributed rotational and electronic generation," in *Proc. 52nd IEEE Conf. Decis. Control*, 2013, pp. 2334–2339.

[21] P. He and S. Jagannathan, "Reinforcement learning-based output feedback control of nonlinear systems with input constraints," *IEEE Trans. Syst, Man, Cybern., Part B, Cybern.*, vol. 35, no. 1, pp. 150–154, Feb. 2005.

[22] G. Huang, "Learning capability and storage capacity of two-hidden-layer feedforward networks," *IEEE Trans. Neural Netw.*, vol. 14, no. 2, pp. 274–281, Mar. 2003.

[23] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[24] J. Schiffer, "Stability and power sharing in microgrids," Ph.D. dissertation, Technische Universität Berlin, Fakultät IV - Elektrotechnik und Informatik, Berlin, 2015, doi: 10.14279/depositonce-4581.

[25] K. Rudion, A. Orths, Z. Styczynski, and K. Strunz, "Design of benchmark of medium voltage distribution network for investigation of DG integration," in *Proc. IEEE Power Eng. Soc. Gen. Meeting*, 2006.

[26] J. W. Simpson-Porco, Q. Shafiee, F. Dorfler, J. C. Vasquez, J. M. Guerrero, and F. Bullo, "Secondary frequency and voltage control in islanded microgrids via distributed averaging," *IEEE Trans. Ind. Electron.*, vol. 62, no. 11, pp. 7025–7038, Nov. 2015.